

Maristella Agosti  
Floriana Esposito  
Stefano Ferilli  
Nicola Ferro (Eds.)

Communications in Computer and Information Science

354

# Digital Libraries and Archives

8th Italian Research Conference, IRCDL 2012  
Bari, Italy, February 2012  
Revised Selected Papers

Editorial Board

Simone Diniz Junqueira Barbosa

*Pontifical Catholic University of Rio de Janeiro (PUC-Rio),  
Rio de Janeiro, Brazil*

Phoebe Chen

*La Trobe University, Melbourne, Australia*

Alfredo Cuzzocrea

*ICAR-CNR and University of Calabria, Italy*

Xiaoyong Du

*Renmin University of China, Beijing, China*

Joaquim Filipe

*Polytechnic Institute of Setúbal, Portugal*

Orhun Kara

*TÜBİTAK BİLGEM and Middle East Technical University, Turkey*

Tai-hoon Kim

*Konkuk University, Chung-ju, Chungbuk, Korea*

Igor Kotenko

*St. Petersburg Institute for Informatics and Automation  
of the Russian Academy of Sciences, Russia*

Dominik Ślęzak

*University of Warsaw and Infobright, Poland*

Xiaokang Yang

*Shanghai Jiao Tong University, China*

Maristella Agosti Floriana Esposito  
Stefano Ferilli Nicola Ferro (Eds.)

# Digital Libraries and Archives

8th Italian Research Conference, IRCDL 2012  
Bari, Italy, February 9-10, 2012  
Revised Selected Papers



Springer

Volume Editors

Maristella Agosti  
Nicola Ferro  
University of Padua  
Department of Information Engineering  
Via Gradenigo, 6/a  
35131 Padua, Italy  
E-mail: {agosti, ferro}@dei.unipd.it

Floriana Esposito  
Stefano Ferilli  
University of Bari  
Department of Computer Science  
Via E. Orabona, 4  
70126 Bari, Italy  
E-mail: {esposito, ferilli}@di.uniba.it

ISSN 1865-0929 e-ISSN 1865-0937  
ISBN 978-3-642-35833-3 e-ISBN 978-3-642-35834-0  
DOI 10.1007/978-3-642-35834-0  
Springer Heidelberg Dordrecht London New York

Library of Congress Control Number: 2012954629

CR Subject Classification (1998): H.3.3-7, H.5.1, H.5.4, J.1, H.2.8, I.7.4, H.4.m

© Springer-Verlag Berlin Heidelberg 2013

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

The use of general descriptive names, registered names, trademarks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

*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 a yearly appointment for Italian researchers, both on the computer science and on the humanities side, interested in digital libraries and related topics. The focus of the eighth conference was on legacy and cultural heritage material. Indeed, digital library systems are becoming increasingly mature and more widely deployed. Not only do they have to grant users effective and personalized access to information, but they also now have to address the need for smoothly processing and including in the DL repositories the available legacy and cultural heritage documents, in addition to born-digital ones. This calls for the ability to deal with compound objects in different media, to provide uniform solutions and methodologies across different cultural heritage institutions, and to take into account preservation, restoration, and curation.

The IRCDL conferences were launched 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.

This volume contains the revised accepted papers selected from among those presented at the 8th Italian Research Conference on Digital Libraries (IRCDL 2012), which was held at the University of Bari ‘Aldo Moro’, Bari, February 9–10, 2012.

The aim 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 16 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 papers presented orally. Those papers underwent a new review process and the results of the selection are the papers appearing in this volume. The topics covered are related to the different aspects needed to support information access and interoperability; among those there are:

- legacy documents and cultural heritage
- systems interoperability and data integration
- formal and methodological foundations of DLs
- semantic web and linked data for DLs
- multilingual information access
- DL infrastructures
- metadata creation and management
- search engines for digital library systems

- evaluation and log data
- handling audio/visual and non-traditional objects
- user interfaces and visualization
- DL quality
- policies and copyright issues in DLs
- scientific data curation, citation and scholarly publication
- user behavior and modeling
- preservation and curation.

Taking into consideration that the Italian research community is involved in different relevant projects related at large to the area of digital libraries, it was decided to report in the volume the most recent results from the diverse projects, in order to transfer them to the international community.

We would like to thank those institutions and individuals who made the conference and this volume possible:

- the Program Committee members,
- the Steering Committee members,
- the Knowledge Acquisition and Machine Learning Lab (LACAM) of the Department of Computer Science of the University of Bari ‘Aldo Moro’, and the members of the same research group who contributed to the organization of the event, namely Teresa M.A. Basile, Claudio Taranto, and Domenico Redavid, and
- the Department of Information Engineering of the University of Padua.

A Call for Participation for IRCDL 2013 will be circulated, but meanwhile we invite all researchers having research interests in digital libraries to start thinking about possible contributions to next year’s conference.

September 2012

Maristella Agosti  
Floriana Esposito  
Stefano Ferilli  
Nicola Ferro

# Organization

IRCDL 2012 was organized by the Knowledge Acquisition and Machine Learning Lab (LACAM) of the Department of Computer Science of the University of Bari ‘Aldo Moro’.

## General Chairs

Maristella Agosti	University of Padua
Floriana Esposito	University of Bari ‘Aldo Moro’

## Program Chairs

Stefano Ferilli	University of Bari ‘Aldo Moro’
Nicola Ferro	University of Padua

## Local Organization Committee

Teresa M.A. Basile	University of Bari ‘Aldo Moro’
Claudio Taranto	University of Bari ‘Aldo Moro’
Domenico Redavid	Artificial Brain S.r.l.

## Program Committee

Nicola Barbuti	University of Bari
Marco Bertini	University of Florence
Francesco Buccafurri	University of Reggio Calabria
Leonardo Candela	ISTI CNR, Pisa
Michelangelo Ceci	University of Bari
Nicola Di Mauro	University of Bari
Costantino Grana	University of Modena
Maria Guercio	University of Urbino “Carlo Bo”
Simone Marinai	University of Florence
Carlo Meghini	ISTI CNR, Pisa
Nicola Orio	University of Padua
Fausto Rabitti	ISTI CNR, Pisa
Giuseppe Serra	University of Florence
Anna Maria Tammaro	University of Parma
Erlide Terenzoni	Soprintendente archivistico per il Veneto, Venice
Mario Vento	University of Salerno

## **IRCDL Steering Committee**

Maristella Agosti	University of Padua
Tiziana Catarci	University of Rome “La Sapienza”
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 2012 benefited from the support of the following organizations:

Department of Computer Science, University of Bari ‘Aldo Moro’

Department of Information Engineering, University of Padua

Centro Interdipartimentale di Logica e Applicazioni, University of Bari ‘Aldo Moro’

University of Bari ‘Aldo Moro’



# Ontology-Based Integration of Cultural Heritage Metadata (Invited Talk)

Christos Papatheodorou

Dept. of Archives and Library Sciences – Ionian University  
72 Ioannou Theotoki str., 49100 Corfu, Greece  
papatheodor@ionio.gr

Managing heterogeneous data is a challenge for cultural heritage institutions, archives, libraries, and museums which usually develop collections with heterogeneous types of material, described by different metadata schemas. For example, the Library of Congress, USA, provides EAD metadata for the archives description, MARC 21 records for the description of a wide variety of material, such as books and photographs, Text Encoding Initiative (TEI) for documenting the text of digital reproductions in the American Memory Collection, etc. The wide use of a number of cultural heritage metadata schemas imposes the development of interoperability techniques that facilitate unified access to cultural resources. One of the widely implemented techniques is the Ontology-Based Integration. Ontologies provide formal specifications of a domain's concepts and their interrelations and act as a mediated schema between heterogeneous sources.

This presentation describes an ontology-based metadata integration architecture. Its components are the mediator, which is based on the CIDOC CRM, the local sources, whose schemas are XML-based metadata schemas, and the mappings between the local sources and the mediator. Four integration scenarios are proposed on this architecture. Moreover a mapping language is demonstrated, called Mapping Description Language (MDL), to define the mappings of the metadata schemas of the local sources to the CIDOC CRM ontology as rules. An application of MDL usage to define the mapping from Encoded Archival Description (EAD) to the CIDOC CRM ontology is exhibited. Finally an algorithm for the transformation of EAD metadata to CIDOC CRM, as well as a query transformation algorithm from XPATH to CIDOC CRM are demonstrated.

# Table of Contents

## Panel

Experiences and Perspectives in Management for Digital Preservation of Cultural Heritage Resources (Panel) . . . . .	1
<i>Maristella Agosti</i>	
Where Do Humanities Computing and Digital Libraries Meet? . . . . .	4
<i>Dino Buzzetti</i>	
The ArchiMEDE Project for an Electronically Digitized Archive of Historical Monographs . . . . .	11
<i>Onofrio Erriquez</i>	
Considerations on the Preservation of Base Digital Data of Cultural Resources . . . . .	13
<i>Nicola Barbuti</i>	

## Papers

Supporting Tabular Data Characterization in a Large Scale Data Infrastructure by Lexical Matching Techniques . . . . .	21
<i>Leonardo Candela, Gianpaolo Coro, and Pasquale Pagano</i>	
Data Interoperability and Curation: The European Film Gateway Experience . . . . .	33
<i>Michele Artini, Alessia Bardi, Federico Biagini, Franca Debole, Sandro La Bruzzo, Paolo Manghi, Marko Mikulicic, Pasquale Savino, and Franco Zoppi</i>	
Annotating Digital Libraries and Electronic Editions in a Collaborative and Semantic Perspective . . . . .	45
<i>Michele Barbera, Federico Meschini, Christian Morbidoni, and Francesca Tomasi</i>	
Empowering Archives through Annotations . . . . .	57
<i>Nicola Ferro and Gianmaria Silvello</i>	
Metadata Inference for Description Authoring in a Document Composition Environment . . . . .	69
<i>Tsuyoshi Sugibuchi, Ly Anh Tuan, and Nicolas Spyrtos</i>	
A Multi-layer Digital Library for Mediaeval Legal Manuscripts . . . . .	81
<i>Monica Palmirani and Luca Cervone</i>	

Extracting Keyphrases from Web Pages . . . . .	93
<i>Felice Ferrara and Carlo Tasso</i>	
Learning to Recognize Critical Cells in Document Tables . . . . .	105
<i>Nicola Di Mauro, Stefano Ferilli, and Floriana Esposito</i>	
Document Image Understanding through Iterative Transductive Learning . . . . .	117
<i>Michelangelo Ceci, Corrado Loglisci, Lucrezia Macchia, Donato Malerba, and Luciano Quercia</i>	
A Domain Based Approach to Information Retrieval in Digital Libraries . . . . .	129
<i>Fulvio Rotella, Stefano Ferilli, and Fabio Leuzzi</i>	
Uncertain (Multi)Graphs for Personalization Services in Digital Libraries . . . . .	141
<i>Claudio Taranto, Nicola Di Mauro, and Floriana Esposito</i>	
Improving Online Access to Archival Data . . . . .	153
<i>Vittore Casarosa, Carlo Meghini, and Stanislava Gardasevic</i>	
Quick and Easy Implementation of Approximate Similarity Search with Lucene . . . . .	163
<i>Giuseppe Amato, Paolo Bolettieri, Claudio Gennaro, and Fausto Rabitti</i>	
Establishing a Digital Library in Wide-Ranging University's Context: The Sapienza Digital Library Experience . . . . .	172
<i>Angela Di Iorio, Marco Schaerf, and Matteo Bertazzo</i>	
Digital Curators' Education: Professional Identity vs. Convergence of LAM (Libraries, Archives, Museums) . . . . .	184
<i>Anna Maria Tammaro, Melody Madrid, and Vittore Casarosa</i>	
A Contribution for the Dissemination of Cultural Heritage Content to a Wider Public . . . . .	195
<i>Maristella Agosti, Lucio Benfante, and Nicola Orio</i>	
Engaging the User: Elaboration and Execution of Trials with a Database of Illuminated Images . . . . .	207
<i>Chiara Ponchia</i>	
Modeling Archives by Means of OAI-ORE . . . . .	216
<i>Nicola Ferro and Gianmaria Silvello</i>	
Reflecting on the Europeana Data Model . . . . .	228
<i>Silvio Peroni, Francesca Tomasi, and Fabio Vitali</i>	

The Europeana Linked Open Data Pilot Server . . . . .	241
<i>Nicola Aloia, Cesare Concordia, and Carlo Meghini</i>	
Managing Authenticity through the Digital Resource Lifecycle . . . . .	249
<i>Maria Guercio and Silvio Salza</i>	
An Innovative Character Recognition for Ancient Book and Archival Materials: A Segmentation and Self-learning Based Approach . . . . .	261
<i>Nicola Barbuti and Tommaso Caldarola</i>	
<b>Author Index</b> . . . . .	271

# Experiences and Perspectives in Management for Digital Preservation of Cultural Heritage Resources (Panel)

Maristella Agosti

Department of Information Engineering, University of Padua, Italy  
agosti@dei.unipd.it

**Abstract.** This paper reports on the panel objectives, on the topics addressed during the panel, and on the following discussion.

A relevant conclusion that emerges from the panel is the need to discuss and define a shared Digital Agenda for Italy.

**Keywords:** digital cultural heritage, cultural heritage resources, digitisation, online accessibility, digital preservation, digital agenda for Italy.

## 1 Panel Objectives

This paper reports on the panel conducted in the context of IRCDL 2012, the eighth edition of the Italian Research Conference on Digital Libraries<sup>1</sup>. The objectives of the panel are consistent with the focus of IRCDL 2012 which is devoted to legacy and cultural heritage material, and in particular to curating the digital cultural heritage resources and making them available to a wider audience of scholars as well as general public.

Digital Libraries have attracted much attention in recent years both from academics and professionals interested in envisaging and designing new tools and systems able to manage diversified collections of documents, artifacts, and data in digital form in a consistent and coherent way<sup>[1]</sup>.

As reported in <sup>[2]</sup>:

The Digital Agenda for Europe seeks to optimise the benefits of information technologies for economic growth, job creation and the quality of life of European citizens, as part of the Europe 2020 strategy. The digitisation and preservation of Europe's cultural memory which includes print (books, journals and newspapers), photographs, museum objects, archival documents, sound and audiovisual material, monuments and archaeological sites (hereinafter "cultural material") is one of the key areas tackled by the Digital Agenda.

---

<sup>1</sup> IRCDL conference series, URL:

<http://ims.dei.unipd.it/websites/ircdl/home.html>

The EU's strategy for digitisation and preservation builds on the work done over the last few years in the digital libraries initiative. The European actions in this area, including the development of Europeana, Europe's digital library archive and museum, were supported by the European Parliament and the Council, most recently in a Parliament resolution of 5 May 2010 and the Council Conclusions of 10 May 2010. The Workplan for Culture 2011-14, established by the Council at its meeting of 18 and 19 November 2010, highlights the need for a coordinated effort in the area of digitisation.

It is the time to build on the experience that has been gained in past recent years with the digital libraries initiatives to invest in systems and tools that make effective the online accessibility of cultural heritage resources. In fact, digitisation is an initial important step towards online accessibility, but it must be supplemented with methods and techniques that permit diversified representations of the contents of the digitised resources. Those diversified representations are necessary for the effective management of the resources through information management systems able to give users online access to the resources.

In recent years, much experience both of digitisation of cultural heritage resources and of design and development of systems able to manage cultural heritage resources have been gained. The systems have been developed in a close relationship with the digitisation of cultural and scientific collections and efforts have been made so that traditionally different communities of specialists become better acquainted with one other. One relevant example of these efforts is Europeana<sup>2</sup> which can be considered a single access point to books, paintings, films, museum objects and archival records that have been digitised throughout Europe.

The new challenge is to succeed in dealing with the fragmentation and specificity of past solutions and envisage new systems able to become part of the cultural exploration and the study of professionals and users of cultural heritage resources. In this way tools will be supplied that incorporate methods that can change the cultural experience of professional and general users.

## 2 The Panelists

The reports of the speeches of the panelists are included in the following volume and they address the different facets related to the necessary differentiated methods and skills.

The panelists and their representative areas are:

- Dino Buzzetti, Humanities Area of the University of Bologna – semantic technologies for representation and management,
- Onofrio Erriquez, Physics Area of the University of Bari – digitisation of historical document archives, and
- Nicola Barbuti, Archives and Libraries Area of the University of Bari – preservation of cultural heritage resources and management.

---

<sup>2</sup> Europeana Web site, URL: <http://www.europeana.eu/portal/>

### 3 Discussion

After the presentations of the panelists a lively discussion took place. Many conference participants contributed to the discussion of the different panel topics, including Floriana Esposito (University of Bari), Carlo Meghini (ISTI-CNR Pisa), Anna Maria Tammaro (University of Parma), Mariella Guercio (University of Rome “La Sapienza”), Carlo Dell’Aquila (University of Bari), Francesca Tomasi (University of Bologna), Stefano Ferilli (University of Bari), and Donato Malerba (University of Bari).

An aspect that emerged from all the contributions to the discussion is the need to discuss and define a shared Digital Agenda for Italy.

**Acknowledgements.** Sincere thanks are due to Floriana Esposito, Stefano Ferilli and Nicola Ferro for the time they spent with the author discussing the aspects related to the management of cultural heritage resources and the way to address them in the context of the panel.

The author would like to thank the Panelists for accepting to participate in the panel and to write a report of their speech, thus actively contributing to laying the foundation for the definition of a new approach to dealing with digital cultural heritage resources.

The work reported has been partially supported by the CULTURA project (reference: 269973<sup>3</sup>), and by the PROMISE network of excellence project (reference: 258191<sup>4</sup>), as part of the Seventh Framework Programme of the European Commission.

### References

1. Agosti, M.: Digital libraries. In: Melucci, M., Baeza-Yates, R., Croft, W.B. (eds.) *Advanced Topics in Information Retrieval*. The Information Retrieval Series, vol. 33, pp. 1–26. Springer, Heidelberg (2011)
2. Recommendations: Commission Recommendation of 27.10.2011 on the digitisation and online accessibility of cultural material and digital preservation. *Official Journal of the European Union* L 283, 39–45 (2011)

---

<sup>3</sup> <http://www.cultura-strep.eu/>

<sup>4</sup> <http://www.promise-noe.eu/>

# Where Do Humanities Computing and Digital Libraries Meet?

Dino Buzzetti

University of Bologna, Italy  
buzzetti@philo.unibo.it

## 1 Introduction

It is in libraries that humanists have always found their basic and essential instrumentation. Libraries can be described as the humanist's lab. Obviously this applies also to digital humanists, who deal with digital objects for research purposes, and to digital libraries that store collections in digital form. But digital objects produced for research purposes are not just inactive artefacts and 'digital library objects are more than collections of bits,' for 'the content of even the most basic digital object has some structure' and to enable access and transactions additional information or 'metadata' is required. [1] So 'if, unlike print,' digital editions 'are also open-ended and collaborative work-sites rather than static closed electronic objects' (p. 77), [2] it can be legitimately asked how a digital repository for objects of this kind can enable effective access to the interactive functionalities they provide. In a digital research context, the issue of how the architecture of a digital library could meet the needs of the working practices increasingly adopted by digital humanists seems therefore of primary importance.

But how can we define humanities computing and what are its requirements? A plausible answer can be found in the final report of a European Thematic Network on Advanced Computing in the Humanities (ACO\*HUM):

[...] we will attempt to define the core in terms of the traditional combination of data structures and algorithms, applied to the requirements of a discipline: (a) the methods needed to represent the information within a specific domain of knowledge in such a way that this information can be processed by computational systems result in the data structures required by a specific discipline; (b) the methods needed to formulate the research questions and specific procedures of a given domain of knowledge in such a way as to benefit from the application of computational processing result in the algorithms applicable to a given discipline. [3]

In this understanding, digital objects representing primary source materials, should be endowed with specific functionalities capable of answering specific research questions. Accessing this kind of resources should not prevent the applicability of such functionalities and that is precisely the point where digital humanities and digital libraries can actually meet.



## 2 The Creation of Digital Resources

The creation of digital resources for the humanities, however, has not remained unaffected by major technological developments. Humanities computing research practices have remarkably changed along with the availability of different computational means. Whereas with the use of mainframes the emphasis was placed chiefly on content processing, with the advent of personal computers and even more so with the introduction of the WorldWideWeb, the interest shifted to the representation of the original source materials. As John Unsworth has timely observed,

we are, I think, on the verge of what seems to me the third major phase in humanities computing, which has moved from tools in the 50s, 60s, and 70s, to primary sources in the 80s and 90s, and now seems to be moving back to tools [...]. I think we are arriving at a moment when the form of the attention that we pay to primary source materials is shifting from digitizing to analyzing, from artifacts to aggregates, and from representation to abstraction. [4]

And again, clearly, the now emerging ‘third phase’ in humanities computing is substantially enhanced by the development of the new Semantic Web technologies. Nowadays research practices in humanities computing are actually moving back from representing sources in digital form to designing tools to process their information content:

We’ve spent a generation furiously building digital libraries, and I’m sure that we’ll now be building tools to use in those libraries, equally furiously, for at least another generation. [4]

But do the needs of advanced digital humanities practice and research find satisfactory support in current digital library environments and architecture? Can digital libraries designers and digital humanists join their efforts to set up a common research agenda?

## 3 The Case of Digital Editions

To better trace these developments, we may consider, by way of example, the case of digital editions. In the late 80’s and early 90’s a digital edition was thought of as a way of representing a text and its entire textual tradition as a database, [5] because at that time, in order to bind passages of text to selections of their manuscript images it was necessary to integrate textual and visual elements in a single DBMS capable of handling both kinds of structured information. Accordingly, and more to the point, the transcription of the original documents was not meant, like a diplomatic transcription, as a means to convey to the reader ‘a closer idea of the nature of the source’ (p. 145), but ‘as data to be processed’; and so, in this understanding, it was assumed that the transcription of a document

becomes an activity of data modelling and encoding in order to elicit as much information as possible from the manuscript and to infer new analytical results. From this point of view, both the image and the transcript are not regarded as physical reproductions referring back to the original document but rather as analytical data pointing toward a new logical representation of the source (p. 148). [6]

The emphasis was still on processing, even after the introduction of graphic user interfaces. A digital text representation was still conceived of as data for further processing rather than as a means to visualise a physical document. But things gradually changed as the emphasis shifted more and more towards visualisation on graphic Web browsers and computer screens. The Web was chiefly meant for remote access and visual display, whereas WYSIWYG systems and page description languages promoted an ever increasing tendency towards the ‘electronic simulation of specific print objects’ (27) [2]. Digitisation projects and the visual representation of primary sources became the prevailing interest in humanities computing.

## 4 The Form of Attention of Digital Humanities

The ‘forms of attention’ of digital humanities – see [7] – shifted then from processing to representation. And the new developments in technology encouraged that process. In computer science, besides the so-called data processing or database community a large and authoritative document community grew up and established itself. [8] Both groups suffered from the problem of having their data ‘trapped’ in proprietary systems. The dissatisfaction of the document community with its early systems led to ‘generalize its markup’ and to endorse the ISO SGML standard, a markup language that was accessible to the writer and allowed to encode not only the ‘presentational aspects of documents,’ but also ‘more general properties of texts’ (p. 26). Since ‘for the document community, the factor of most permanence was the document,’ that community ‘chose to standardize the representation of data.’ On the other hand, ‘for the database community, the factor of most permanence was the semantics of applications,’ and so that community ‘chose to standardize the semantics of data.’

These different leanings proved decisive for the choices of the scholarly community. Three foremost humanities computing associations, the Association for Computational Linguistics (ACL), the Association for Literary and Linguistic Computing (ALLC) and the Association for Computers and the Humanities (ACH) decided to promote the Text Encoding Initiative (TEI) and to adopt the ISO SGML standard for the encoding of texts. ‘Data semantics was not irrelevant to the document community,’ but the definition of semantics ‘did seem to be a difficult problem’ (p. 27). And also ‘attempts to define semantics in the scholarly community, most notably the Text Encoding Initiative, similarly met with resistance.’ Thus, ‘the route proposed by SGML’ seemed ‘a reasonable one’ and the scholarly community conformed to it:

promote the notion of application and machine independence, and provide a base on which semantics could eventually be developed, but avoid actually specifying a semantics (p. 28).

As a consequence, the centre of attention moved over from processing information content to mere data representation.

## 5 The Web and Its Languages

The same repercussions can be noticed by observing the expansion of the Web. It is not by chance, that the languages employed in the construction of the Web, HTML and now ever increasingly XML, are basically data representation languages. They express the structure of the representation, not the structure of what is represented, unless the two structures match and can be put into a one-to-one correspondence. The processing of the documents accessible on the Web depends on the structure these languages assign to them, and thus on the constraints of a hierarchical tree structure. XSLT, the language introduced to process data in XML format, 'takes a tree structure as its input, and generates another tree structure as its output.' [9] It preserves the structure of the document and what it can process is not the structure of the information content it conveys.

Since it allows easy access and excellent visualisation, the Web has been confidently envisioned as a potential universal library. In this conviction, a number of large-scale digitisation projects were begun, such as the Million Book Project (also known as the Universal Library), led by Carnegie Mellon University and started in 2002, the Google Book Search Project started in 2004, and the Microsoft's MSN Book Search, announced in 2005 and subsequently discontinued. But, as it has been observed by Deegan and Sutherland [2], 'the paradigm for the universal library' they enforce 'is not a library at all, it is the Internet' (p. 151). And the Internet really is a different kind of information space from a library. In the Internet the 'professional organisational principles' that belong to the library science tradition 'do not appear to be carried over' (p. 150); in the information space created by the Internet, 'order' is virtually neglected and so 'one of the major benefits that libraries bring to the almost boundless intellectual space that is our literate culture is lost' (p. 149). All in all, as Deegan and Sutherland maintain, 'Google "Book" Search (note our inverted commas) is not providing electronic text, it is providing books' (p. 147). The emphasis is again on the document – the book – and not on its information content – the text. Mass digitisation projects show, once more, that in the Web 'the potential of the computer as visualisation tool' has probably overtaken its analytical and, for many humanists, more appropriate 'computational' uses (p. 75).

## 6 Major Technological Innovations

How, then, can we explain that major technological innovations such as the introduction of personal computers and the expansion of the Web produced almost paradoxical

effects on humanities computing? How could they hold back the development of its methods and research practices? We may assume an evolutionary point of view to look for a possible answer. What matters more for the evolution of biological organisms are not so much their external features, but rather their physiological capabilities and functions. In a similar way, if a digital object can now be visualised as the reproduction of a corresponding physical object, it can also be evaluated for its functionalities and the available facilities to process the information content it conveys. Functional as opposed to visual features are what really matters.

Now, on the one hand, ‘what humanities computing has been doing, implicitly, for years’ can in many ways be described as ‘knowledge representation.’ [10] But, on the other hand, if knowledge representation is legitimately seen as ‘a medium for pragmatically efficient computation,’ [11] representing information and processing information cannot be regarded as separate activities, each one opposed to the other. The form of a knowledge representation can actually be thought of as depending on the computational procedures aimed at processing its information content.

It is precisely for their concern over processing information content that humanities computing research practices are now aligning with those of relevant neighbouring fields. In the specific domain of knowledge organisation and subject indexing, Vanda Broughton, an expert in faceted classification systems and thesaurus construction, observes:

Current co-operative work with scholars in the area of humanities computing suggests that, in combination, facet analytical and text encoding methods may offer a solution to improving the usability of metadata tools and providing more subtle and sophisticated means of subject representation (p. 193). [12]

Here knowledge organisation and humanities computing concur expressly on the analysis of information content, which is exactly what the new Semantic Web technologies are aiming for. Thus it is indeed the technological evolution of the Web what can help recover that partially neglected aspect of humanities computing which its nascent construction momentarily and almost paradoxically contributed to obscure.

## 7 The Semantic Web

With the help of these new technologies humanities computing can get back to its original inspiration: the ‘attention’ that in a successive phase of its development was mostly directed to the ‘representation of primary source materials’ goes back again to building ‘tools’ for processing their information content. [4] Humanities scholars too recognise ‘the semantic web’ as their ‘future’ and humanities computing is thus bound to produce ‘formal representations of the human record’ suitable for automatic processing. For, as John Unsworth again reminds us,

those representations – ontologies, schemas, knowledge representations, call them what you will – should be produced by people trained in the humanities. Producing them is a discipline that requires training in the humanities, but also in elements of mathematics, logic, engineering, and computer science [...]. There is a great deal of work for such people to do – not all of it technical, by any means. Much of this map-making will be social work, consensus-building, compromise. But even that will need to be done by people who know how consensus can be enabled and embodied in a computational medium. [13]

New developments induced by Semantic Web technologies can also be observed in the field of digital libraries. An interesting example is offered by the so-called semantic digital libraries, [14] whose declared purpose is to integrate Semantic Web and social networking technologies into a digital library management system. The basic assumption, here, is that ‘semantic technologies can offer more efficient solutions for building robust, user-friendly ways of accessing content and metadata.’ [15] Semantic technologies, it is averred, can supply ‘efficient discovery techniques in the new, interconnected information space’ of digital resources accessible on the Web. The use of ontologies produces new forms of information and knowledge organisation, that do not reduce themselves to a ‘mere specification of metadata schemata’ previously established, but allow ‘metadata to become more open, unstructured, and what is most important, highly interlinked’ (p. 78-79). [16] The use of lightweight tag ontologies ‘provides the possibility for machine-processable representations that can be shared across social tagging systems.’ [17] The practice of social tagging can then usefully help to integrate valuable sources of semantic annotations in a digital library platform that provides linked data services.

The application of semantic annotation technologies both in digital library systems and humanities computing applications clearly shows that in both fields a need for common functionalities is actively felt. The case could easily be generalised and may wishfully prompt a closer reflection on the prospects of a common research agenda for digital libraries and digital humanities.

## References

1. Arms, W.Y.: Key Concepts in the Architecture of the Digital Library. *D-Lib Magazine* 1(1) (1995), <http://www.dlib.org/dlib/July95/07arms.html>
2. Deegan, M., Sutherland, K.: *Transferred Illusions: Digital Technology and the Forms of Print*. Ashgate, Farnham (2009)
3. Thaller, M.: Defining humanities computing methodology. In: de Smedt, K., et al. (eds.) *Computing in Humanities Education: A European Perspective*, ch. 2.3, University of Bergen-HIT Centre (1999), <http://www.hd.uib.no/AcoHum/book/fm-chapter-final.html>
4. Unsworth, J.: *Forms of Attention: Digital Humanities Beyond Representation*. Paper Presented at the 3rd Conference of the Canadian Symposium on Text Analysis (CaSTA). *The Face of Text: Computer-Assisted Text Analysis in the Humanities*, McMaster University (2004), <http://people.lis.illinois.edu/~unsworth/FOA/>

5. Buzzetti, D.: Masters and Books in 14th-century Bologna: An edition as a database. In: Bocchi, F., Denley, P. (eds.) *Storia & Multimedia*, Proceedings of the 7th International Congress of the Association for History and Computing, August 29-September 2, pp. 642–646. Grafis Edizioni, Bologna (1994)
6. Buzzetti, D.: Image Processing and the Study of Manuscript Textual Traditions. *Historical Methods* 28(3), 145–154 (1995)
7. Kermode, F.: *Forms of attention*. The University of Chicago Press, Chicago (1985)
8. Raymond, D., Tampa, F., Wood, D.: From data representation to data model: Meta-semantic issues in the evolution of SGML. *Computer Standards & Interfaces* 18, 25–36 (1996)
9. Kay, M.: What Kind of Language is XSLT? An analysis and overview (2005), <http://www.ibm.com/developerworks/library/x-xslt/>
10. Unsworth, J.: Knowledge Representation in Humanities Computing. Paper Presented at eHumanities. NEH Lecture Series on Technology & the Humanities, Lecture I, Washington, DC, vol. 4 (2001), <http://people.lis.illinois.edu/~unsworth/KR/>
11. Davis, R., Shrobe, H., Szolovits, P.: What is a Knowledge Representation? *AI Magazine* 14(1), 17–33 (1993)
12. Broughton, V.: Finding Bliss on the Web: Some problems of representing faceted terminologies in digital environment. In: Gnoli, C., Mazzocchi, F. (eds.) *Paradigms and Conceptual Systems in Knowledge Organization*, pp. 188–194. Ergon-Verlag, Würzburg (2010)
13. Unsworth, J.: What is Humanities Computing and What is Not? *Jahrbuch für Computerphilologie* 4, 71–84 (2002), <http://computerphilologie.tu-darmstadt.de/jg02/unsworth.html>
14. Semantic Digital Libraries: Bringing Libraries to Web 3.0, <http://sem1.info/>
15. Kruk, S.R., Westerski, A., Kruk, E.: Architecture of Semantic Digital Libraries. Digital Enterprise Research Institute (DERI), National University of Ireland, Galway, Work Package, vol. 4, pp. 1–12 (2008)
16. Kruk, S.R., Westerski, A., Kruk, E.: Architecture of Semantic Digital Libraries. In: Kruk, S.R., McDaniel, B. (eds.) *Semantic Digital Libraries*, pp. 77–85. Springer, Berlin (2009)
17. Kim, H.L., et al.: The state of the art in tag ontologies: A semantic model for tagging and folksonomies. In: *DCMI 2008: Proceedings of the 2008 International Conference on Dublin Core and Metadata Applications*, Dublin Core Metadata Initiative (2008), <http://dc2008.de/wp-content/uploads/2008/09/kim-scerri-breslin-decker-kim.pdf>

# The ArchiMEDE Project for an Electronically Digitized Archive of Historical Monographs

Onofrio Erriquez

University of Bari, Italy  
erriquez@fisica.uniba.it

First of all I would like to thank the organizers for having invited me to this round table discussion, and in particular my colleagues Floriana Esposito, whom I studied with at university, and Maristella Agosti, whom I have had the pleasure of meeting for the first time on this occasion.

As you are all well aware, there are numerous initiatives involving the digitization of cultural heritage, which have the dual aim of preserving cultural heritage, as proposed by the topic of this round table discussion, and improving its accessibility.

Allow me to mention just two examples: the MiBAC CulturalItalia portal and the MICHAEL project (Multilingual Inventory of Cultural Heritage in Europe). These two initiatives also happen to be the main suppliers of Italian digital resources for the European Union's Europeana Portal.

It could be said that the University of Bari, which was founded in 1925, is far too young to have books worthy of being part of a digitization project. However, thanks to numerous donations and several acquisitions, the university can boast a rather prestigious collection, which includes over 2,000 volumes from the 16th century.

In particular, there are the volumes on Roman Law included in the donation made by the Stella-Maranca Foundation, which belonged to the first Dean of our Faculty of Law, and the volumes of the Chioventa Foundation acquired by the former Botany Institute.

This is the setting which gave rise to the idea of the ArchiMEDE Project (Archivio Monografie d'Epoca Digitalizzate Elettronicamente – Archive of Electronically Digitized Historical Monographs) which was approved and funded by the Fondazione Cassa di Risparmio di Puglia with a EUR 75,000 grant and will draw to a close in October 2012.

As mentioned earlier, the aim of the project is to preserve and make accessible the considerably prestigious cultural heritage possessed by our university.

In order to achieve this aim, the project involves the following:

- start-up of the digitization of the collection of 16th century volumes (the first 500 of over 2,000) using a planetary scanner acquired by our university in the setting of the "Unknown Heritage" project of Nicola Barbuti and involving specially selected and trained external staff;
- cataloguing (historical book) and classification of the digitized heritage;
- inclusion in the university's catalogue and the relative OPAC;

- creation of a consultable archive;
  - in the geographic network (resolution 72 dpi);
  - in the local network (resolution 300 dpi);
  - on DVD (resolution 600 dpi).

The hope but also the conviction is that this heritage – in part unique, especially in Southern Italy – may emerge from the limbo of poorly known or even completely unknown heritage and return to life thanks to scholarly consultation.



# Considerations on the Preservation of Base Digital Data of Cultural Resources

Nicola Barbuti

Dipartimento di Scienze dell'antichità e del tardoantico, University of Bari, Italy  
nicola.barbuti@uniba.it

## 1 Introduction

This paper does not aim to thoroughly list and discuss issues which are well-known in research circles working towards defining the correct strategies for the preservation of cultural heritage digital resources, nor is it an attempt to suggest possible solutions to the current problems, as this is a burdensome task which has been tackled by individuals much more professionally and theoretically qualified than myself.

This paper does, however, provide food for thought and raises pertinent questions which come up on a daily basis for those who through their research or work encounter the aforementioned issues. The sole aim, if at all possible, is look upon the issue from a different perspective to those adopted in the discussions and theoretical debates on the delicate and still unresolved question of digital preservation.

## 2 State of the Art

The common definition of digital preservation is ‘the collection of activities and instruments which guarantee that digital documents are kept accessible, usable (legible and intelligible) and authentic (unambiguously identifiable and intact) in the medium and the long term, in a technological environment which is definitely different from its original environment.’<sup>1</sup>

For years the pressing need for planning common and definitive strategies has been the source of major concern and repeated appeals by archivists and librarians the world throughout. Lately, with the recent heady progress in the adoption of digitisation even in institutional and public administration settings, this has become a primary need and a real emergency. There is in fact a growing awareness that if we continue with the current intuitional and scientific confusion and indifference, the legacy of knowledge we will leave future generations regarding the beginning of this millennium will be next to nothing.

It is worth recalling the extent and the significance of the problem within the scientific community with a summary of some of the most important theoretical contributions published on the subject in recent years.

---

<sup>1</sup> «L'insieme delle attività e degli strumenti che assicurano che i documenti informatici siano mantenuti accessibili, utilizzabili (leggibili e intelligibili) e autentici (univocamente identificabili e integri) nel medio e nel lungo periodo, in un ambiente tecnologico certamente diverso da quello originario.» M Guercio, *La conservazione digitale nello scenario europeo e internazionale. Principi, metodi, progetti*, Rome, November 2003, p. 2 (url: <http://eprints.erpanet.org/archive/00000064/01/e-book.pdf>).

It is readily apparent that the studies, proposals and speculations put forward by renowned scholars and researchers such as Mariella Guercio, Perluigi Feliciati, Stefano Allegrezza and Paolo Franzese, just to mention a few, are all bound by a common thread: they note with clear and incontrovertible argumentation and documentation that in contrast to the enormous flurry and mobility at the international level, the problem of preservation in Italy is still unresolved because it is ignored in its entire extent. The debate and fragmentary projects which are underway are suffering under the conditions of complete backwardness due to the incapacity and, worse still, the unwillingness of the political institutions and universities to take on the task<sup>2</sup>. Not to mention the perennial shortage of financial resources which are made available for studies and projects in the sector.

Taking it for granted that we all consider the proposals and the considerations outlined in the abovementioned studies of vital interest, I would now like to briefly outline some considerations made on the topic from 2008 onwards which emerged following the setting up of a project for the creation of a multimedia library specialised in ICT research for archival and library cultural resources<sup>3</sup> and the subsequent setting up of a spin-off activity which I currently head in the ICT sector for cultural heritage, with specific reference to archival and library cultural resources.

### 3 Questions without Answers

As I stated earlier, the main problem which seems to plague digital preservation in Italy is that political and university circles in this country are not interested in committing human let alone financial resources towards serious cooperation in international initiatives promoting digital preservation, or when they do show an interest money is wasted without producing results worthy of mention, or the sum production is a total failure<sup>4</sup>.

---

<sup>2</sup> In particular, see the following: M. Guercio, *Archivi digitali e conservazione a lungo termine. Un quadro di sintesi sulle strategie internazionali e nazionali*, Archiexpo, 12-15 December 2006 (url: <http://ebookbrowse.com/archiexpo-guercio-ppt-d26376049>); P. Franzese, *Archiviazione e conservazione delle risorse digitali. Les archives électroniques. Manuel pratique* edited by the Directorate of the Archives of France (February 2002), Rome, October 2006 (url: [http://media.regesta.com/dm\\_0/ANAI/.../000/.../ANAI.000.0113.0012.pdf](http://media.regesta.com/dm_0/ANAI/.../000/.../ANAI.000.0113.0012.pdf)); S. Allegrezza, *Informatica di base. Conoscere e comprendere le risorse digitali nella società dell'informazione*, Edizioni SIMPLE, Macerata, 2009; P. Feliciati, *Il nuovo teatro della memoria. Informatica e beni culturali in Italia, tra strumentalità e sinergie*, «Il Capitale culturale. Studies on the Value of Cultural Heritage», vol. 1, 2010, p. 83-104.

<sup>3</sup> The “Unknown Heritage” project, together with the Unknown Heritage workshop of which I am currently the Chief Scientist.

<sup>4</sup> P. Feliciati, p. 86: ‘This instrumental relationship [between information processing techniques and activities related to cultural heritage], due to haste, limited competence or lack of vision which goes beyond the use of extensive resources to obtain – at the best of times – special effects which last no more than the length of a legislative period, has to date only widened the gap rather than produce synergies’. He adds shortly thereafter (p. 88): ‘Except for some sporadic examples of excellence, the rather disapproved “cultural deposits” are noteworthy for the waste of resources in projects which are ad hoc, isolated and without any worthwhile duration or real utility regarding the digital objects obtained ... Moreover, they are all projects which are cited in the literature as poor examples of long-term preservation of digital resources, since with only a few exceptions most of the data which were gathered at such a high price are by now totally lost or useless’.

Another reason which has been put forward is the lack of professionally trained personnel able to take up the arduous challenge, due to severe inadequacies in the national university system. Still today these inadequacies have not been resolved<sup>5</sup>.

Nonetheless, while recognising the clear distinction currently present between digital humanities in Italy and abroad, it can be said that there has been increasing involvement at the international level of archiving and library circles in studies and research on information science and digital libraries, all the while noting the lamentable backwardness characterising the Italian scene with respect to international scenarios, even though in recent years a more solid foundation seems to have been laid down<sup>6</sup>.

On the backdrop of this scenario, when the pilot project “Unknown Heritage” was launched in 2008 the first problem we faced regarded the definition of criteria able to guarantee an acceptable percentage of probability that the databases planned as the output of the project would be usable on the web for at least a five-year period after their creation.

After several months spent evaluating the state of research regarding the market sector particularly in Italy, we were forced to accept the reality which had been so well described in the studies mentioned above: approximation and fragmentation of the scientific research, inhomogeneity in the best practices theorised even at the international level, products and services touted as innovative and revolutionary which after closer examination proved to be totally incapable of satisfying the needs of university project, and so on.

In short, the problem was still there and remains unresolved still today. Nonetheless, I continue to surf the web on a daily basis, consulting specific bibliographies in search of signs foreshadowing the decisive momentum needed to achieve possible solutions.

At any rate, despite our perplexity and awareness of the risks we were up against, the project saw the creation of two databases of equal content but different software architecture which are both usable today on the web without them needing to be updated<sup>7</sup>.

The problem of preservation has however become a pressing issue for me, in that as stated above the project gave rise to a university spin-off D.A.BI.MUS. s.r.l. – Digitalizzazione di Archivi Biblioteche e MUSEi (Digitisation of Library and Museum Archives), the activities of which are digital ICT for cultural resources, with specific reference to archival, library and museum heritage.

Between 2010 and 2011, the spin-off in fact planned and created an innovative application which is currently pending patent. The application is a digital recognition

---

<sup>5</sup> P. Feliciati, p. 89, 90.

<sup>6</sup> P. Feliciati, p. 86, 90.

<sup>7</sup> On the pilot project “Unknown Heritage”, see N. Barbuti, Valorizzare tutelando. Il Laboratorio multimediale e la banca dati digitale “Patrimoni Sconosciuti” dell’Università di Bari, «Biblioteche Oggi», No. 3, April 2011, pp. 38-44. Four years down the track the two databases are still visible and totally accessible at the following URLs:

<http://digilibrary.patrimonisconosciuti.uniba.it> e

<http://virtualibrary.patrimonisconosciuti.uniba.it>

suite, which includes functions of Intelligent Character Recognition, Intelligent Word Recognition, Optical Character Recognition and Graphic Pattern. The application is capable of operating with a high level of efficacy and efficiency on the basis of digital data of antique printed documents, manuscripts and books<sup>8</sup>.

The main difficulties which arose during the planning phase of the system and which were made apparent by various parties during the presentation of the suite include the question of its spatiotemporal duration and the need for multiplatform functioning. We are currently working towards achieving these two objectives: the functions of the suite, which currently are only available for Windows, will be extended for Unix/Linux and MacIntosh platforms, and the algorithm will be developed so that it functions on all image formats currently in use, but above all on the Flexible Image Transport System (FITS) format.

The choice of insisting on this image format as a standard for the development of both the university research and the products of the spin-off is well grounded. FITS has in fact been in use at NASA for almost 50 years as an image format for astronomical photographs. Indeed, it presents all of the necessary characteristics to guarantee the base digital data will enjoy the maximum duration and usability over time without too many risks of destruction and without excessive costs for maintenance or updating.

FITS is a non-proprietary image format which we believe, thanks to its characteristics of usability and portability in time and space, could validly constitute a real starting point for developing strategies aimed at definitively resolving the problem of the preservation of digital originals and the metadata associated with them, and therefore their safety and consultability over time.

Some might argue that it is uncertain whether this image format, which is as I mentioned commonly used for astronomical images, is similarly valid for the reproduction of other materials and in particular cultural resources. We counter that the Vatican Library, after years of tests and checks, was the first in the world to adopt FITS as its main format for the project of digital reproduction of its immense manuscript collection. Based on this initiative, on 5 July 2012, during the EWASS 2012 Conference held at the Pontificia Università Lateranense, the Vatican Library organised a special session entitled *Long-term preservation... from the stars? File format assessment and technical issues in preservation projects for cultural resources*, in which the excellent results were presented. Indeed the Vatican Library has rightly become a standard bearer with its adoption of FITS as an image format shared at the ecumenical level.

Of course, it remains to be seen whether FITS is a format which can also be used for the production of native digital document. Nonetheless, the initiative of the Vatican Library marks a watershed and a significant step forward in the definition of the strategies for digital preservation, and it is worth undertaking serious research to ensure that this beginning does not remain, as often happens, an isolated phenomenon.

---

<sup>8</sup> On the function of graphic pattern, implemented by the spin-off in the setting of a project currently underway in cooperation with the Vatican Library, the following paper is currently undergoing revision: N. Barbuti, T. Caldarola, *Graphic Matching in Historical Manuscripts*.

It is worth examining at this point the concept of original mentioned above in that, when tackling the problem according to an archival/library science approach, the question is not so much one of the preservation of the image as an end in itself, but rather the preservation of the originals, which is a pivotal, inescapable and irreplaceable concept in the doctrine mentioned above.

Indeed, the distinction between original and copy, which is well defined for analogic objects, is extraordinarily subtle for digital resources. From the moment of its creation an electronic or digital document undergoes rapid modifications which irreversibly alter its original structure. From the moment of its creation to the moment of its publication the so-called 'definitive' document ends up being the copy of various and subsequent copies germinating from the original, which has been inevitably and irreversibly lost along with its subsequently revised versions. The document which reaches the end user is only the final interface, and nothing is left of its construction and the original which preceded it in the first phase of its creation. Nonetheless, the definitive document is by definition considered to be the original, and as such it is archived on a stable support and preserved so that consultation and manipulation do not compromise its integrity and possibility of survival over time.

We are all well aware that there are various causes of irreversible destruction of an electronic or digital document/archive which have a decidedly frequent incidence<sup>9</sup>.

To avoid the problems due to the different forms of obsolescence which could compromise their survival over time, the base data need to undergo a process of periodic updating<sup>10</sup>.

Now, reflecting for a moment on this necessary procedure, the following question comes to mind: every time that a digital document undergoes updating, what happens to the original? Is the document created by the update considered the original, or is it a copy of the previous document which nonetheless maintains the security data and the metadata unchanged? And above all, are the metadata and the security data truly preserved in their entirety and perfectly identical in the new document? Several doubts remain, and we are patiently awaiting answers which are clear, thorough and above all definitive.

In order to clarify what we believe to be the extent of the problem, let us now briefly examine the change in the transposition of the collective memory of human endeavour onto a transferable support, and how this change has today reached a point of no return, which is slipping through our fingers unnoticed, or at least unmentioned.

---

<sup>9</sup> We are referring to the three major problems which create difficulties for the preservation of digital content: obsolescence of hardware and software technology, obsolescence of the support and obsolescence of the formats. In addition there can be accidental causes, such as prolonged exposure to heat, or those due to inadequate environmental preservation. See in this respect Allegrezza, p.2.

<sup>10</sup> See S. Allegrezza, p. 4: 'Over the last fifteen-twenty years the problems of digital preservation have been tackled from numerous viewpoints and a variety of preservation strategies have been put forward suggesting a solution. The main ones include: output to analogue media; technology preservation; emulation; refreshing and migration; and digital archaeology'.

For thousands of years mankind has been an avid inventor of new methods with which it has entrusted its evolutionary history so that it can be passed on to future generations. And for thousands of years these methods have had the common characteristic of tremendous stability over time and space, characterised as they are by a rate of decay which is either virtually non-existent (stone, pliable materials) or extremely slow (papyrus, parchment, paper). All of these materials have proven their ability to resist natural calamities and the disasters of warfare and thus allow us still today to study their content and acquire even greater knowledge about who we are and where we come from.

With the advent of the analogic age between the 19th and the 20th centuries (wax, charcoal, vinyl, photographic film, audio and video), the ability of the materials to last over time has noticeably decreased with respect to the past, all the while maintaining still acceptable levels.

However, it has been the advent of the electronic and then the digital age, characterised by recording processes which are unstable and volatile by definition, which has marked the beginning of an irreversible disappearance of a significant quantity of contemporary collective memory.

Consider for example that while I am writing this paper, millions of billions of data created by mankind throughout the world are disappearing into the ether, taking with them the collective memory that they contain. Consider for example that this very paper has already been rewritten numerous times in many of its parts, its many errors and typos corrected in real time, such that in the end it has become a final and very different copy of what would have been the original project, if instead I had decided to first write it down with a pen and paper, and then in a word-processed document, and lastly made a comparison of the two products.

Electronic mail and on-line communication, chat and social networks have definitively blown away all practice of handwritten interpersonal communication, a heritage which for centuries has allowed us to understand the life and culture of those who with their lives and their works have written our history.

With the word processed document it is impossible to identify the path running from the gestation of any thought through to its birth, elaboration and publication. The digitisation of the public administration will make it impossible in a short space of time to reconstruct contemporary social, economic, health and demographic history. For future generations living in an age of hypertechnology even more volatile than our own, if that is at all possible, such a reconstruction could be useful for understanding their role on this planet.

This is not a post-apocalyptic scenario from some science fiction screenplay. Instead it is the reality which passes daily before our eyes and which for years we have called the technological miracle, without realising that we are passively succumbing to the rapid destruction of contemporary collective memory, a drama in which we are both actors and directors. If we fail to shore up this flood towards oblivion, we run the risk of becoming the historical age without a clearly identifiable past and with no collective memory of the present capable of creating a future: an endless present, the first and only true dark age in the history of humanity. The dark centuries of the Middle Ages will pale in comparison.

Therefore, we feel that the problem of digital preservation is no longer a problem of the capacity of real or virtual hardware space made available for preserving our digital collective memory. Nor is it a problem of distinguishing between an original and a copy, which nonetheless is a question of primary importance in the choice of what needs to be preserved what can be eliminated. It is not even a problem of the obsolescence of magnetic, optical or who knows what other type of support which will appear on the scene in the next five years or so.

The problem of preservation, a problem I have been wrestling with since I became interested in information technology for cultural heritage some ten years ago, is becoming something much larger. Indeed the problem involves society as a whole, in a setting where it appears there is ignorance of the fact that already a large part of what was created in the last 20 years was destroyed at the moment of its creation, and much of what has survived is being destroyed at a tremendous rate. It is a problem of preservation of collective memory in a digital format, and not one of the “simple” – but as we already know very complex – digital preservation. If we fail to carefully focus on this analytical perspective of the problem, we feel that it will be very difficult to develop common strategies capable first of restricting the flow and then of stemming it entirely with appropriate and timely planning.

Moreover, perhaps even as a result of the difficulty of framing the issue in its true extent, in this scenario researchers, scholars and operators active in the various information, cultural and administrative sector, who despite their daily ringing of alarm bells regarding the worrying situation and who apply themselves with a passion and in some cases a high degree of professionalism to provide hypothetical solutions to the problem, in reality seem more interested in justifying their own membership to their respective sectors and the supremacy of each over the others.

By now the scientific papers and studies and the research published on the different problems surrounding digital culture and its preservation are on the daily agenda, published above all by computer scientists and archivists/librarians who have chosen to broaden their knowledge of the new technologies and who have, so to say, lent themselves to computer science. However, an analysis of these studies reveals that the approaches to these problems and the possible solutions are still diametrically opposed. The computer scientists are sunk in their endeavours to develop algorithmic structures or theoretically perfect networks, which nonetheless are often practically unusable and therefore destined to remain contemporary pipe dreams which will soon be forgotten or overtaken by new theories, which will also be perfect and unusable. The cultural scholars are instead hunkered down in their own defensive positions built on the few and certain results they have obtained and their supposed eternal validity, voluntarily unaware that without programmes and projects built on synergies between qualified professionals originating from both sectors, those results will be destined to a life much shorter than what is need as they are the result of fragmentary, inhomogeneous policies often brought about by more of a need to put on a show than to really make cultural resources available to the present and future generations.

The approach to the problems of the preservation of collective memory in a digital format can only be interdisciplinary and must cut across the various cultural, scientific and social forces, and only with a major effort and results which are worthy of note

will it be possible to achieve a clear vision of what is happening, and as a result, provide the planning of preservation policies which are shared, certain, effective, efficient and long lasting in time and space.

In conclusion, let us finish with this bitter sweet literary digression:

‘From the right the sound of a trumpet is heard,  
from the left [still no] sound is heard in reply.’<sup>11</sup>

And yet, faithfully we shall wait, and in the meantime we will continue do research, to study, to compare and to grow.

---

<sup>11</sup> Translator’s note: Alessandro Manzoni, *Il Conte di Carmagnola*, «S’ode a destra uno squillo di tromba / a sinistra [ancora non, nda] risponde uno squillo», with the text ‘ancora non’ inserted in square brackets by the author. The verse refers to two armies facing each other on the battlefield. They appear to mirror each other, and it is noted that neither is an invading force – the question of brotherhood is raised as the armies are composed of Venetians and Milanese.



# Supporting Tabular Data Characterization in a Large Scale Data Infrastructure by Lexical Matching Techniques

Leonardo Candela, Gianpaolo Coro, and Pasquale Pagano

Istituto di Scienza e Tecnologie dell'Informazione "Alessandro Faedo"  
Consiglio Nazionale delle Ricerche  
Via G. Moruzzi, 1 – 56124, Pisa, Italy  
{candela,coro,pagano}@isti.cnr.it

**Abstract.** Digital Libraries continue to evolve towards research environments supporting access and management of multiform Information Objects spread across multiple data sources and organizational domains. This evolution has introduced the need to deal with Information Objects having traits different from those characterizing Digital Libraries at their early stages and to revise the services supporting their management. Tabular data represent a class of Information Objects that require to be efficiently managed because of their core role in many eScience scenarios. This paper discusses the tabular data characterization problem, i.e., the problem of identifying the reference dataset of any column of the dataset. In particular, the paper presents an approach based on lexical matching techniques to support users during the data curation phase by providing them with a ranked list of reference datasets suitable for a dataset column.

**Keywords:** tabular data management, data curation, large-scale data infrastructure, lexical similarity.

## 1 Introduction

Digital Libraries have evolved a lot during the last twenty years while maintaining and further strengthening their central role in knowledge sharing [6]. Digital Libraries are revolutionizing the whole knowledge management lifecycle. They are no longer perceived as a means to discover cultural heritage only, rather are nowadays conceived as innovative, dynamic, and ubiquitous research supporting environments. In such environments *communities of practice* [15,25] are expected to be able, through their Web browsers, to seamlessly access and exploit data, services, and processing resources managed by diverse systems in separate administration domains.

This evolution continues to enlarge the domains Digital Libraries are called to serve that presently include *eScience*, *cultural heritage*, and others [5,21,29,12,24]. Current Digital Library developers are called to develop complex systems that

have to give solutions to “traditional” issues, e.g., existing data providers federation, distributed retrieval, and long-term preservation, as well as “new” issues, e.g., social network models, large-scale computing, and micro information. Furthermore, they have to face scaled-up versions of the above issues with respect to various axes, e.g., number and variety of actors to be served, size and variety of content to be managed, and diversity of systems and technologies to be integrated. Very often the content they are requested to manage falls under the “*data*” category and their implementation actually requires the realization of Data Infrastructures.

The term “data” itself, although very common, is difficult to define since it may be given different meanings, both in the digital and in the real world. Actually, the act of recognising or understanding that “something” – e.g., observations, statistics, artefacts, records – constitutes data is an intellectual activity that is usually driven by a certain goal. Data is collected for many purposes, via different approaches and very often it is difficult to interpret once exploited in contexts other than its initial one [3,4]. Digital Libraries are called to manage data ranging from traditional research outputs, mainly papers and experimental data, to living reports [7,5], executable research papers [10,19], and scientific workflows [20]. Very often such data fall into the category of “*big data*” [23], i.e., data characterised by (i) *volume*, i.e., its dimension in terms of bytes is huge; (ii) *velocity*, i.e., its speed requirements for collecting, processing and using is demanding; and (iii) *variety*, i.e., its heterogeneity in terms of data types to be managed and data sources to be merged is high.

This paper discusses one of the problems arising when dealing with *tabular data*<sup>1</sup> management where management needs (i) to support collaboration among multiple users and organizations; (ii) to appeal to a broad audience of users who are not technically skilled; and (iii) to guarantee data completeness and correctness as to enable effective data analysis; i.e., to solve the problem of identifying, verifying and associating the actual controlled vocabularies that might have been used by the data provider while producing the dataset. Tabular data are mainly stored in CSV (Comma Separated Values) files where little or no emphasis is posed on representing and standardizing the characterization of the single columns they consist of. However, knowing the “type” of values a column is expected to contain (*controlled vocabulary*, *code list* or *reference dataset* rather than basic types such as string or integer) is a fundamental aspect when datasets have to be effectively managed for, e.g., certification of compliance, comparison, integration and analysis. To this aim, this paper proposes an approach for supporting an end user during the operations to transform a “*raw dataset*” – i.e., a dataset consisting of its data only – into a “*characterized dataset*” – i.e., a dataset where each column is characterized by the controlled vocabulary from which its values have been selected. The effectiveness of such an approach is discussed in the context of a Data Infrastructure.

---

<sup>1</sup> Tabular data is a very common format for a plethora of data ranging from observations to specimen records, catch statistics, surveys, etc.

The remainder of the paper is organized as follows. Section 2 characterizes the major challenges of the problem identified above. Section 3 describes the proposed approach. Section 4 assesses the effectiveness of the proposed approach. Finally, Section 5 concludes the paper and summarizes its results.

## 2 The Tabular Data Characterization Problem

Data-intensive science as well as approaches expecting to rely on data require three basic activities: data *capture*, *curation*, and *analysis*. In these scenarios, data come in all scales and shapes covering: large international experiments; cross-laboratory, single-laboratory, and individual observations; and also individuals lives [12].

In these settings it is fundamental to equip collected datasets with additional information aiming at characterizing each dataset and making it possible to interpret the dataset even in contexts other than its initial one. This additional information may range from *bibliographic*-oriented metadata to *provenance*-, *coverage*-, *certification*-oriented metadata. Enriched and standardized datasets are, in fact, simpler to be managed and allow for exploiting more predefined functionalities as to get high performances on analysis and processing.

Tabular data represent a very common format for many datasets in many different scenarios, e.g., statistical data, surveys, observations. A fundamental piece of information that should equip tabular data is the one characterizing the “*data type*” of any column a dataset contains. However, the actual notion of data type goes well beyond the expected ones like string or integer. In fact, the compilation of datasets commonly relies on existing *controlled vocabularies*, *code lists* and *reference datasets* [2]. For instance, in compiling a dataset on catch statistics or specimen records it is worth to refer to reference datasets for species names and zones. Such reference datasets usually contain a complete record for each of the instances the reference dataset is about, as well as links with other reference datasets. By linking a dataset with the reference datasets its values come from, the actual information contained in the dataset is multiplied. The motivations of this are similar to those of *Linked Data* [1].

Although reference datasets are used or alluded during datasets capture phase, any information about them is usually discarded when the tabular dataset is stored in a CSV file for management purposes. Moreover, this capture phase is usually performed in very diverse technological and organizational settings, thus leading to a very heterogeneous set of tabular datasets. Because of this, it is expected that a curation phase reconciles the “raw dataset” with its “characterized” / “curated” version when the datasets are aggregated in a common information space aiming at promoting their consumption.

Common issues that may arise when a user wants to “curate” a given dataset are the following:

---

<sup>2</sup> In the remainder of the paper the term reference dataset will be used to represent any dataset whose values are recognized instances of the elements the dataset is about, e.g., species, zones, countries.

- The raw dataset contains entries which might be misspelled with respect to the intended reference values;
- The raw dataset contains too many entries to be controlled by hand;
- The reference datasets are too many to be manually searched and then be associated with the dataset under curation;
- Potentially, many reference datasets might be associated to a given dataset (high level of ambiguity).

A complete comparison between a raw dataset and all the reference datasets would need high computational requirements. Moreover, it is not appropriate if a quick (almost real time) response time is expected, as it happens when the user is asking a web application to propose a reference dataset suitable for the dataset she/he is managing.

On the other hand, even a “*greedy*” approach is not so easy to identify because of the issues just discussed, e.g., a simple match between string data cannot be used because it is incapable to overcome the misspelling problems.

In the remainder of the paper, an approach for supporting an end user during the curation phase is proposed. It consists in an “helper” facilitating the identification of the reference datasets that have been actually used while compiling the “raw dataset”. This approach is conceived to be fast and effective with respect to the issues discussed above.

### 3 An Approach for Tabular Data Characterization

The proposed approach is based on two algorithms: *(i)* a revised version of the Minimum Edit Distance (MED) and *(ii)* a constant complexity ranking procedure aiming at proposing a ranked list of suitable reference datasets given a column of a tabular dataset.

The Minimum Edit Distance (or Levenshtein Distance) algorithm was firstly introduced in [16]. It is a metric for measuring the amount of difference between two character sequences. It is defined as the minimum number of edits needed to transform one string into the other, the allowed edit operations being insertion, deletion, or substitution of a single character. The algorithm is based on a dynamic programming procedure introduced in [18] and has a computational complexity that is linear with respect to the product of the length (number of characters) of the strings to be compared. However, there exist several approaches for computing the “distance” between two strings or sequences of symbols. Some well known *similarity metrics*, i.e., measures for similarity or dissimilarity between two text strings for approximate matching or comparison, include: *(i)* the Hamming distance [11], which calculates the number of positions at which the corresponding symbols are different; *(ii)* the Needleman-Wunsch distance [18], which is used in bioinformatics to align protein or nucleotide sequences; and *(iii)* the Smith-Waterman distance [22], which is a variation of the previous one and performs local sequences alignment. Other techniques are used in various domains ranging from biology to phonetics, e.g., *(i)* the Jaro-Winkler distance [13,26], which is mainly used in the area of duplicates detection; *(ii)* the

Block or L1 distance [14], which introduces a new geometry for distance calculation, where Euclidean geometry is replaced by a new metric in which the distance between two points is the sum of the absolute differences of their coordinates; and (iii) the Soundex distance [17], which is a phonetic algorithm for indexing names by sound, as pronounced in English. Among the existing algorithms, we selected the MED one as it is the most common method for string comparison, its implementation is straightforward and it fits well with the characteristics of the proposed approach.

The constant complexity ranking procedure proposed is called *Lexical Guesser*. This is an approach defined by relying on the edit distance, which uses the lexical similarity scores for limiting the computational extent of the ranking procedure of a given column of a dataset. From that point on, a given column of a dataset which has been selected for curation purposes is called “*target dataset*”. The Lexical Guesser uses similarities, instead of exact matching, in order to avoid to perform all the comparisons between the target dataset entries and all the entries of all the recognized reference datasets. The basic underlying ideas are:

- if the target dataset contains entries which are misspelled, errors can be recovered by using MED (actually, a revised version of it);
- if the target dataset is syntactically correct, then the computation can be limited by assuming that by picking some random chunks (samples) from the *right* reference dataset, these chunks will probably be lexically similar to the target dataset. For instance, a target dataset containing entries like ‘North Atlantic Ocean’, ‘South Pacific Ocean’, etc. would always get a non-minimal score when compared to the ‘Oceans English Names’ reference dataset because the latter also contains entries like ‘Indian Ocean’ or ‘North Pacific Ocean’ which share some lexical similarities with the target dataset entries. It is assumed that the recall of the search for the target dataset can include all those reference datasets presenting lexical similarities (over a certain *threshold*);
- the proposed approach is expected to be an helper for an activity that should remain semiautomatic, i.e., the algorithm reduces the search space of the possible reference data, while the final choice about the reference dataset to use is a duty of the user.

According to the above premises, the MED algorithm was modified and then incorporated into a ranking procedure realising the *Lexical Guesser*.

Actually, the MED algorithm has been enriched with a set of check rules and parameters aiming at enhancing its performances for the overall classification process. From a preliminary analysis, it has been noticed that the standard MED algorithm is not sufficient for calculating distances in the target scenarios. Some boosting rules have been added to raise or lower the scores in some cases.

The distance between two strings  $x$  and  $y$  is calculated as follows:

$$d(x, y) = \begin{cases} 0 & \text{if } \frac{\max l_n}{\min l_n} > 1.5 \\ \min\left(\frac{\min l_n}{\max l_n} * 1.5, 0.9\right) & \text{if } \text{contains}(x, y) \vee \text{contains}(y, x) \\ 1 - \frac{\text{MED}(x, y)}{\max l_n} & \text{otherwise} \end{cases} \quad (1)$$

where:

- $maxln = \max(\text{length}(x), \text{length}(y))$ ;
- $minln = \min(\text{length}(x), \text{length}(y))$ .

The constant values in the formula above are the result of an experimental activity. The limitation to 0.9 for the value of  $d(x, y)$  when a string contains another one is a penalty score which aims to lower the distance value in the cases when strings are really close but not equal.

The ranking procedure consists in computing a similarity score  $S(T, R_i)$  between the *target dataset*  $T$ , i.e., the values of a given dataset column, and every recognized reference dataset  $R_i$  as a product of three factors, namely (i) a *distance score*  $D(T, R_i)$ , (ii) a *coverage score*  $C(T, R_i)$ , and (iii) a *weight score*  $W(R_i)$ , by actually relying on samples of both the datasets, i.e.,  $\overline{T}$  and  $\overline{R_i}$ . A score  $\alpha$  is computed to estimate the representativeness of the sample  $\overline{R_i}$  as follows:  $\alpha = |\overline{R_i}|/|R_i|$ .

Given a *target dataset*  $T$ , for each reference dataset  $R_i$  the similarity score  $S(T, R_i)$  is calculated by the following formula:

$$S(T, R_i) = D(T, R_i) * C(T, R_i) * W(R_i) \quad (2)$$

where

1. the *distance score*  $D(T, R_i)$  is computed as the average distance between all the pairs of the selected samples  $\{(t_k, r_{i_j}) | t_k \in \overline{T} \wedge r_{i_j} \in \overline{R_i}\}$  where the distance is greater than an “acceptance threshold”  $\tau$  as follows:

$$D(T, R_i) = \frac{\sum \{d(t_k, r_{i_j}) | d(t_k, r_{i_j}) > \tau\}}{|\{(t_k, r_{i_j}) | d(t_k, r_{i_j}) > \tau\}|} \quad (3)$$

2. the *coverage score*  $C(T, R_i)$  is computed by multiplying the  $\alpha$  score aiming at indicating the representativeness of the sample  $\overline{R_i}$  by a factor aiming at indicating the similarity between  $\overline{R_i}$  and  $\overline{T}$  as follows:

$$C(T, R_i) = \alpha * \frac{S}{|\overline{R_i}|} = \frac{S}{|R_i|} \quad (4)$$

where  $S = |\{r_{i_j} | r_{i_j} \in \overline{R_i} \wedge \exists t_k | t_k \in \overline{T} \wedge d(r_{i_j}, t_k) > \tau\}|$

3. the *weight score*  $W(R_i)$  is computed (i) by comparing the “size” of  $R_i$  with respect to the size of the whole set of recognized datasets and (ii) mitigating the impact of “big” dataset via logarithmic transformation as follows:

$$W(R_i) = \begin{cases} \frac{|R_i|}{\sum |R_j|} * 100 & \text{if } \frac{|R_i|}{\sum |R_j|} * 100 \leq 1 \\ \log\left(\frac{|R_i|}{\sum |R_j|} * 100\right) & \text{otherwise} \end{cases} \quad (5)$$

It is evident that the higher is each factor value, the higher the similarity score. This means that a very high score could imply a good overall similarity among

the single entries but even that the elements in  $T$  cover a big percentage of the  $R_i$  set.

Given a *target dataset*  $T$ , the list of recommended reference datasets is produced by sorting the set of reference dataset according to the values of the similarity score  $S(T, R_i)$  and pruning those whose score differs from the top-ranked element in the list (i.e., the best score) for more than a given customizable threshold (Maximum Difference from Best Threshold or MDBT).

Overall, the complexity of the procedure depends from the number of string comparisons to be performed. If  $k$  is the number of reference datasets recognized and  $|\overline{T}| = m$  and  $\forall i, |\overline{R_i}| = n$ , then the overall number of comparisons to be performed is  $k * m * n$ . However, because of its characteristics, the proposed approach is incline for parallelization both with respect to the reference datasets (every  $S(T, R_i)$  can be calculated by an independent process) as well as with respect to the single reference dataset (independent processes can be used to calculate factors of the same  $S(T, R_i)$ ).

The procedure can then be tuned in order to get results in an acceptable time, e.g., by establishing proper values for  $n$  and  $m$  as well as for the thresholds and the rest of parameters discussed above. The higher is the number of comparisons, the higher will be the complexity of the calculation as well as the accuracy. These aspects are discussed in the next Section.

## 4 Experiment and Results

The experiment we performed to validate the approach is based on tabular datasets and reference datasets expected to be managed in the context of the large scale data infrastructure implemented by D4Science and D4Science-II projects [8]. In particular, the settings are those resulting from an environment aiming at providing fisheries statisticians with a set of tools to manage tabular data on catch statistics. Tabular data usually are time series coming from observations about fishery periodic catches in terms of quantities and costs. When an user wants to manage a new time series, in order to use all the facilities offered by the environments for time series analysis and consumption, she/he has to curate such dataset by recognizing the reference datasets it exploits. Such operation involves the correction of misspelled entries, the identification of the data types for the columns and a validation of the coherence of the dataset contents. In this phase, the user is expected to rely on facilities helping the identification of the most suitable reference datasets. These facilities are based on the Lexical Guesser.

In this scenario, the set of recognized reference datasets is about information on marine species, e.g., species names, geographical areas, economic zones. It consists in 326 reference datasets, containing from 5 to 39,000 elements. These reference datasets can be classified as follows:

- *no overlap* – reference datasets that are *disjoint* from each other;
- *medium overlap* – reference datasets that present a *medium degree of intersection* with other ones, i.e., 20-50% of their entries overlap with entries

in at least another reference dataset (e.g., FAO area names, the ocean and sub-ocean names and the geographical names);

- *high overlap* – reference datasets that have a *large degree of intersection* with others, i.e., 80-90% of their entries overlap with entries in at least another reference dataset (e.g., species names coming from different species databases).

Each experiment reported in the remainder of this paper was executed by using 50 different target datasets for 20 times per input. The average score is reported in the tables.

In order to test the performances of the proposed approach, the following well known measures have been exploited:

$$Accuracy = \frac{TrueNegatives + TruePositives}{TotalNumberofReferenceDatasets} \quad (6)$$

$$Precision = \frac{TruePositives}{TruePositives + FalsePositives} \quad (7)$$

$$Recall = \frac{TruePositives}{TruePositives + FalseNegatives} \quad (8)$$

$$F1 = 2 * \frac{Precision * Recall}{Precision + Recall} \quad (9)$$

where:

- *True Negatives* indicates the number of classifications which are correctly classified as *not suitable*,
- *True Positives* indicates the number of reported classifications which are really suitable for labeling an unknown target dataset;
- *False Negatives* and *False Positives* are defined by complement of the above.

The experiment configuration was set as follows:

- one sample of 25 elements was taken from a target dataset;
- a sample of 625 elements was taken from each reference datasets;
- the threshold for pruning the ranked list of  $S(T, R_i)$ , i.e., MDBT, was set to 30%;

The performance of the following tree approaches have been assessed:

- *Lexical Guesser* – i.e., the approach proposed in Sec. 3, based on the ranking of similarity scores  $S(T, R_i)$  by formula 2;
- *Simple Matcher - Constant Complexity* – i.e., an approach based on the same ranking procedure (with pruning) where the distance  $d(x, y)$  is based on exact string matching;
- *Simple Matcher - High Complexity* – i.e., an approach based on the ranking of  $S(T, R_i)$  for all the reference datasets where the distance  $d(x, y)$  is based on exact string matching.



**Table 1.** Results on a column of a dataset that exactly matches a reference dataset (results are expressed in percentages)

	Lexical Guesser	Simple Matcher Constant Complexity	Simple Matcher High Complexity
No overlap			
Accuracy	100	100	100
Precision	100	100	100
Recall	100	100	100
F1	100	100	100
Medium overlap (20%-50%)			
Accuracy	99.18	99.18	99.40
Precision	28.89	28.89	38.89
Recall	100	100	100
F1	44.44	44.44	44.44
High overlap (80%-90%)			
Accuracy	99.77	99.85	99.92
Precision	70.83	75	87.50
Recall	100	100	100
F1	79.17	83.33	91.67

Table 1 reports the results for target datasets that match exactly one reference dataset. In the case of *no overlap*, all the approaches get a 100% of accuracy. The task is quite trivial and the ranking procedure with pruning does not influence the performances. In the case of *medium overlap*, the ‘Simple Matcher - High Complexity’ approach is expected to perform better than the others, while errors are experienced with the ‘Simple Matcher - Constant Complexity’. The ‘Lexical Guesser’ performs as good as the ‘Simple Matcher - High Complexity’, thus the flexibility of  $d(x, y)$  does not help in this case. In the case of *medium overlap*, the Lexical Guesser performs worst than the others, however the performances are still acceptable for the application scopes.

Table 2 presents the performances when the experiment focuses on target datasets containing misspelled entries and entries that do not occur at all in reference datasets for the 50 to 100% of their entries. The ‘Lexical Guesser’ always outperforms the other two approaches. Moreover, the ‘Simple Matcher - Constant Complexity’ introduces errors. Performances are appreciable both in terms of accuracy and recall, which means that the approach is always able to return the right reference datasets to the user. The recall of approaches based on simple matching is always lower than that of the Lexical Guesser because in some cases the target dataset may be ambiguous, so that more than one reference dataset is suitable for it. In this case the choice necessarily is on the user’s side, as she/he only knows the real nature of her/his data. A simple match tends to find few columns, while the proposed approach uses the flexibility of the comparisons in order to propose more reference datasets. The precision score indicates that in presence of either low or high ambiguity, the Lexical Guesser is able to extract the correct information, while with medium ambiguity, the

**Table 2.** Results on a column of a dataset which does not match exactly reference datasets (results are expressed in percentages)

	Lexical Guesser	Simple Matcher Constant Complexity	Simple Matcher High Complexity
No Superpositions			
Accuracy	100	99.69	94.89
Precision	100	66.67	35.29
Recall	100	44.44	55.56
F1	100	53.33	30.37
Medium Superpositions (20%-50%)			
Accuracy	99.54	99.39	99.54
Precision	58.33	100	100
Recall	100	45.83	62.50
F1	73.33	60	70
High Superpositions (80%-90%)			
Accuracy	99.54	99.23	99.23
Precision	100	50	50
Recall	50	16.67	16.67
F1	65	25	25

statistical nature of the algorithm begins to be evident. This happens because for some samples lexical similarities are found, while for others they are not retrieved. As for the F1 measure, it can be noted that it gives an estimation of the overall functioning, and the value for the Lexical Matcher is always higher.

## 5 Conclusion

The evolution of Digital Libraries calls for innovative, dynamic, and ubiquitous research supporting environments where communities of practice can seamlessly access data, software, and processing resources managed by diverse systems in separate administration domains through their Web browsers. In these environments data are multiform and their management demand for new methods.

This paper has discussed one of the problems arising when dealing with tabular data management where management occurs in scenarios characterized by these needs: (i) supporting collaboration among multiple users and organizations; (ii) appealing to a broad audience of users who are not technically skilled; and (iii) guaranteeing data completeness and correctness as to enable effective data analysis; i.e., giving solution to the problem of identifying, verifying and associating the actual reference datasets that might have been used by the data provider while producing the dataset.

It has been proposed an approach supporting an end user during the massaging of a “raw dataset” to transform it into a “characterized dataset” defined by associating the proper reference datasets that might have been used while capturing the data. This approach is based on (i) a similarity measure aiming

at estimating the similarity among the entries of the target dataset and the entries of the reference dataset by overcoming misspelling issues and (ii) a ranking approach appropriate for a real time use and aiming at providing the end user with a sorted list of reference datasets suitable for a given target dataset.

The experimental results show that the proposed approach actually outperforms other approaches in presence of misspelled entries, even if it loses in performances with respect to an approach based on exact string matching when user's data completely agree with some of the reference dataset.

**Acknowledgments.** The work reported has been partially supported by the *D4Science-II* project (FP7 of the European Commission, INFRA-2008-1.2.2, Contract No. 239019) and the *iMarine* project (FP7 of the European Commission, FP7-INFRASTRUCTURES-2011-2, Contract No. 283644). The authors would like to thank M. B. Baldacci (ISTI-CNR) for many helpful comments.

## References

1. Bizer, C., Heath, T., Berners-Lee, T.: Linked data - the story so far. *International Journal on Semantic Web & Information Systems* 5(3), 1–22 (2009)
2. Blanke, T., Candela, L., Hedges, M., Priddy, M., Simeoni, F.: Deploying general-purpose virtual research environments for humanities research. *Philosophical Transactions of the Royal Society A* 368, 3813–3828 (2010)
3. Borgman, C.: Research data: Who will share what, with whom, when, and why? In: *China-North America Library Conference*, Beijing (2010)
4. Borgman, C.: The Conundrum of Sharing Research Data. *Journal of the American Society for Information Science and Technology*, 1–40 (2011)
5. Candela, L., Akal, F., Avancini, H., Castelli, D., Fusco, L., Guidetti, V., Langguth, C., Manzi, A., Pagano, P., Schuldt, H., Simi, M., Springmann, M., Voicu, L.: DILIGENT: integrating Digital Library and Grid Technologies for a new Earth Observation Research Infrastructure. *International Journal on Digital Libraries* 7(1-2), 59–80 (2007)
6. Candela, L., Castelli, D., Pagano, P.: History, Evolution and Impact of Digital Libraries. In: Iglezakis, I., Synodinou, T.-E., Kapidakis, S. (eds.) *E-Publishing and Digital Libraries: Legal and Organizational Issues*, ch. 1, pp. 1–30. IGI Global (2011)
7. Candela, L., Castelli, D., Pagano, P., Simi, M.: From Heterogeneous Information Spaces to Virtual Documents. In: Fox, E.A., Neuhold, E.J., Premssmit, P., Wu-wongse, V. (eds.) *ICADL 2005*. LNCS, vol. 3815, pp. 11–22. Springer, Heidelberg (2005)
8. Castelli, D.: D4Science-II - An e-Infrastructure Ecosystem for Science. *ERICIM News* 79, 9 (2009)
9. Crane, G., Babeu, A., Bamman, D.: eScience and the humanities. *International Journal on Digital Libraries* 7(1-2), 117–122 (2007)
10. Gorp, P.V., Mazanek, S.: SHARE: a web portal for creating and sharing executable research papers. *Procedia CS* 4, 589–597 (2011)
11. Hamming, R.W.: Error detecting and error correcting codes. *Bell System Technical Journal* 29(2), 147–160 (1950)

12. Hey, T., Tansley, S., Tolle, K.: *The Fourth Paradigm: Data-Intensive Scientific Discovery*. Microsoft Research (2009)
13. Jaro, M.A.: Advances in record linkage methodology as applied to the 1985 census of tampa florida. *Journal of the American Statistical Society* 84(406), 414–420 (1989)
14. Krause, E.F.: *Taxicab Geometry*. Dover Publications (1987)
15. Lave, J., Wenger: *Situated Learning: Legitimate Peripheral Participation*. Cam (1991)
16. Levenshtein, V.: Binary Codes Capable of Correcting Deletions, Insertions and Reversals. *Soviet Physics Doklady* 10, 707–710 (1966)
17. National Archives and Records Administration. *The Soundex Indexing System* (2007)
18. Needleman, S.B., Wunsch, C.D.: A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology* 48(3), 443–453 (1970)
19. Nowakowski, P., Ciepiela, E., Harezlak, D., Kocot, J., Kasztelnik, M., Bartynski, T., Meizner, J., Dyk, G., Malawski, M.: The collage authoring environment. *Procedia CS* 4, 608–617 (2011)
20. Roure, D.D., Goble, C.A., Stevens, R.: The design and realisation of the my<sub>experiment</sub> virtual research environment for social sharing of workflows. *Future Generation Comp. Syst.* 25(5), 561–567 (2009)
21. Shen, R., Vemuri, N.S., Fan, W., Fox, E.A.: Integration of complex archaeology digital libraries: An ETANA-DL experience. *Information Systems* 33(7-8), 699–723 (2008)
22. Smith, T.F., Waterman, M.S.: Identification of common molecular subsequences. *Journal of Molecular Biology* 147, 195–197 (1981)
23. Stapleton, L.K.: Taming Big Data. *IBM Data Management Magazine* 16(2), 12–18 (2011)
24. Wallis, J.C., Mayernik, M.S., Borgman, C.L., Pepe, A.: Digital libraries for scientific data discovery and reuse: from vision to practical reality. In: *Proceedings of the 10th Annual Joint Conference on Digital Libraries, JCDL 2010*, pp. 333–340. ACM, New York (2010)
25. Wenger, E.: *Communities of Practice: Learning, Meaning and Identity*. Cambridge University Press (1998)
26. Winkler, W.E.: String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage. In: *Proceedings of the Section on Survey Research Methods (American Statistical Association)*, pp. 354–359 (1990)

# Data Interoperability and Curation: The European Film Gateway Experience

Michele Artini, Alessia Bardi, Federico Biagini, Franca Debole, Sandro La Bruzzo, Paolo Manghi, Marko Mikulicic, Pasquale Savino, 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.** Film archives, containing collections of cinema-related digital material, have been created in many European countries. Today, the EC Best Practice Network Project EFG (European Film Gateway) provides a single access point to 59 collections from 19 archives and across 14 European countries, for a total of 640,000 digital objects. This paper illustrates challenges and solutions in the realization of the EFG data infrastructure. These mainly concerned the curation and interoperability issues derived by the need of aggregating metadata from heterogeneous archives (different data models, hence metadata schemas, and exchange formats). EFG designed a common data model for movie information, onto which archives data models can be optimally mapped. It realizes a data infrastructure based on the D-NET software toolkit, capable of dealing with data collection, mapping, cleaning, indexing, and access provision through web portals or standard access protocols. To achieve its objectives EFG has extended D-NET with advanced tools for data curation.

**Keywords:** Data Infrastructure, Aggregation System, Metadata Formats, Data Interoperability, Data Curation, Data Cleansing, Audio Video, D-NET.

## 1 Introduction

Nowadays, many digital film archives are available in Europe, thanks to a significant effort performed in digitizing existing collections of images, videos, and cinema-related material (e.g., audio documents, photographs, posters, drawings, text documents). These archives make their collections available to the community through repository platforms or similar technologies, which support web portals to search, browse, and visualize cinema-related metadata and relative digital objects. Although the information dissemination service they offer is extremely useful, their autonomy still represents a limit to the urgent demand of immediate and global access to information required by today's communities.

The EFG (European Film Gateway) Best Practice Network [1], funded by the European Commission under the eContent*plus* programme [2], provides community

users with a single entry point from which content of several archives can be searched in a uniform fashion, abstracting over their differences and peculiarities. Specifically, EFG delivers a data infrastructure whose aim is to aggregate content from the most prominent European film archives and cinematheques in order to make it available to end users and authorized third-party consumers, including Europeana (European Digital Library) [3]. The project started in September 2008 and was completed in October 2011 and its two-year continuation will kick-off on February 2012. It includes 20 partner institutions from 14 European countries, and today provides direct access to about 640,000 digital objects including films, photos, posters, drawings, and text documents, plus authority files for film works, persons and corporate bodies.

Although film archives contain similar digital movie-related objects, their data models (and relative metadata schemas) may be very different in structure and semantics, as well as their content be subject to errors or be duplicated. In this paper, we describe the solutions to the *data interoperability* and *data curation* challenges faced in EFG in order to deliver a unified, homogeneous, high-quality, and unambiguous European information space of movie metadata.

*Data interoperability.* The EFG infrastructure has adopted a bottom-up approach to data aggregation where interoperability is achieved by (i) defining a common data model and relative metadata schema together with domain specific vocabularies, and (ii) implementing the technology to collect, transform onto the common schema, and harmonize metadata records collected from the archives. The EFG data infrastructure technology is powered by the D-NET [15] software toolkit, which provides a rich and customizable set of data management services capable of coping with issues such as metadata collection, storage, indexing, transformation, and cleaning. D-NET also offers services for the deployment of portals that can be configured according to the target community requirements, hence enabling end-users to search/browse the information space. Moreover, the D-NET toolkit includes mediation services for systems to access the space through standard protocols, such as OAI-PMH [22] and SRW/CQL, and several exchange formats.

*Data curation.* Once metadata records are aggregated into a structurally and semantically homogenous information space, the EFG infrastructure enables archive experts to perform data curation actions by delivering easy-to-use tools for metadata validation, editing, de-duplication (e.g. the same persons and movies entities collected from different repositories). To this aim, the authors extended D-NET with services implementing the data curation functionalities for content and vocabulary checking, metadata editing, and authority file management (i.e., record de-duplication).

**Paper Outline:** Section 2 gives an overview of the problem and introduces the adopted solution. Section 3 describes the main characteristics of the EFG common metadata schema. Section 4 describes the D-NET software toolkit. Section 5 presents the EFG D-NET-based infrastructure and its extension with D-NET data curation services. Finally, Section 6 concludes the paper.

## 2 Overview of the Problem and Adopted Solution

The EFG data infrastructure delivers two main requirements as identified by the user community:

- *Single access point to the European movie archives*: it supports advanced search and browse over all different types of collections (videos, images, textual documents), visualization of detailed metadata descriptions, and metadata export to third-party services, including Europeana.
- *High-quality metadata descriptions*: the EFG information space does not contain documents with poor descriptions and avoids duplication of information.

As mentioned in the introduction, these requisites are hindered by the highly heterogeneous nature of the archives. In fact, content of different archives generally conforms to different metadata models and XML schemas, whose structure may vary from complex element trees to simple flat sets of elements. Moreover, such content may describe different entities or the same entities, but with distinct semantics; e.g., different vocabularies of terms and format representation standards for dates, names, time durations.

To tackle such heterogeneity, EFG delivered two main outcomes: the EFG common data model and relative XML schema, onto which archive metadata records can be mapped; the EFG data infrastructure, whose services offer functionality for (i) collecting XML records from the archives and transforming them onto records matching the common XML metadata schema, and (ii) curating the resulting records by identifying and fixing semantic errors and duplicates. The data infrastructure was realized adopting the D-NET Software Toolkit [15] and extending it with new services for data curation.

The data ingestion workflow (sketched in Fig. 1.) consists of four phases and requires an interaction between domain experts and infrastructure administrators, adequately supported by the infrastructure services. These actors are driven by a detailed methodology, whose aim is to enable a controlled data ingestion life-cycle which will incrementally lead to the publication in production of a high-quality information space. Such workflow consists of four phases:

**Phase 1: Metadata Mapping Definition.** Domain experts from the archives analyze the metadata they provide to determine how such information may structurally and semantically map onto the EFG metadata schema. The relative structural and semantic mapping rules are handed over to infrastructure administrators, who encode them in the form of D-NET scripts.

**Phase 2: Metadata Transformation and Cleaning.** Archive metadata records are collected via OAI-PMH or FTP protocols to be processed through the mapping scripts produced in phase 1 and generate corresponding EFG records. The resulting records are not immediately available for access, but stored in a “pre-production” information space, where the Phase 3 of the workflow can take place. As we shall see, the Phase 1 and Phase 2 may be fired several times to refine the mapping rules and achieve the best metadata quality.

**Phase 3: Metadata Quality Control and Enrichment.** Records in the pre-production Information Space can be validated and inspected to identify mapping errors, mistakes (e.g., typos), and duplicates. Specifically, the Content Checker Tool can be used to verify that structural mapping was properly performed, the Vocabulary Checker Tool notifies data providers about EFG records not yet complying with the common vocabularies, and the Authority File Manager (AFM) identifies possible record duplicates. This quality control process may lead to the redefinition of the mapping rules (Phase 1), the adjustments of the mapping scripts (Phase 2), or to a subsequent data enrichment process. The Metadata Editor Tool enables curators to edit EFG records, while the AFM can fire record merge actions and effectively remove the duplicates.

**Phase 4: Metadata Publishing.** EFG records which passed Phase 3 are moved to the production Information Space, where they become visible from the EFG portal and can also be exported to third-party providers, such as Europeana.

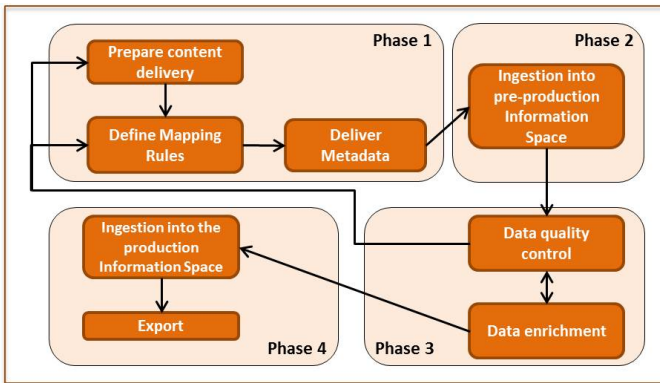


Fig. 1. Phases of the EFG data ingestion workflow

### 3 EFG Common Metadata Model and XML Schema


The EFG Common Metadata Model was designed after the analysis of the metadata models and schemas adopted within various organisations operating in the audio/video domain, starting from the data providers of the EFG consortium. This study took into consideration standards such as EAD [25], FRBR [4] and Dublin Core [5], as well as more film-specific standards such as the Cinematographic Works Standards EN 15907 [6]. As a result, eight interrelated entities have been defined in the EFG Common Metadata Model [19][24]:

- The *AVCreation* contains the properties of a cinematographic work: the film title, the record source (archive), the country of reference, the publication year, etc.
- The *AVManifestation* contains the information about the physical embodiment of an audiovisual creation. Examples are archival copies (analogue or digital) and database files. Properties of an *AVManifestation* include language, dimension, duration, coverage, format, rights holder, and provenance.



- The *NonAVCreation* describes all non audiovisual creations that can be represented in EFG. These are pictures, photos, correspondence, books or periodicals. The properties of NonAVCreations are: title, record source, keywords, description, date of creation and language.
- The *NonAVManifestation* entity keeps track of copies of non-audiovisual objects. It has properties such as type (e.g. text, image, sound), specific type (e.g. photograph, poster, letter), language, dates (i.e. a date or period associated with the issue of the manifestation), digital format (including its status, size, resolution), physical format, geographic scope, rights holder.
- The *Item* entity points to the digital file held in the source archive. Its attributes are *isShownBy* (i.e. the URL reference to the digital object on the content provider's web site), *isShownAt* (i.e. the URL reference of the object in its information context), digital format, provider and country.
- The *Agent* is defined as an entity that can perform an action. The model includes three agent types: Person, Corporate Body and Group. For example, the Person Agent has the following properties: name (composed of prefix, forename and family name), type of activity, date (which specifies the temporal properties of the person in relation with his activity), place (where the activity was performed), sex. Similar properties are defined for Corporate Body and Group.
- The *Event* is an entity that can occur within the lifecycle of an audiovisual or non-audiovisual creation. Examples of Events are Physical Event (e.g. a public screening or a broadcast), Decision Event (e.g. when a manifestation of a creation was evaluated by a censorship body), IPR registration, Award (i.e. the award obtained by an audiovisual creation or an agent), Production event (e.g. dates and places where castings took place, dates and locations of shooting).
- The *Collection* is defined as a compilation of creations (audiovisual or non-audiovisual).

In order to better illustrate the model and the relationships it defines among the above entities, we show a real-case example about the film “2001: A Space Odyssey” directed by Stanley Kubrik. We may have a record description of the AVCreation as follows:

 <p>A classic cinema of adventure and exploration</p> <p>2001: a space odyssey</p>	<p>Title: “2001: A Space Odyssey”</p> <p>Record Source: IMDB</p> <p>Identifying Title: “2001: A Space Odyssey”</p> <p>Country of Reference: USA</p> <p>Production Year: 1968</p> <p>Keywords: Science Fiction, HAL, intelligent computer</p> <p>Description: “Mankind finds a mysterious, obviously artificial, artifact buried on the moon and, with the intelligent computer HAL, sets off on a quest”</p>
-----------------------------------------------------------------------------------------------------------------------------------------------------------------------	--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

The record description includes some metadata elements plus a thumbnail describing the AVCreation. We will have several AVManifestations associated to the AVCreation, such as all national versions of the movie, for example the Italian and the

American versions. At the same time we may have several Agents related to this movie. As an example, we show a record description for the movie director, Stanley Kubrick:

	Record Source: IMDB Name: Stanley Kubrick Region of Activity: UK Sex: male Type of Activity: director ViewBiography
-----------------------------------------------------------------------------------	------------------------------------------------------------------------------------------------------------------------------------

Furthermore we may have NonAVCreations such as posters and film reviews. All these entities are connected through relationships (see Fig. 2). The metadata record associated to each entity will be used to retrieve the archived object, while the relationships will be used to support browsing. As an example, it is possible to search for all movies directed by Stanley Kubrick in the '50s and browse all received awards, biographies of actors, etc.

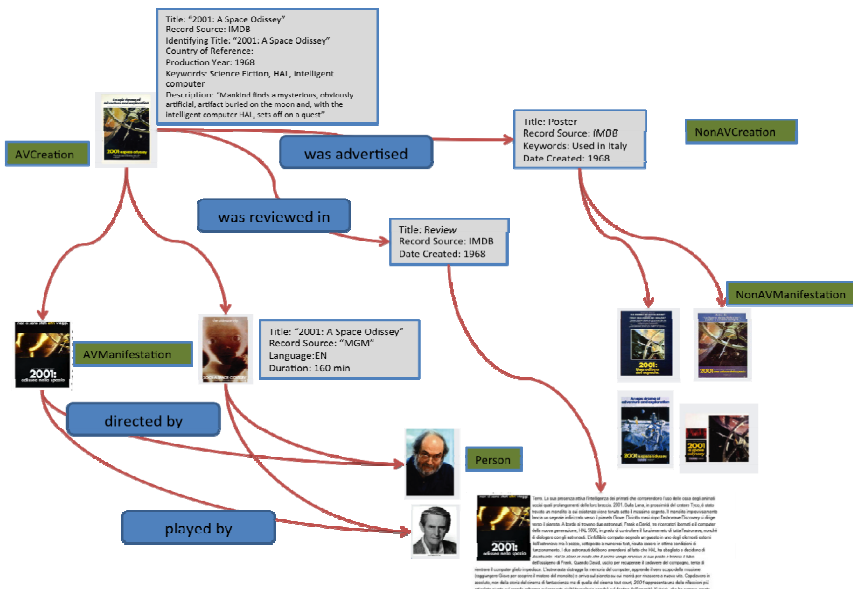


Fig. 2. Example of metadata associated for the film "2001: A Space Odyssey"

The EFG Common Metadata XML Schema [19] implements the common model described so far. It defines XML element types and attributes for all the eight entities and their relevant properties. The common schema is conceived as the type union of eight XML schemas (one for each entity) in such a way that one EFG XML record represents one entity together with its relationships to other entities. Furthermore, the

schema defines the so-called “controlled elements”, which are the XML elements whose values must comply with a given vocabulary of terms.

#### 4 Enabling Data Infrastructures: The D-NET Software Toolkit

In the last decade, as witnessed by several national initiatives (e.g., BASE 7, DAREnet [8], OAIster [9]) and EC projects (e.g., Europeana [3], Bricks [10], ScholNet [14], DILIGENT [11], D4Science [12], DRIVER [16], OpenAIRE [17], CLARIN [13], HOPE [18]), the diffusion of Digital Libraries which took place in the last ten-twenty years in several communities, has been followed by an urgent need for integrating and aggregating content from such DLs to make it available through a single access point. In the last three Framework Programme calls, the European Union initiated the so called *knowledge infrastructure vision*, inspired by the same goal of unifying data resources of all kinds available in Europe. The idea was that of devising *data infrastructures*, which are environments through which several organizations can share, process, aggregate their data resources by adopting an economy of scale approach. Several technological solutions [20] were devised in such projects, to offer functionality for collecting data from heterogeneous data sources (e.g. repository systems, archives, databases), curating such data to form a homogeneous information space, and offering customized portal services to operate over such space; e.g. search, inference of references between publications, citation calculation, etc.

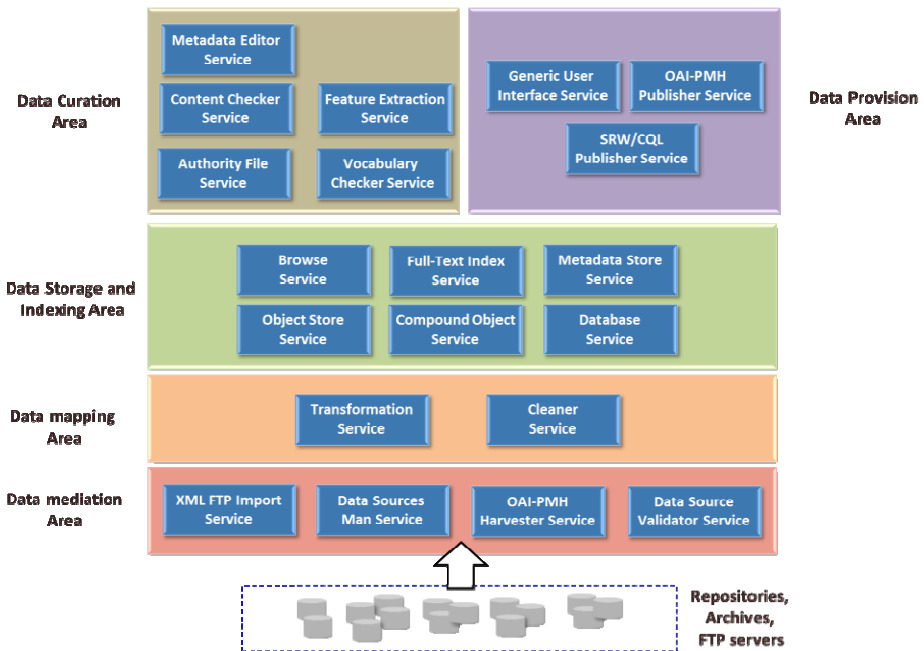


Fig. 3. D-NET service architecture

Of particular interest to Digital Libraries is the *D-NET software toolkit*, resulting from the experience of DRIVER, DRIVER-II, and OpenAIRE EC projects. D-NET is an open source solution specifically devised for the construction and operation of customized data infrastructures. D-NET provides a service-oriented framework where data infrastructures can be constructed in a LEGO-like approach, by selecting and properly combining the required D-NET services (such architectural concept was devised at CNR-ISTI by some of the authors of this paper). The resulting infrastructures are customizable (e.g., transformation into common metadata formats can be configured to match community preferences), extensible (e.g. new services can be integrated, to offer functionality not yet supported by D-NET), and scalable (e.g., storage and index replicas can be maintained and deployed on remote nodes to tackle multiple concurrent accesses or very-large data size). D-NET offers a rich set of services (see Fig. 3) targeting aspects such as data collection (mediation area), data mappings from formats to formats (mapping area), and data access (provision area). Services can be customized and combined to meet the data workflow requirements of a target user community. As proven by the several installations [15] and adoption in a number of European projects (DRIVER, DRIVER II, OpenAIRE, HOPE), D-NET represents an optimal and sustainable solution [21] for the realization of the EFG infrastructure. In the context of the EFG project, D-NET has been successfully extended with further generic and configurable services (curation area) for advanced curation and validation of XML metadata records.

## 5 EFG Data Infrastructure

The EFG data infrastructure consists of the D-NET services shown in Fig. 3, appropriately combined to support the data ingestion workflow presented in Section 2. In particular, the services in the Data Curation are resulted from the project activities. They were devised in order to meet the requirements of EFG archive partners, but engineered to support their functionalities when operating over arbitrary XML schemas.

### 5.1 Metadata Mapping Definition, Transformation, and Cleaning

Archives and their experts joining the EFG data infrastructure are supported with a methodology that facilitates the definition of *structural mappings* from their archive schema onto the EFG common metadata schema and *semantic mappings* from their vocabularies onto the common vocabularies. A mapping consists in a set of rules, which serve as input to the infrastructure administrators to configure the services in the Data Mapping Area. Here, the Transformator Service and the Cleaner Service run PERL scripts which parse, validate and transform the source records into EFG records according to the defined rules.

The Transformator Service is responsible for the application of *structural rules*. Such rules define the correspondence among elements and attributes of the archive schema and elements and attributes of the EFG schema. Structural mapping is not as

trivial as it may seem, due to the fact that input XML records are typically mapped onto several interrelated EFG records, representing different EFG data model entities. More in detail, a structural mapping rule consist of the following information:

1. *Source element*: xpath identifying the schema element relative to the input value;
2. *Target element*: xpaths identifying the schema elements (and the sub-entity) onto which the source value should be mapped;
3. *Mandatory element*: states if the source element is mandatory (if not, the record is rejected);
4. *Element multiplicity*: states if the source element is repeatable;
5. *Comment*: description of the mapping rule.

The Cleaner Service is instead responsible for the application of *semantic rules*. Such rules identify an element of the archive schema and the corresponding element of the EFG schema (i.e., source element and target element of structural rules), and define the correspondence between the terms of the respective vocabularies.

## 5.2 Metadata Quality Control and Enrichment

For the realization of the EFG data infrastructure the D-NET software toolkit has been extended with the following services, constituting the D-NET Data Curation Area.


*Content Checker*. The Content Checker (see Fig. 4) is a validation tool that allows low-level searching and browsing the pre-production Information Space in order to check if metadata records have been correctly harvested and mapped.


The screenshot shows the EFG Content Checker interface. At the top left is the EFG logo (European Film Gateway). At the top right is the title 'EFG Content Checker' and a welcome message 'Welcome Franco.zoppi | EFG Tools | Logout'. Below the logo is a 'Query' box containing 'itemType: image' and 'repositoryName: Istituto Luce'. A 'Revise Search' button is visible. The main content area is divided into search filters on the left and search results on the right.


**Search Filters:**

- decade**: 330006 (Total), 1960-1969 (5256), 1930-1939 (4683), 1920-1929 (4513), 1950-1959 (4278), [more](#)
- efgtype**: noniscreation (351315)
- itemtype**: image (351315)
- subject**: Esercito Italiano (11352), Africa Orientale Italiana (10706), Occupazione Italiana dell'Albania (6188), cronaca mondiale (5962), manifestazioni del regime fascista (5790), [more](#)

**Search Results:** Documents found (351315). Pages: [1] 2 3 4 5 6 7 8 9 10 >> (total)

**Record 1:**  **Attilio Piccioni all'inaugurazione della Fiera di Roma - Totale (Main title)**  
Record Type: noniscreation  
Cinecittà Luce S.p.A.  
[View the Record](#)

**Record 2:**  **Attilio Piccioni all'inaugurazione della Fiera di Roma - Campo lungo (Main title)**  
Record Type: noniscreation  
Cinecittà Luce S.p.A.  
[View the Record](#)

**Record 3:**  **Attilio Piccioni all'inaugurazione della Fiera di Roma - Piano americano (Main title)**  
Record Type: noniscreation  
Cinecittà Luce S.p.A.  
[View the Record](#)

**Record 4:**  **Attilio Piccioni e Urbano Ciocchetti all'inaugurazione della Fiera di Roma; tra i presenti Rebecchini - Campo medio (Main title)**  
Record Type: noniscreation  
Cinecittà Luce S.p.A.

Fig. 4. EFG content checker

*Vocabulary Checker.* The Vocabulary Checker gives access to the metadata records that do not satisfy the constraints imposed by the common metadata schema and vocabularies after the transformation and cleaning phases. The Vocabulary Checker displays the number, the types and the positions of errors in the records of the Information Space. Thanks to the browse by error typology functionality, curators can decide if an error can be solved directly in the Information Space via the Metadata Editor Tool or in the original source archive.

*Metadata Editor Tool.* The Metadata Editor Tool (MET) is a cataloguing tool for the enrichment of the Information Space. It allows data curators to add, edit and delete metadata records in the Information Space, as well as to establish relationships between existing (authority) records, even if coming from different sources. The MET is aware of controlled vocabularies, hence supports data curators while editing controlled elements by proposing a drop down list with all and only the terms defined by the associated controlled vocabulary. For example, let us suppose the Det Danske Filminstitut (DFI) EFG data provider provides a metadata record relative to the movie “*Olsen Banden over alle bjerge*”, which features the actor *Ove Sprogøe*, but the actor is not mentioned in the metadata record. In order to make the record retrievable through the EFG portal to end users searching for “*Ove Sprogøe*”, the movie record must be enriched with such information. The MET allows data curators to construct a relationship between the DFI movie metadata record and the person record, be the latter provided by harvesting other archives or created by data curators themselves.

*Authority File Manager.* The Authority File Manager (PACE [23]) is an advanced tool that curators can use to merge duplicate records and disambiguate the information space. The tool is capable of automatically identifying the pairs of records candidate for merging based on a multi-sort version of the sorted neighbourhood algorithm and a record similarity function that is customizable by data curators (they can choose between a range of similarity functions and assign different weights to the record fields). After one run of the candidate identification process, record pairs are displayed in descending order with respect to a 0...1 similarity distance. The curator has the responsibility of merging the two records (i.e., deciding if the two records are indeed representing the same entity). In the EFG scenario, the AFM has been configured to merge metadata records relative to persons and film works (AVCreation). Once the information space is disambiguated, authoritative records can also be linked to international ontologies such as VIAF [27] or transformed according to standard encodings, such as EAC-CPF [26], for external re-use.

### 5.3 Metadata Publishing

The EFG Portal is available at [1]. Facilities like advanced metadata search and browse (by collection, provider, date, language and media type), search results filtering, video streaming, photo gallery and news highlights enhance the user experience in the phases of search and access. Moreover, D-NET offers services to export metadata records through OAI-PMH, OAI-ORE, and SRW/CQL protocols. EFG operate

such services to automatically serve its information space to third-party consumers, above all the Europeana project [3], of which EFG is a direct feeder.

## 6 Conclusions and Future Work

We described the solutions adopted in the EFG Best Practice Network to achieve a complete integration of different national audio/video archives. The solution is based on the creation of a metadata schema that is an expressive interoperability metadata schema integrating well-known standards with peculiarities of data providers' idiosyncratic schemas. The schema has therefore both the power to preserve the input metadata quality and the simplicity to enable simple mappings from all different archives. Metadata aggregation is based on the use of the D-NET software toolkit, a data infrastructure enabling software. D-NET offers services for metadata collection, transformation, and provision. Its service-oriented framework allows for the addition of new services, to add domain specific missing functionalities. In EFG this resulted in the realization and integration of advanced curation and validation services: the Content Checker, the Vocabulary Checker, the Metadata Editor Tool and the Authority File Manager.

The current limitations of the EFG data infrastructure relate to the manual effort required in the phases of mapping rule definition and implementation and of metadata quality control and enrichment. Whilst some of the operations cannot be fully automatized, because archive administrators want to have control on data manipulation processes, we foresee some enhancements to (i) facilitate domain experts in the definition of mappings, (ii) (partly) automate the script-implementation of those mappings, and (iii) support experts and system administrator to ensure better metadata quality. Data provider experts currently define mappings by filling prefabricated Excel worksheets. Such files are then manually processed by infrastructure administrators to generate the corresponding transformation scripts. We could simplify this workflow by supporting data providers with a mapping definition tool, equipped with a GUI that shows a visual representation of their metadata schema and the common schema, and allows them to draw mappings by "dragging and dropping" elements of the first to elements of the second. The same tool could "generate" transformation scripts, at least when mappings can be reduced to a sequence of rule templates. Finally, we believe that the number of iterations of the transformation/cleaning and validation workflow could be reduced by providing a mapping test environment, where domain experts and infrastructure admins can verify the result of their mappings over a set of sample records.

**Acknowledgements.** This work is partly funded by the EFG Best Practice Networks project: grant agreement ECP 517006-EFG, call: FP7 EU eContentplus 2007. Its completion would have not been possible without the precious cooperation of Marco Rendina (Istituto Luce, Italy), Georg Eckes and Francesca Schultze (Deutsches Filminstitut, Germany) for the design of the data model.

## References

1. European Film Gateway project, <http://www.europeanfilmgateway.eu>
2. eContentPlus framework,  
[http://ec.europa.eu/information\\_society/activities/econtentplus/index\\_en.html](http://ec.europa.eu/information_society/activities/econtentplus/index_en.html)
3. Europeana, <http://www.europeana.eu>
4. Functional requirements for bibliographic records: final report/IFLA Study Group on the Functional Requirements for Bibliographic Records. UBCIM Publications; New Series, vol. 19, Saur, K.G., München (1998)
5. Weibel, S.L.: Metadata: The Foundations of Resource Description. D-Lib Magazine (1995), <http://www.dlib.org/dlib/July95/07weibel.html>
6. Cinematographic Works Standard. Committee, Technical (2005)
7. BASE: Bielefeld Academic Search Engine, <http://www.base-search.net>
8. DAREnet: Digital Academic Repositories, <http://www.darenet.nl/>
9. OAster Official Site, <http://www.oaister.org>
10. Bricks Project, <http://www.brickscommunity.org/>
11. DILIGENT Project, <http://diligent.ercim.eu/>
12. D4Science Project, <http://www.d4science.eu/>
13. CLARIN Project, <http://www.clarin.eu/>
14. ScholNet Project,  
<ftp://ftp.cordis.europa.eu/pub/ist/docs/rn/scholnet.pdf>
15. D-NET Software Toolkit,  
<http://www.d-net.research-infrastructures.eu>
16. DRIVER Project, <http://www.driver-community.eu/>
17. OpenAIRE Project, <http://www.openaire.eu/>
18. HOPE Project, <http://www.peoplesheritage.eu>
19. Balzer, D., Debole, F., Savino, P.: Common interoperability schema for archival resources and filmographic descriptions. Deliverable D2.2 EFG Project
20. Manghi, P., Mikulicic, M., Candela, L., Artini, M., Bardi, A.: General-Purpose Digital Library Content Laboratory Systems. In: Lalmas, M., Jose, J., Rauber, A., Sebastiani, F., Frommholz, I. (eds.) ECDL 2010. LNCS, vol. 6273, pp. 14–21. Springer, Heidelberg (2010)
21. Manghi, P., Mikulicic, M., Candela, L., Castelli, D., Pagano, P.: Realizing and Maintaining Aggregative Digital Library Systems: D-NET Software Toolkit and OAster System. D-Lib Magazine 16(3/4) (2010)
22. Lagoze, C., Van de Sompel, H.: The making of the open archives initiative protocol for metadata harvesting. Library Hi Tech 21(2), 118–128 (2003)
23. Manghi, P., Mikulicic, M.: PACE: A General-Purpose Tool for Authority Control. In: García-Barriocanal, E., Cebeci, Z., Okur, M.C., Öztürk, A. (eds.) MTSR 2011. CCIS, vol. 240, pp. 80–92. Springer, Heidelberg (2011)
24. Savino, P., Debole, F., Eckes, G.: Searching and browsing film archives. The European Film Gateway Approach. In: 4th International Congress on Science and Technology on the Safeguard of Cultural Heritage in the Mediterranean Basin, Cairo, Egypt, December 6-8 (2009)
25. Encoded Archival Description, <http://www.loc.gov/ead>
26. Encoded Archival Context Corporate Bodies, Persons, and Families,  
<http://eac.staatsbibliothek-berlin.de>; Virtual International Authority File, <http://www.viaf.org>



# Annotating Digital Libraries and Electronic Editions in a Collaborative and Semantic Perspective

Michele Barbera<sup>1</sup>, Federico Meschini<sup>2</sup>, Christian Morbidoni<sup>3</sup>, and Francesca Tomasi<sup>4</sup>

<sup>1</sup> Net7, Pisa, Italy, SpazioDati, Trento, Italy

barbera@netseven.it, barbera@spaziodati.eu

<sup>2</sup> Department of Humanities, Communication and Tourism, Tuscia University, Italy

fmeschini@unitus.it

<sup>3</sup> Semedia Group, Università Politecnica delle Marche, Ancona, Italy

christian.morbidoni@gmail.com

<sup>3</sup> Department of Classical Philology and Italian Studies, University of Bologna, Italy

francesca.tomasi@unibo.it

**Abstract.** The distinction between digital libraries and electronic editions is becoming more and more subtle. The practice of annotation represents a point of convergence of two only apparently separated worlds. The aim of this paper is to present a model of collaborative semantic annotation of texts (SemLib project), suggesting a system that find in Semantic Web and Linked Data the solution technologies for enabling structured semantic annotation, also in the field of electronic editions in Digital Humanities domain. The main purpose of SemLib is to develop an application so to make easy for developers the integration of annotation software in digital libraries, which are different both for technical implementations and managed contents, and provide to users, indifferently from their cultural backgrounds, a simple system which could be used as a front-end. We present, for this purpose, a final example of semantic annotation in a specific context: a digital edition of a literary text and the issues that an annotation task involves.

**Keywords:** ontologies, Open Collaboration, Linked Data, TEI, RDF.

## 1 Introduction

In the Library of Alexandria, the distinction between philologists and librarians was almost not existent, since the functions of acquisition, cataloguing and preservation were strictly related to an editorial work which main aim was to give to the texts the best possible rigour and accuracy [1]. The progressive specialization of skills has brought, as a natural consequence, the loss of a global view, also in strongly linked sectors like the ones quoted above. Let's only think about the different meanings that the term bibliography can assume when used with different prefixes such as 'analytic' or 'descriptive'. This same loss has been denounced by Vannevar Bush [2] and therefore his vision of the Memex was a possible solution for this situation. It is therefore quite a paradox that in digital information systems and frameworks published on the World Wide Web, representing on one side Digital Libraries (DLs) and, on the other,

Electronic or Digital Editions, this difference is, if possible, even sharper, having generated two different scholarly communities, which, even though overlapping, presents their own distinctive features.

In fact, by concentrating on cataloguing and digitizing collections rather than analysing the content of individual items in a collection, the DL is mostly focused on publishing mechanisms (or to better say the dissemination ones) — an opposite extreme from the focus of electronic editions on textual encoding, in particular the one based on the TEI standard [3]. The DL, instead, is by nature agnostic towards the contents it has to manage, since they could be very heterogeneous, preferring therefore to ignore the granularity of an encoded text, which is fundamental for an electronic edition. Moreover, a DL contains reproduction of physical objects or content *born-digital*, while in the current situation the primary sources upon which an electronic edition is based have an almost and exclusively analogical origin. But interaction between these two paradigms is actually taking place, even though, in a not very organized and co-ordinated way, following what clearly are physiological patterns, but in a deeper and more dynamic modality than the one allowed by the analogical dimension<sup>1</sup>.

The electronic edition and the DL are modelled on the needs, experiences and uses of two different communities, even though related, and this is a natural expectation. What is not completely expected is the (re)definition of the respective natures, and overall modalities of interaction of these two entities, in particular from the point of view of their (re)modelling, following computational principles. In fact now the differences have a logical base rather than a physical one.

Even though the technologies are necessarily the same, this does not assure a complete homogeneity at the level of methodologies, approaches and solutions, in other words it does not guarantee the perfect overlapping and compatibility of the hypothetical semantic models. This discrepancy is mitigated by observing two opposed and complementary movements that are currently taking place. On one side the electronic editions and archives are expanding, becoming more and more complex and stratified, while on the other side DLs are becoming more and more granular (together with their natural tendency for progressive growth). In fact, of the two currently available formal models for DLs, the *5S* [7] and the *DELOS Reference Model* [8], the former, being directly based on first-order logic, is granular and expressive enough to model also electronic editions, both the textual encoding level and the very different relationships existing between the witnesses which made up the textual tradition.

Building on the facts presented until now, the most logical consequence is that a focus on semantic and formal models seems the only way for breaking the barriers between these DLs and electronic editions, even though, in this latter case the

---

<sup>1</sup> For instance, there is the progressive adoption of *IFLA FRBR* model, in projects such as *Perseus* <<http://www.perseus.tufts.edu>> [4] or the *Canonical Text Service* protocol <<http://chs75.chs.harvard.edu/projects/diginc/techpub/cts>> [5] or standards such as METS. On the digital libraries side, an interesting case is the use of the publishing framework *Cocoon*, inside *Dspace*, in a component called first *Manakin* and then *XMLUI*, used to customize the user interface. The *XML* schema used to transport the data, *DRI Schema*, is based on *TEI* for what concerns the actual contents [6].

prevailing diffusion of the TEI encoding, notwithstanding its many advantages, is a potential hindrance, since it brings the focus mostly on data structures which lack a formal semantics [9], [10].

The annotation task becomes then crucial. In the context of both DLs and electronic editing, the term annotation indicates the process of adding some kind of information to an existing digital resource. It can be a tag, a comment or some kind of structured metadata. The task of librarians is to provide sets of high quality annotations for each library resource, in order to help organizing the knowledge gathered in each library. These metadata are usually designed in accordance with standard library science practices and are meant to facilitate knowledge discovery by the generic library user. A DL can be accessed by several user communities, each one with a specific vision of the world and each one interested in a specific topic or aspect of knowledge. On the contrary, built-in metadata in DLs are often generic information (e.g. year of publication, author, historical period, artistic wave, etc.), or reflects a single viewpoint. They do not capture all the aspects the users might be interested in, thus being often of poor value with respect to interesting resources discovery.

A similar problem exists in more specialized types of digital publications, such as critical text editions. In digital editions it is common practice that the editors enrich the text by annotating it with TEI or other forms of markup. These annotations are then used to deliver a richer reading and searching experience for the final user. Despite this practice has certainly proved useful, it also suffers from a similar limitation to the one outlined above for generic DLs. Annotations made by the editors are intrinsically static and relative to a particular view of the world or school of thought. Differently than paper artifacts, digital resources can be easily exploited as social objects around which communities can collaboratively and continuously enrich the digital artifacts with different interpretations.

According to authoritative studies [11], [12], [13] DLs, and more specific collections of digital objects, should allow their users to annotate resources and leave comments. They should also let users share their index and classification schemata with other users [14]. We believe that DLs and more specialized collections in the Digital Humanities field, can greatly benefit from the availability of Web annotation tools based on Semantic Web and Linked Data technologies and the aim of our research is to show in which way.

The paper is structured as follows. Section 2 presents related works in the field of annotation. Section 3 discusses the approach taken within the context of the SemLib project, whose prototypal intermediate results (the project will publicly release its final results in December 2012) represent an interesting experiment in this direction.

Section 4 gives a specific case study of semantic annotation in a digital edition.

## **2 Related Works**

Annotating Web documents like Web pages, part of Web pages, images, audios and videos is one of the most spread technique to create interconnected and structured metadata on the Web. In the last years several automatic, semi-automatic and manual

systems have been proposed that provide support for creating annotations at different levels and in diverse scenarios. Some applications have been developed as extensions of social bookmarking tools and have become a popular service over the Web with application as Delicious<sup>2</sup> or StumbleUpon<sup>3</sup> that count millions of registered users. Other tools have been more specifically conceived for manually creating and sharing annotations in specific domains, including Digital Humanities and Cultural Heritage [15], [16]. Early implementations of manual annotation tools have been mostly developed as desktop applications or browser plugins (such as Zotero<sup>4</sup> and others). With the growing availability of powerful client side Web programming tools and techniques, annotation tools then evolved in fully fledged Web applications such as EuropeanConnect Media Annotation Prototype [16] based on Annotea [17], One Click Annotator [18], the Open Knowledge Foundation's Annotator project<sup>5</sup>, SharedCopy<sup>6</sup>, A.nnotate<sup>7</sup> and many others. Another widely applied method to create annotations is to use automatic and semi-automatic tools based on euristics like natural language processing, image recognition, audio and video segmentation. For textual content there are several widespread commercial services that automatically perform a light type of annotation known as entity extraction with a constantly improving degree of relevance (e.g. Opencalais<sup>8</sup>, Zemanta<sup>9</sup>, AlchemyAPI<sup>10</sup>).

While some of the existing tools address ease of use and wide adoption, they hardly provide support for expressing non trivial semantics, as establishing precise (typed) relations among digital objects or referring to specific entries in domain thesauri and vocabularies. In Semlib, the goal is to build different annotation GUIs to address different levels of expressivity, from simple tags to structured conceptual graphs, carefully balancing ease of use and expressivity. The other idea behind SemLib is that of representing annotations (simple or complex ones) in a uniform way (as RDF graphs), and expose them via REST APIs so to enable effective reuse of collaboratively created knowledge, for example to further enrich DLs.

### 3 Collaborative Semantic Annotation of Texts: The SemLib Approach

One of the main goals of the SemLib project [19] is the design and implementation of a semantic aware annotation system that can be easily used in conjunction with different DLs, requiring as less modification as possible to the existing DL software infrastructures, and that can be flexible enough to address different needs of specific

---

<sup>2</sup> <http://delicious.com/>

<sup>3</sup> <http://www.stumbleupon.com/>

<sup>4</sup> <http://www.zotero.org>

<sup>5</sup> <http://okfn.org/projects/annotator/>

<sup>6</sup> <http://www.sharedcopy.com/>

<sup>7</sup> <http://www.a.nnotate.com/>

<sup>8</sup> <http://www.opencalais.com/>

<sup>9</sup> <http://www.zemanta.com/>

<sup>10</sup> <http://www.alchemyapi.com/>

communities. Such differences both reside in required expressivity and complexity of annotations, which might range from simple tags to non trivial semantic relations among media content and other kind of entities, and in the use of different domain dependent terminology and vocabularies.

A core requirement in SemLib is that of enabling reuse of annotations, for example to leverage them as a crowd-sourced structured knowledge that might be used to enrich DLs themselves. While interoperability at data representation level is certainly a key feature with respect to this goal, the system has to provide effective ways to meaningfully consume such data, for example allowing external applications to search annotations and to obtain “slices” of the overall annotations (e.g. obtaining annotations from trusted users only, or those that involves relevant resources only, etc.).

### 3.1 From Tagging to Semantically Structured Annotations

The simple form of annotation, widely understood and adopted by the majority of Web users, is tagging. Keywords based tagging, however, has several disadvantages (between them: no explicit meaning and explanation; polysemy; synonymy; base form variation; specificity gap; reused in different systems). These poor semantics expressed by “traditional” tags prevents in fact the use of annotations to produce reusable structured knowledge, which is the core goal of SemLib. To overcome such limitations, SemLib supports “semantic tagging”, where each tag corresponds to an entry in a controlled vocabulary or ontology and it is a Web resource in itself, thus being resolvable into a natural language description by dereferencing its URL. A similar approach has been already experimented in the Common Tag initiative ([CommonTag.org](http://CommonTag.org)). The current prototype allows users to transparently search for entities (semantic tags) in [Freebase.com](http://Freebase.com), providing auto-complete suggestions and resulting in external web resources to be associated to text fragments or pictures in a web page.

Interestingly, such web resource happens to be Linked Data sources. This means that they can be used to retrieve further information about the entities, allowing external applications, which consume such annotations, to immediately use such additional data in intelligent ways. In addition, existing APIs, such as DBpedia spotlight, are used to suggest simple forms of automatic tagging.

SemLib also supports more advanced types of annotations that exploit all the expressive power of the RDF data model, which goes beyond simple semantic tagging. As an example, suppose Alice is a scholar studying Italian literature. She finds a DL with some interesting novels. While reading one of them, she highlights a paragraph and creates an annotation specifying that such a paragraph cites Alessandro Manzoni and that attempts to give a definition of “Historical novel”. Semantic tagging as described above, does not allow to specify the relation (e.g. cites, defines) between the text and the related entities (e.g. “Alessandro Manzoni” and “Historical novel”), which is needed to answer queries like “what are the paragraphs that cites a given author?” or “What definitions of historical novel does the system know?”

The SemLib annotation tool supports the creation of such complex annotations by allowing user to collect different kind of “items” (they can be terms from a vocabulary, web pages, or fragments of them, e.g. sentences and pictures) and then connect them via semantically typed relations, thus in fact creating a semantic graph. The annotation tool can be configured to use custom vocabularies to accommodate the

needs of different DLs. Vocabularies can be published on the web in a simple JSON based format and then ingested by the tool by simply specifying their URL.

Cross-references annotations are an interesting special case that is often required by scholars. It consists in establishing a semantic relations between two media fragments, such as a sentence and other one in a different page (possibly in a different DL), or a sentence and a specific region of an image. From a conceptual and data representation point of view, they are equivalent to other semantic relations, however they raise new challenges at the user interaction level. In SemLib such kind of annotations are made possible by allowing user to bookmark media fragments, to surf to other web pages and then to reuse such bookmarked items in annotations.

### 3.2 Data Model and API

The representation of semantic annotation is composed by two distinct parts: the annotation context and its semantic content. The first represents information such as the author of the annotation, while the latter represents the actual meaning or knowledge that the user wanted to express in the annotation. The data model is illustrated in Fig. 1 as an RDF graph.

Different RDF based data models for representing web annotations have been proposed in literature. In SemLib we decided to base on the Open Annotation Collaboration (OAC) data model [20]. The OAC ontology is used to represent annotations contexts, which specifies, through the *oac:hasTarget* property, the web resources involved in an annotation. An additional concept used in the SemLib model is that of notebook, which is an aggregation of annotations. Each user can have multiple notebooks, e.g. to group annotations pertaining to different tasks or contexts. Notebooks have a central role in the overall functioning of the system, as they constitute the granularity where user privileges are attached and annotations are shared among users.

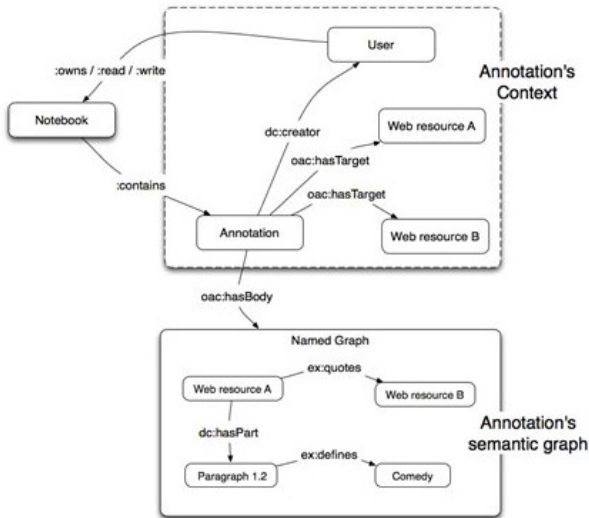


Fig. 1. The annotation data model

The *oac:hasBody* property links the annotation to its content. In SemLib the body, rather than being a text or a web page (as it happens in most of the examples given in the OAC specifications) is a RDF graph itself. The content graph is made addressable by using named graphs. As shown in Figure 1, a named graph is used as value of the *oac:hasBody* property. As annotations are stored in a quad-store and the SPARQL standard query language natively supports named graphs, such an approach results in more flexibility in consuming annotations.

Named graphs are sub-graphs of an RDF graph that can be merged and queried as single graphs. The use of named graphs allows, for example, querying only those graphs that belongs to annotations in a given notebook. The same can be done for collections of annotations grouped via other criteria (e.g. from the same user, involving the same resource, etc.). A set of RESTful APIs are exposed by the annotation server to provide an easy way of consuming slices of the overall annotations set in various RDF serialization formats, and additional custom queries can be performed via standard SPARQL endpoint.

While the annotation context is represented using a fixed ontology (an extension of the OAC ontology), semantic content of annotations can use any ontology that a certain domain requires (e.g. a TEI-derived ontology in the Digital Humanities main<sup>11</sup>).

### 3.3 Enabling Collaboration in the Digital Humanities

One of the most relevant limitations of annotation systems based on embedded markup, such as HTML or XML, is the tight coupling between the annotation and the annotated object [21]. In the Digital Humanities community the problem is acute, partly because most of the times digital edition projects do not publish the XML annotated version of the text, but rather a derivative HTML or PDF version produced from the original XML. In these cases annotation ceases to be metadata to become part of the digital objects, thus not being reusable, preventing collaboration and further enrichments of the text.

Even when the XML source files are distributed along with the human readable version, they are seldom reused or integrated outside the boundaries of the systems in which they were originated. The semantics of the annotations is in most cases based on local interpretations and local extensions of the core vocabulary. Additionally, vocabularies and thesauri are not shared and the semantics of local extensions is not machine-readable. As a result, textual resources, their interpretations and enrichments, remain siloed within the boundaries of individual projects. Data is not shared, derived information and interpretation is hidden within the browsing applications and not addressable or reusable. As a result, digital collaboration practices are rather weak if compared to other disciplines where information sharing and reuse is more common (e.g. ‘hard’ sciences).

Stand-off semantic annotations based on the RDF standard, as described in the previous section, have the potential to overcome many of these limitations.

---

<sup>11</sup> TEI ontologies SIG <http://www.tei-c.org/SIG/Ontologies>

Annotations are useful in organizing and adding information to digital content, supporting single users in studying and exploring online resources. However, a lot more value can be added if annotations are shared with others, enabling, for example, virtual communities to perform collaborative tasks. In SemLib, both annotations and vocabularies can be published on the Web in machine-readable format (RDF) so that any authorized application can interpret annotations. This basically decouples the applications used to access the DL from the content itself, allowing for multiple “views or interpretations” of the same content to be published in a decentralized way. Another interesting side effect of the Semantic Web approach, is that it naturally enables collaboration: as long as users have the possibility to upload their annotation somewhere on the Web, they are free to keep enriching the content stored on the DL without the need of any coordination with the content holder. Users communities can share their work by simply exchanging URLs that point to their annotation graphs. At a later stage, user annotations can also be made authoritative by the library curators, by incorporating them into the DL. Annotations are provided with provenance meta-data, so that it is always possible to determine who made a specific annotation and when.

While the prototype has not yet been released to the public, some online screen-casts demonstrate the core functionalities and user interactions<sup>12</sup>.

#### 4 Adding Meaning. A Case Study

The case study we present here is a digital edition of a XV century collection of letters now held in different libraries and, mostly, archives. The purpose of the edition is to experiment a concrete case of semantic annotation starting from a sequence of XML/TEI files, regarding the same field.

The correspondence documents the professional relationships managed by the Florentine librarian and copyist Vespasiano da Bisticci, who was also the leader of a school of copyists, maker of some European libraries’ manuscript repositories. The correspondence is with notable people of that period and the content regards mostly the trade of manuscripts copied, proposed or requested by/to Vespasiano. A lot of these manuscripts had been identified in codices now held in various libraries in all Europe. From the letters we can learn about features of these manuscripts: the materials, the copyists, the costs, but also the names of latin, greek and humanistic authors and texts that were the most fashionable at the time.

The purpose of the digital edition, moving toward a DL of digital objects, is on one side to represent the information that is implicitly connected inside the source (people with manuscripts, manuscript with lexicon); on the other to create semantic links between the information inside the letters and correlated Web resources, useful in order to describe this information (people, manuscripts, lexicon). The first purpose is thus to create relationships between people and manuscripts related to people at some level

---

<sup>12</sup> [http://www.youtube.com/watch?v=gVA\\_v152Qn0](http://www.youtube.com/watch?v=gVA_v152Qn0),  
<http://www.youtube.com/watch?v=z1hXr5K3kTM>



(the copyist, the owner, the requester) and the manuscripts with the lexicon used in order to describe them; this information born naturally from the letters and have to be expressed with semantic assertions through external annotations. The second aim is to create relationships between the specific document and other documents: people and resources useful to describe people (public prosopography), mentioned manuscripts with the existent codices (catalogues of manuscripts), the lexicon with repositories of technical words (thesauri) using the Linked Data system.

#### 4.1 The Embedded Markup

The markup model we present covers all the different aspects of the edition, that is the model reflects the different access points to the letters' content. The focus is on: persons mentioned in the letters; manuscripts realized by Vespasiano's school; the technical lexicon of the copy and of librarian trade.

At the markup level it is quite easy to represent, with the appropriate TEI elements, all this information. The base TEI markup let us identify: proper name (`<persname>` and `<placename>`) and referring string (`<rs>` with `@type` for specification), mentioned manuscripts (`<bibl>`) and related author (`<author>`) and title (`<title>`), technical term (`<term>`) on various field (`@type`).

The `@ref` value in the instances of `<persname>`, `<placename>` and `<rs>` allows a first identification of individuals being mentioned: missing parts of names are solved outside and variants are associated to the same instance. The same attribute `@ref` was also used for the element `<bibl>` and `<term>`, with a TEI customization. The `@ref` value points to a specification of the item stores in a place outside the document. Therefore, within these elements, an access point is defined as an element owning a URI reference that points to `@rdf:about` attribute in the external repository. In this way each pertinent string of characters or fragment has an URI.

If the URI pointing gives us the possibility to formally describe annotated elements in the external representation, we need a system to create connections between these annotated elements. These connections must answer to questions like: which relation exists between a person and a manuscript? And which one between the same manuscript and a technical word used to describe the manuscript? But we need to answer also to questions like: which exemplar of the manuscript has been realized? Is it still available in any library in Europe? What a specific technical word means? There are any other occurrences in other repositories? Who is a mentioned person?

For the first set of questions we need some more formal semantics in order to describe relationships between annotated data elements. For the second set of questions we have to use techniques useful in order to connect the edition with existent resources on the Web.

#### 4.2 The External Information

Once the texts are annotated, and each pertinent string of characters has an URI, it is necessary firstly to focus on what kind of additional information it is possible to define in the separated documents. Making each occurrence (persons, manuscripts and

lexicon) accessible via URI-based pointing we can create a public authority list. Starting from Vespasiano's letters it became possible to expose authority information about people, manuscripts, technical lexicon regarding XV century culture, in order to start to create an open and public authority list, to be integrated in authority records, as a set of Web resources.

We need firstly to decide which kind of relevant added data it is possible to specify for each of the three categories. Then we have to reflect on which kind of relationships we could define, expressing the content through stand-off semantic annotations.

Persons. At this level we define relationships: with other persons; with places; with dates; with other resources (like multimedia data); and with mentioned manuscripts (i.e. a person could be owner of a codex created by Vespasiano's school).

Manuscripts. With regard to manuscripts mentioned in the letters it is possible to create link to other repositories and establish relationships with both persons and terms. We have to create relationships: with the codex repository; with the codicological description; with a digital image; with the digitalized full text; with a person relevant to it (i.e., the owner, the requester, the copyist, etc.).

Lexicon. The analysis of the technical lexicon used by Vespasiano is an interesting exploration in the history of the book and in the actual trade of the copy, and it is sometimes connected to the manuscripts realized. At this level it is possible to define relationships between manuscripts and lexicon used.

### 4.3 The Knowledge Base

All this external information could be described in a formal way through RDF assertions, like similar researches did (i.e. [22], [23]) and the annotations could be accessible over the Web as Linked Data sets. Mostly we need to focus on the fact that between these concepts (persons and manuscripts; manuscripts and lexicon) we can define relationships that provide greater conceptual depth and that can be easily expressed in a formal language, creating a good model for the representation of the content of the letters: we can establish unambiguously the different relationships existing between a person, a manuscript and a term, such as a person being the owner, the copyist, or the client of a codex and this manuscript is described with a specific term (p.e. Piero de' Medici is owner of Plutarco's *Vite*, which are realized in "chordovani"). Mapping the specific created class (persons, manuscripts, lexicon and the different kind of relationships) with predicates defined in existent model a suitable ontology could be distributed. We are now analyzing different predicates for internal connection, starting from these consideration: a person could be *owner-of*; *copyist-of*; *illuminator-of*; *requester-of* a manuscript; a manuscript could be *created-for*; *requested-by*; *copied-by*; *illuminated-by* a person / *described-with* terms; the lexicon is *related-to* a manuscript.

But like we explain in the previous section the external information could be linked to some external resources. We are studying some predicates. For person, TEI prosopography integrated with CIDOC-CRM<sup>13</sup> for relationships between Agents, Physical things, Events and Places. For manuscripts an ontology properly extending and

---

<sup>13</sup> CIDOC Conceptual Reference Model, <http://www.cidoc-crm.org/>

customizing the FRBR<sup>14</sup> can be used to capture the subtle difference between the physical codex, the content it has and the work the codex is a *manifestation-of*. For the lexicon SKOS<sup>15</sup> provides a basic ontological foundation for a terminological thesaurus and the relationships of the terms defined therein and elsewhere.

But we mostly need to reflect on the fact that the mentioned manuscripts are specific codices that can be found nowadays in a specific library, or a person has an iconographic representation that could be found in a certain cultural institute. With URI references and RDF representation we started to create relationships between people, manuscripts and lexicon of the letters and the related concept all over the Web, using existent Linked Data sets and exposing our annotation as Linked Data sets in open collaboration.

## 5 Conclusions

Having started from a general description of current epistemological and ontological differences between digital libraries and electronic editions, together with the related underlying rationale, the choice of focusing on open and semantic annotation has been considered by the authors a strategic one, since as it has been demonstrated this “functional primitive” has the potential to be a bridge between these two different but strongly related worlds. Therefore both the act of annotating a text, thus giving it the status of “scholarly” and the addition of (meta)information to a generic resource can greatly benefit from the adoption of a common formal model, which at the same time make explicit both their actual natures and potentialities.

The aim of SemLib project is to use standard technologies to create a really usable application, easy to use for every kind of user and easy to integrate in heterogeneous digital libraries, therefore filling up a current gap in the landscape of annotation applications on three different levels: actually usability, ease of use, integration. For these reasons we are organizing focus groups which would consolidate the theoretic foundations which led the implementation of the application.

**Acknowledgements.** The research leading to these results has received funding from the European Union's Seventh Framework Programme managed by REA-Research Executive Agency ([FP7/2007-2013][FP7/2007-2011]) under grant agreement n. 262301.

## References

1. Ore, E.S.: ... they hid their books underground. In: Deegan, M., Sutherland, K. (eds.) Text Editing, Print and the Digital World, pp. 113–125. Ashgate, Farnham (2009)
2. Bush, V.: As We May Think. *Atlantic Monthly* (1945)
3. Burnard, L., Bauman, S. (eds.): TEI P5: Guidelines for Electronic Text Encoding and Interchange. TEI Consortium (2007)

---

<sup>14</sup> Functional Requirements for Bibliographic Records, <http://www.frbr.org/>

<sup>15</sup> Simple Knowledge Organization System, <http://www.w3.org/TR/skos-reference>

4. Mimno, D., Crane, G., Jones, A.: Hierarchical Catalog Records: Implementing a FRBR Catalog. *D-Lib Magazine* 11(10) (2005)
5. Smith, N.: Citation in Classical Studies. *Digital Humanities Quarterly* 3(1) (2009)
6. Phillips, S., Green, C., et al.: Manakin: A New Face for DSpace. *D-Lib Magazine* 13 (2007)
7. Gonçalves, M.A.: Streams, Structures, Spaces, Scenarios, and Societies (5S): A Formal Digital Library Framework and Its Applications. *Computer Science and Applications*. Virginia Polytechnic Institute and State University, Blacksburg, Virginia, U.S.A. PhD: VII-153 (2004)
8. Candela, L., Castelli, D., et al.: The DELOS Digital Library Reference Model. *Foundations for Digital Libraries Version 0.98, DELOS Network of Excellence on Digital Libraries* (2007)
9. Raymond, D.R., Tompa, F.W., Wood, D.: Markup Reconsidered. In: *First International Workshop on Principles of Document Processing*, Washington, DC (1992)
10. Buzzetti, D.: Rappresentazione digitale e modello del testo. In: *Il Ruolo Del Modello Nella Scienza e Nel Sapere*, pp. 127–161. *Accademia Nazionale dei Lincei*, Roma (1998)
11. Agosti, M., Ferro, N.: Annotations: Enriching a Digital Library. In: Koch, T., Sølvberg, I.T. (eds.) *ECDL 2003. LNCS*, vol. 2769, pp. 88–100. Springer, Heidelberg (2003)
12. 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)
13. Agosti, M., Ferro, N.: An Information Service Architecture for Annotations. In: *DELOS Workshop Digital Library Architectures*, pp. 115–126 (2004)
14. Kruk, R., Decker, S.: Semantic Social Collaborative Filtering with FOAFRealm. In: Decker, S., Park, J., Quan, D., Sauermann, L. (eds.) *Proc. of Semantic Desktop Workshop at the ISWC*, Galway, Ireland, November 6, vol. 175 (2005)
15. Bradley, J.: Pliny: A model for digital support of scholarship. *Journal of Digital Information (JoDI)* 9(1) (2008)
16. Haslhofer, B., Momeni, M., et al.: Augmenting Europeana Content with Linked Data Resources. In: *6th International Conference on Semantic Systems* (2010)
17. Kahan, J., Koivunen, M.R.: Annotea: An Open RDF Infrastructure for Shared Web Annotations. In: *Proceedings of the 10th International Conference on World Wide Web*, pp. 623–632 (2001)
18. Heese, R., Luczak-Rösch, M., et al.: One Click Annotation. In: Williams, G.T., Grimnes, G.A. (eds.) *CEUR Workshop Proceedings*, vol. 699 (February 2010)
19. Morbidoni, C., Grassi, M., et al.: Introducing SemLib Project: Semantic Web Tools for Digital Libraries. In: *International Workshop on Semantic Digital Archives, TPD, Berlin, Germany, September 29* (2011)
20. Sanderson, R., Van de Sompel, H. (eds.): *Open Annotation: Alpha3 Data Model Guide* (October 15, 2010), <http://www.openannotation.org/spec/alpha3/>
21. Thompson, H.S., McKelvie, D.: Hyperlink semantics for standoff markup of read-only documents. In: *SGML Europe* (1997)
22. Vieira, J.M., Ciula, A.: Implementing an RDF/OWL Ontology on Henry the III Fine Rolls. In: *OWLED 2007, Innsbruck, Austria* (June 2007)
23. Tummarello, G., Morbidoni, C., Pierazzo, E.: Toward Textual Encoding Based on RDF. In: *9th ICCCE International Conference on Electronic Publishing, Leuven-Heverlee, Belgium* (June 2005)

# Empowering Archives through Annotations

Nicola Ferro and Gianmaria Silvello

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

**Abstract.** The paper presents an integration and visualization service to enhance the use of annotations and to empower the role of the user and research community in the archival context. We show how this service allows us to address the interoperability between diversified digital archive and annotation systems. Furthermore, it propels the use of annotations to enhance the user experience and to exploit the archivists expertise both in the description and consultation phases.

## 1 Motivation

One of the main goal of the research on *Digital Library (DL)* is to supporting the creation of innovative applications and services to access, share and search our cultural heritage.

An important challenge in this field is to transform DL into a new type of information infrastructure that can be user-centered and able to support content management tasks together with tasks devoted to communication and cooperation [11]. DL can enable the intellectual production process and support user cooperation and exchange of ideas; in this way, DL not only foster access to knowledge, but they are also part of knowledge creation and evolution. The evolution and transmission of knowledge has always been an interactive process between scientists or field experts, and annotations have been one of the main tools for this kind of interaction. In the digital era, annotations are still a relevant means of intellectual collaboration and thus, one of the main collaboration tools exploited by DL [5].

The informative context enclosed by digital libraries is multifaceted and comprises many realities of interest such as libraries, archives and museums. In this paper we focus on the archives and archival metadata which are the basic means for accessing and consulting archival resources in a digital environment [18]. Annotations foster collaboration between archivists, researchers and general users by playing a central role both in the phase of *creation* and in the phase of *consultation* of archival metadata. Indeed, in the creation phase archivists have to select and describe the archival material and annotations allow them to explain and discuss their choices enabling users to properly access and consult the archival metadata. In the consultation phase, annotations are exploited to find out relationships between different parts of an archive or between different archives; for instance, users can exploit annotations to move from one archive to another guided by the expertise of the archivists that annotated them. In order to properly exploit annotations in the archival context we have to take

into account the heterogeneous environment composed of digital archive systems and annotation systems which are often grounded on different methodological and technological approaches. The archival community has developed “content and data structure standards” [15] to facilitate the description, management and access to the archival resources; however, these standards can be difficult for archivists to use [19] and are often implemented in ways that can negatively affect their description activity [20]. Thus, there has been a proliferation of digital archival systems based on diversified descriptive methodologies and metadata; also from the annotation point-of-view a lot of research has been done that has led to the design and development of variegated annotation systems [4].

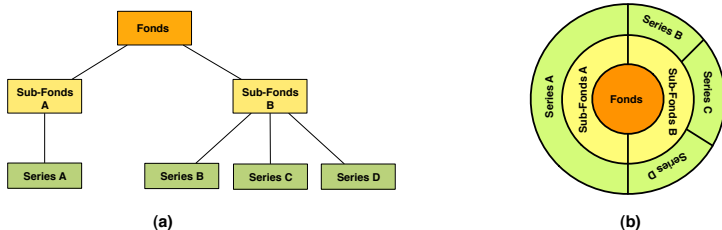
This heterogeneity turns into an interoperability problem when we need to access and consult archival metadata managed by different digital archive systems and annotations created and handled by different systems. On the other hand, every digital archive system has to respect some fundamental archival principles – i.e. the hierarchical organization of the documents and their descriptions [6]; moreover, also annotations under certain conditions can be opportunely organized in a hierarchical way [4]. We exploit these facts to define a common basis for addressing interoperability issues and for designing an integration and visualization service for annotated archives. To this end we use the *NEsted SeTs for Object hierArchies (NESTOR) Model* [8] and the *Flexible Annotation Service Tool (FAST)* annotation model [4] to:

- propose a methodology which provides us with a unified, coherent, and concise view of heterogeneous archival metadata and annotations;
- design a service allowing users to consult different archives within the relative annotations and find out the relations between different archives connected by annotations;
- develop a Web-based visualization tool based on this service which helps users to access and consult archival metadata and annotations.

The paper is organized as follows: Section 2 gives a brief background about archives, archival metadata, and annotation services highlighting the concepts we exploit in the rest of the work. Section 3 reports on the heterogeneity of archival metadata. Section 4 presents the methodology which by using the NESTOR Model and the FAST annotation model allows us to represent archives and annotations in an integrated and coherent way. Section 5 describes the proposed architecture of the integration and visualization service. In Section 6 we present the functioning of the Web-based visualization tool prototype. Finally, in Section 7 we conclude and present some future works.

## 2 Background

**Archives.** An archive is 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. Archives have to keep the context in which their records have

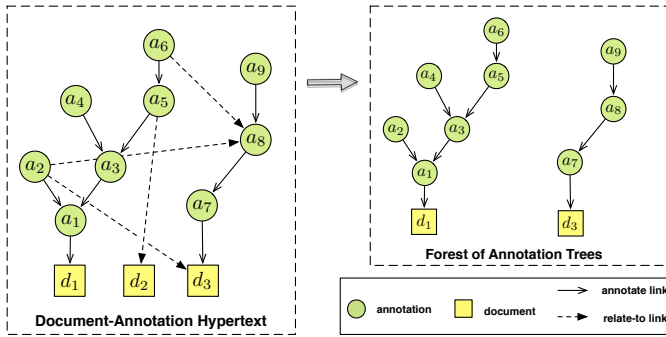


**Fig. 1.** The structure of a sample archive represented by: (a) a tree; (b) a Doc-Ball

been created and the network of relationships between them in order to preserve their informative content and provide understandable and useful information over time [10]. 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 then by sub-series and so on – see Figure 1a for an example; 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 (e.g. a fonds, a sub-fonds, etc.). 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* [12] that have to reflect the peculiarities of the archive [6].

**Digital Archives and the NESTOR Model.** 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 [10]. Archives can be modeled by means of the NESTOR Model which relies on two set data models called *Nested Set Model* (NS-M) and *Inverse Nested Set Model* (INS-M) [3]. Both these set data models, formally defined in the context of set theory, can be used to model an archive by means of nested sets [8]. An extensive analysis of the NESTOR Model and its applications in the context of DL and archives can be found in [3]; in this paper we exploit the functionalities of the INS-M and thus we focus our presentation on this model.

The most intuitive way of understanding how the INS-M works is to see how a sample tree is mapped into an organization of nested sets based on the INS-M. We can say that a tree is mapped into the INS-M by transforming each node into a set, where each parent node becomes a subset of the sets created from its children. The set created from the tree’s root is the only set with no subsets and the root set is a proper subset of all the sets in the hierarchy. The leaves are the sets with no supersets and they are sets containing all the sets created from the nodes composing the tree path from a leaf to the root. We can represent in a straightforward way the INS-M by means of the “*DocBall representation*” [17] – see Figure 1b. It is worthwhile to understand how the DocBall is used because the graphical tool we are going to present is based on this idea. The DocBall



**Fig. 2.** A document-annotation hypertext and its subgraph composed of the “annotate” edges which is a forest composed of two trees

is composed of a set of circular sectors arranged in concentric rings; each ring represents a level of the hierarchy with the center representing the root. In a ring, the circular sectors represent the nodes in the corresponding level. We use the DocBall to represent the INS-M, thus for us each circular sector corresponds to a set; for instance, referring to Figure 1b, it is possible to say that section “Series C” is a direct superset of section “Sub-Fonds B”.

**Annotations and the FAST Annotation Model.** Research on annotations has given rise to different data models, systems and services. An example is the MPEG-7<sup>1</sup> which is a standard for annotating and describing multimedia content data; the Semantic Web is another example of where annotations are exploited, in particular in the context of the Annotea project developed by the W3C<sup>2</sup>. In the context of DL an example is *Collaboratory for Annotation Indexing and Retrieval of Digitized Historical Archive Material* (COLLATE) [16], which supports the collaboration among film scientists and archivists.

Another relevant example is FAST which adopts and implements the formal model for annotations proposed in [4]. FAST distinguishes between *documents* – which are generic digital objects managed by a DL – and *annotations*. Annotations can be associated with a digital object by two types of link: **annotate link** and **relate-to link**. An annotate link allows an annotation to be linked to a part of a digital object; through this link it is possible to express *intra-digital object relationships* between different parts of an object. A relate-to link is intended to allow an annotation only to relate to one or more parts of other digital objects, but not the annotated one; therefore it expresses *inter-digital object relationships*. From these definitions annotations can be seen as a means of linking digital objects. Annotations permit us to create new relationships between the components of a digital object, between different digital objects of the same DL or between digital objects belonging to different DL. As shown in [4] the set

<sup>1</sup> Please refer to *ISO/IEC 15938-1:2002*.

<sup>2</sup> <http://www.w3.org/2001/Annotea/>



of digital objects and annotations form a labeled directed acyclic graph called document-annotation hypertext. Furthermore, each annotation must annotate only one digital object, and it has been shown [4] that for each document there is a **unique tree of annotations** constituted by “annotate” edges that can be rooted in the document. In Figure 2 we can see an example of document-annotation hypertext and the trees formed by the “annotate” links.

### 3 Heterogeneity of Archival Metadata

The standard format of metadata for representing the hierarchical structure of the archive is the *Encoded Archival Description (EAD)* [13], which reflects the archival structure and holds relations between entities in an archive. In addition, EAD has a flexible structure, encourages archivists to use collective and multilevel description, and has a broad applicability. On the other hand, the EAD permissive data model may undermine the very interoperability it is intended to foster and it must meet stringent best practice guidelines to be shareable and searchable [15]. Furthermore, 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 [7] by means of DL standard technologies like the *Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH)* [3].

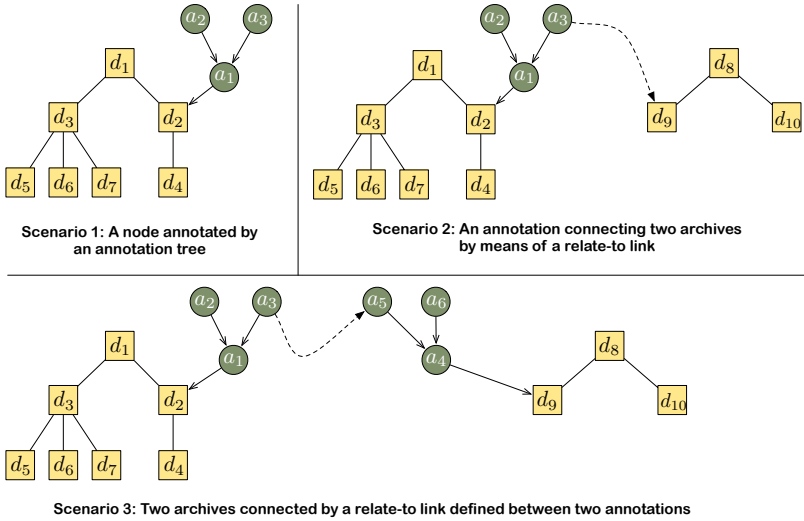
Although EAD is the archival description standard, several other modeling methodologies and metadata formats have been developed. Indeed, we may consider the “Tree-based Metadata” approach in which archives are described by a collection of lightweight metadata – e.g. Dublin Core Application Profiles [4] – one for each archival resource, connected to each other by means of links to a third-party file – e.g. an external XML file – which maintains the archival structure [14]; alternative instantiations of this approach maintain the archival structure by means of an opportunely designed relational database [15]. These approaches differ from EAD both in the way in which they express the structure and the content of the archive. Furthermore, outside EAD boundaries, there is no common agreement on which metadata fields should be used to describe archival resources.

There is also the possibility of representing the archival structure by means of the INS-M [7]. It has been shown [8] that an archive can be modeled by means of the INS-M and then instantiated in such a way that allows the use of the OAI-PMH architecture to enable a variable granularity access and exchange of the archival metadata. Furthermore, [7] describes a methodology to map an EAD file into the NESTOR Model and preserve the full informative power of the metadata. Mapping an EAD file into the NESTOR Model means that we make use of a methodology that maps the EAD structure into the INS-M and a collection of lightweight metadata containing the content information retained by EAD. In this way the INS-M preserves the archival structure while the metadata

---

<sup>3</sup> <http://www.openarchives.org/>

<sup>4</sup> <http://www.dublincore.org/>



**Fig. 3.** Annotations: Three possible scenarios in the archival context

belonging to its sets preserve the content of archival descriptions [7]. In the same way, this methodology is adopted with the “Tree-based metadata” approach, where the structure retained by an external XML file or by a relational database is mapped into the INS-M [3].

## 4 An Integrated View of Archives and Annotations

Our goal is to make available a uniform and integrated view of archives described and managed by means of heterogeneous digital archive systems together with their annotations which in turn can be handled by different annotation systems. To this purpose we rely on the NESTOR Model and on the FAST annotation model to address interoperability at the archival level and to show how annotations can be enclosed in the “NESTOR view” of the archives.

We present three possible scenarios showing how annotation trees can be attached to an archive and then we show how they can be modeled through the INS-M and represented by means of the DocBall. Figure 3 presents the scenarios; in this figure an archive is represented as a document tree where the nodes are named as “ $d_1, d_2, \dots$ ” for convenience; for the same reasons annotations are indicated as “ $a_1, a_2, \dots$ ”. In the first scenario we consider an archival tree where the node  $d_2$ , annotated by  $a_1$ , is the root of an annotation tree composed of three annotations. The second scenario shows that  $a_3$  which is part of an annotation tree annotating  $d_2$  is connected to a second archive by means of a “relate-to” link. In the third scenario, we can see two archives connected by a relate-to link defined between two annotations – i.e. a relate-to link between  $a_3$  and  $a_5$ . Figure 4 shows by means of the DocBall representation how these scenarios are handled by the INS-M; we adopt the DocBall as a graphical means to describe

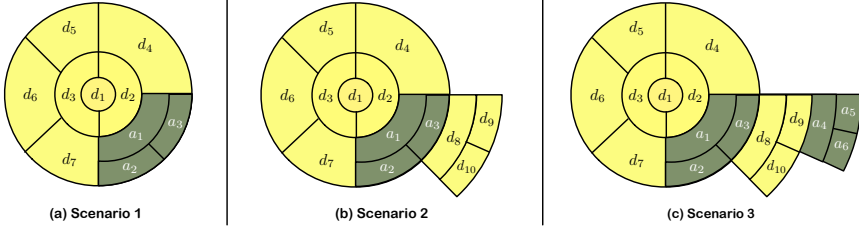


Fig. 4. DocBall representations of archive annotation scenarios

and explain how archives and annotations are joined together by means of the INS-M.

In the first scenario we need to join an “archival DocBall” representing the archive and an “annotation DocBall” representing the annotation tree originally attached to node  $d_2$  of the archive – see Figure 3a. The resulting DocBall is shown in Figure 4a, where  $a_1$  is a superset of  $d_2$ . The second scenario presents the same annotated archive we have seen in the first scenario enriched by the relationship of annotation  $a_3$  with the node  $d_9$  of a second archive. In this case, we use a DocBall representing the first archive within its annotations – call it “DocBall A” (see Figure 4a) – and a DocBall representing the second archive – call it “DocBall B”. In order to join these two DocBalls connected by annotation  $a_3$ , we add the inner sector of DocBall B – i.e.  $d_8$  – to DocBall A as a superset of  $a_3$ . The resulting DocBall (see Figure 4b) provides us with of an integrated view of the two archives connected by the annotation tree rooted in  $a_1$ . The third scenario enhances this idea; indeed, in this case both “DocBall A” and “DocBall B” represent annotated archives that have to be joined together. So, we follow the methodology presented for scenario 2 by taking the inner sector of DocBall B – i.e.  $d_8$  which represents the root of the second archive – and adding it to DocBall A as a superset of the annotation – in this case  $a_3$  – which relates the two archives together. The general methodology of joining two DocBall together can be summarized as follows; let  $D_A$  and  $D_B$  be two DocBall, where section  $s_A$  of  $D_A$  is related to section  $s_B$  of  $D_B$ . To join  $D_A$  with  $D_B$ , the inner section of  $D_B$  must be added to  $D_A$  as a superset of  $s_A$ .

## 5 Architecture and Functionalities of the Integration and Visualization Service

In order to accomplish the purposes of this work, the integration and visualization service must be non-intrusive, scalable and flexible. Indeed, it has to be *non-intrusive* to model the archives and annotations by means of the INS-M without interfering with the organization and the functioning of the local systems. It has to be *scalable* to collect resources in a distributed environment, and it has to be *flexible* to integrate archives and annotations together satisfying user needs.

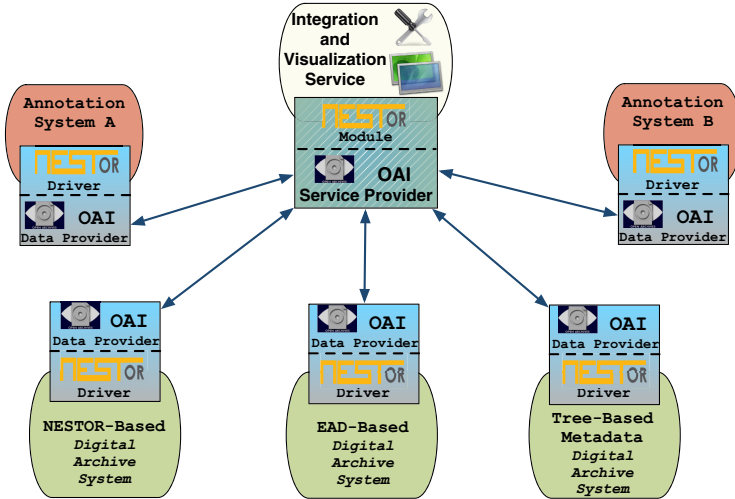
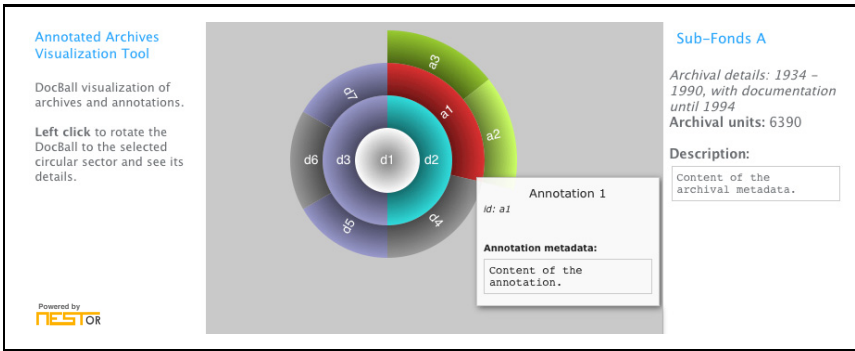


Fig. 5. Proposed architecture of the integration and visualization service

In Figure 5 we can see the proposed architecture of the integration and visualization service. We consider three different digital archive systems: the first is based on the NESTOR model, the second on EAD and the third on the “Tree-based metadata” approach. Furthermore, we consider two generic annotation systems: “Annotation system A and B”. Each digital archive and annotation system should be equipped with a software module divided into two main components. The first component is called “NESTOR driver” and the second is an OAI Data Provider. The NESTOR driver is a lightweight component that has to map the archival metadata into the INS-M and prepares them to be exchanged by means of OAI-PMH. If we consider the NESTOR-based system in Figure 5, the NESTOR driver has to check if the archival metadata are modeled by means of NS-M or INS-M and in the first case it has to map the archive from the NS-M into the INS-M [8]. For the EAD-based archive system, the NESTOR driver has to map the EAD files into the INS-M [7] and in the “Tree-based metadata” system it has to map the XML file or the relational schema preserving the archive structure into the INS-M [3]. The NESTOR driver does the same operations with the annotation trees by mapping them into the INS-M [9].

In this way the NESTOR driver addresses the heterogeneity between different digital archive and annotation systems in a non-invasive and transparent way: the local systems handle archives and annotations within their own policies and expose them coherently with the INS-M. Furthermore, we know that the sets and the metadata defined by the INS-M can be straightforwardly exchanged by means of OAI-PMH [8]. Thanks to this feature we can exploit the OAI-PMH architecture to exchange the archival metadata and the annotations between the local systems and the centralized integration and visualization service. As we can see in Figure 5, the NESTOR driver is configured as a component of an



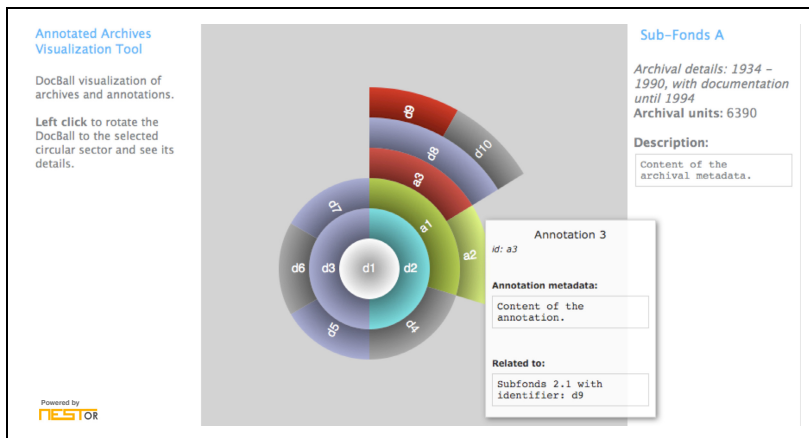
**Fig. 6.** Prototype of the visualization Tool: DocBall representation of Scenario 1

OAI Data Provider; in this way the presented architecture draws on OAI-PMH scalability and flexibility and the NESTOR driver can be configured as a plug-in of already existing and widely-diffused software modules.

The integration and visualization service can be developed over an OAI Service Provider which harvests the archival metadata and the annotations. The service utilizes a “NESTOR module” that acts as a mediator between the requests of the service and the harvested metadata and annotations. According to the three scenarios presented in the previous section, if the service requires just the archival metadata together with their annotations – i.e. scenario 1 – the NESTOR module embeds the archive with its annotations and returns the INS-M represented by the DocBall in Figure 4a. The NESTOR module returns a DocBall like the ones in Figure 4b and 4c when the service needs to exploit the relationships established by the annotations between different archives. The role of the visualization tool is to enhance the relationships between an archive and its annotations and between different archives connected by annotations. Especially in the second and third scenarios, the visualization tool needs to have an effective interface to help the users to infer and exploit the relationships between the resources.

## 6 Web-Based Visualization Tool

The visualization tool is the front-end component of the integration and visualization service; it relies on the archives and annotations modeled by means of the INS-M. We show and discuss several screenshots of the initial prototype of the visualization tool based on test data; Figure 6 shows how the service addresses the first scenario. We can see that the DocBall is similar to the one in Figure 4a and it shows an archive where section  $d_2$  is annotated by an annotation tree composed of three annotations. In the left column we have general information about the service. The DocBall is in the center of the canvas and when we move the pointer over a circular section a tooltip appears showing the content of this section; if we click on a section, the DocBall rotates and the selected section is



**Fig. 7.** Prototype of the visualization Tool: DocBall representation of Scenario 2

highlighted. In this figure we selected section  $d_2$  the content of which is shown in the right column and the tooltip shows the content of  $a_1$ . In this way the user can select an archival section, see its content in the right column and view the content of annotations or other archival divisions by means of the tooltip.

Figure 7 shows a screenshot where the visualization tool addressed the second scenario; the annotation ( $a_3$ ) which annotates an archival section ( $d_2$ ) is related to the archival section ( $d_9$ ) of a second archive. The tool highlights the related sections; indeed, when  $d_2$  is selected, the DocBall rotates in such a way that its annotation tree moves on the top of the DocBall, and annotation  $a_3$  together with section  $d_9$  are colored in red revealing the connection between them. The user can explore the content of these sections by means of the tooltip while visualizing the content of  $d_2$  in the right column. In Figure 7 we captured the tooltip related to  $a_3$ ; we can see that it reports the content of the annotation and the information about its relationship with section  $d_9$ .

Figure 8 shows the last scenario where we exploit the relationship between two annotations – i.e.  $a_3$  and  $a_5$  – to relate two different archives. The service works as in the second scenario but in this case it highlights the two annotations; the user can visualize the content of the annotations of the first and second archive as well as the content of the second archive contextually with the content of the selected archival section.

We can see that archival documents and annotations are represented as circular sectors with different colors in the DocBall. The use of colors may be an effective way to distinguish between the sectors which are documents and those which are annotations. Furthermore, the DocBall could become ineffective if there are many sectors that have to be represented. In this case an expand/compress strategy could be adopted as well as it is used to show the branches of very large trees.

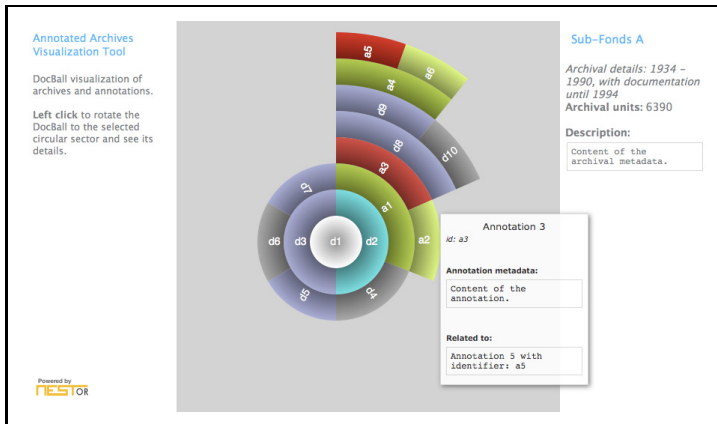


Fig. 8. Prototype of the visualization Tool: DocBall representation of Scenario 3

## 7 Conclusion

In this paper we propose the architecture of an integration and visualization service that exploits the NESTOR Model and the FAST annotation model to provide us with a unified view of archives and annotations that can come from diversified systems. This service can address interoperability issues between different digital archive systems and annotation systems in a flexible and scalable way by exploiting existing and widely-diffused software modules – i.e. OAI Data and Service Provider – and extending them by means of lightweight software modules – i.e. the NESTOR driver. The presented prototype of the service enables a comprehensive view of archival structure and content together with its annotations; furthermore, it highlights the relationships between different archives. This service can enhance the role of annotations in the archival context and the expertise of archivists in the description as well as in the consultation phase of the archives.

Future work foresees the adoption of this service in the context of a project of the Italian Veneto Region<sup>5</sup>. The main aim of the project is to make available a regional archival information system which allows the management of the resources of archives present in the Region.

**Acknowledgments.** The work reported has been carried out in the context of an agreement between the Italian Veneto Region and the University of Padua. CULTURA<sup>6</sup> (Grant agreement no. 269973) and the PROMISE network of excellence<sup>7</sup> (contract n. 258191) projects, as part of the 7th Framework Program of the European Commission, have partially supported the reported work.

<sup>5</sup> <http://www.regione.veneto.it/>

<sup>6</sup> <http://www.cultura-strep.eu/>

<sup>7</sup> <http://www.promise-noe.eu/>

## References

1. Christensen-Dalsgaard, B., Castelli, D., Ammitzbøll Jurik, B., Lippincott, J. (eds.): ECDL 2008. LNCS, vol. 5173. Springer, Heidelberg (2008)
2. Agosti, M., Esposito, F., Thanos, C. (eds.): IRCDL 2010. CCIS, vol. 91. Springer, Heidelberg (2010)
3. 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)
4. Agosti, M., Ferro, N.: A Formal Model of Annotations of Digital Content. *ACM Trans. Inf. Syst.* 26(1) (2007)
5. Agosti, M., Ferro, N.: Annotations: A Way to Interoperability in DL. In: [1], pp. 291–295
6. Duranti, L.: *Diplomatics: New Uses for an Old Science*. Society of American Archivists and Association of Canadian Archivists in association with Scarecrow Press, Lanham, Maryland, USA (1998)
7. Ferro, N., Silvello, G.: A Methodology for Sharing Archival Descriptive Metadata in a Distributed Environment. In: [1], pp. 268–279
8. Ferro, N., Silvello, G.: The NESTOR Framework: How to Handle Hierarchical Data Structures. In: Agosti, M., Borbinha, J., Kapidakis, S., Papatheodorou, C., Tsakonias, G. (eds.) ECDL 2009. LNCS, vol. 5714, pp. 215–226. Springer, Heidelberg (2009)
9. Ferro, N., Silvello, G.: FAST and NESTOR: How to Exploit Annotation Hierarchies. In: [2], pp. 55–66
10. 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 (2000)
11. Kani-Zabihi, E., Ghinea, G., Chen, S.Y.: Experiences with Developing a User-Centered Digital Library. *IJDL* 1(1), 1–23 (2010)
12. Pearce-Moses, R.: *Glossary of Archival And Records Terminology*. Society of American Archivists (2005)
13. Pitti, D.V.: Encoded Archival Description. An Introduction and Overview. *D-Lib Magazine* 5(11) (1999)
14. Prom, C.J., Habing, T.G.: Using the Open Archives Initiative Protocols with EAD. In: Proc. 2nd ACM/IEEE Joint Conf. on Digital Libraries, pp. 171–180. ACM Press, USA (2002)
15. Prom, C.J., Rishel, C.A., Schwartz, S.W., Fox, K.J.: A Unified Platform for Archival Description and Access. In: Proc. 7th ACM/IEEE Joint Conf. on Digital Libraries, pp. 157–166. ACM Press, USA (2007)
16. Thiel, U., Brocks, H., Frommholz, I., Dirsch-Weigand, A., Keiper, J., Stein, A., Neuhold, E.J.: COLLATE - A Collaboratory Supporting Research on Historic European Films. *Int. J. on Digital Libraries* 4(1), 8–12 (2004)
17. Vegas, J., Crestani, F., de la Fuente, P.: Context Representation for Web Search Results. *Journal of Information Science* 33(1), 77–94 (2007)
18. Vitali, S.: Archival Information Systems in Italy and the National Archival Portal. In: [2], pp. 5–11
19. Yakel, E., Shaw, S., Reynolds, P.: Creating the Next Generation of Archival Finding Aids. *D-Lib Magazine* 13(5/6) (May/June 2007)
20. Yako, S.: It's Complicated: Barriers to EAD Implementation. *American Archivist* 71(2) (Fall/Winter 2008)



# Metadata Inference for Description Authoring in a Document Composition Environment

Tsuyoshi Sugibuchi, Ly Anh Tuan, and Nicolas Spyratos

Laboratoire de Recherche en Informatique, Université Paris-Sud 11, France  
{Tsuyoshi.Sugibuchi,Anh.Tuan.Ly,Nicolas.Spyratos}@lri.fr

**Abstract.** In this paper, we propose a simple model for metadata management in a document composition environment. Our model considers (1) composite documents in the form of trees, whose nodes are either atomic documents, or other composite documents, and (2) metadata or *descriptions* of documents in the form of sets of terms taken from a taxonomy. We present a formal definition of our model and several concepts of inferred descriptions. Inferred descriptions can be used for *term suggestion* that allows users to easily define and manage document descriptions by taking into account what we call *soundness* of descriptions.

## 1 Introduction

Today's growth of digital publishing is bringing about not only media migration from atom to bit, but also more flexibility in authoring and customizing digital documents *after* their publication. For example, several non-profit projects and commercial companies start to offer *open textbook* platforms that intend to allow textbook authors, educators and students to create and customize textbooks. An interesting example is the *Connexions project* [1] funded by Rice University. In the Connexions' repository, every textbook is managed as a collection of individual learning objects called *modules*. The Connexions' website allows users not only to read textbooks but also to create and customize textbooks by composing modules taken from a variety of existing textbooks.

To make a new textbook by composing fragments of existing textbooks, authors need to find appropriate fragments from textbook repositories. At present, most open textbook platforms adopt description based document management. In such systems, each document and its fragments are associated with their descriptions, also called *metadata*. Usually metadata contains free-text information including title, short description and free keywords, and information based on controlled vocabularies, or *taxonomies*, including subject category, topic group, etc. Information based on controlled vocabularies is useful for more accurate and intelligent content retrieval, if metadata is properly created and maintained.

If we intend to allow users to take fragments from textbooks with smaller granularity, the cost of authoring many metadata for each textbook fragment might be a problem. A clue for reducing such metadata authoring cost is the

fact that each fragment of a textbook is usually part of a bigger context. For instance, a section of a textbook usually has a previous or next siblings, and a parent chapter that encloses child sections. By taking into account such relationships among fragments, we can infer metadata of new fragments from the metadata of existing fragments. Metadata of textbook fragments should be manually made by human-beings, but machines can also “suggest” inferred metadata. Such metadata suggestions will help users in easily making metadata.

In this paper, we propose a simple metadata management model for document composition environments. Our work is based on the metadata inference model for composite documents proposed in [2]. The model described in [2] mainly focuses on document *sharing*. In the present paper, we focus on the actual usage of metadata for *authoring* document descriptions. We give the formal definition of our metadata model and demonstrate how it can be used to suggest terms for descriptions based on descriptions of existing documents.

In the rest of this paper we first review some related studies (Section 2). Then we describe our metadata model and some algorithms for inferring metadata (Section 3). Based on this model, we introduce a criterion that every description should satisfy, and then we explain how our “term suggestion” by using this criterion (Section 4).

## 2 Related Work

A lot of efforts have been devoted recently to develop languages and tools to generate, store and query metadata. Some of the most noticeable achievements are the RDF language, RDF schemas and several standards for representing controlled vocabulary including OWL [3] and SKOS [4]. By using such languages and standards, several controlled vocabularies for metadata have been developed and are widely used in practice. These vocabularies include Gene Ontology [5] (genomics), AAT [6] (arts and architectures), DBPedia Ontology [7] (cross-domain ontology) and others. Most of these vocabularies are structured as general graphs including cycles. Even then most of these vocabularies also include hierarchically organized “is-a” relationships of terms. In this paper, we focus on taxonomy-based annotations [8] to describe the content by using such hierarchically organized sets of terms. Generation of such annotations still remains mostly a manual process, possibly supported by acquisition software (for instance [9]). Many of such annotation supports are performed by text analysis techniques (for instance [10]) and some researches deal with *annotation propagation* to infer metadata of derived contents from those of the original based content authoring processes [11][12]. The work in [2] which is the basis of our study also proposes a metadata inference model for composite documents. However, the inference model of [2] is mainly intended for document repository management. In contrast, the inference model that we propose here is intended for document description authoring, including creation and modification.

### 3 The Model of Composite Documents and Descriptions

#### 3.1 Documents and Composite Documents

First of all, our model does not consider contents of documents. Our model deals only with structures of document composition and document descriptions. Therefore, we focus only on a document representation consisting of an identifier and a set of *parts*, as this is sufficient for our metadata management. Therefore, hereafter, when we talk of a document we shall actually mean its representation by an identifier and a set of parts.

**Definition 1 (The representation of a document).** A document consists of an identifier  $d$  together with a set of documents, called the *parts* of  $d$  and denoted as  $parts(d)$ . If  $parts(d) = \emptyset$  then  $d$  is called *atomic*, else it is called *composite*.

For notational convenience, we shall often write  $d = d_1 + d_2 + \dots + d_n$  to stand for  $parts(d) = \{d_1, d_2, \dots, d_n\}$ . Based on the concept of parts, we can now define the concept of *component*.

**Definition 2 (Components of a document).** Let  $d = d_1 + d_2 + \dots + d_n$ . The set of *components* of  $d$ , denoted as  $comp(d)$ , is defined recursively as follows:

- if  $d$  is atomic, then  $comp(d) = \emptyset$
- else  $comp(d) = parts(d) \cup comp(d_1) \cup comp(d_2) \cup \dots \cup comp(d_n)$ .

In this paper, we assume that every composite document  $d$  is a tree in which  $d$  is the root and  $comp(d)$  is the set of nodes. Our choice is justified by the fact that (1) the tree is the most suitable structure for representing traditional books that are hierarchically organized, and (2) the tree is also a common structure adopted by many existing document composition environments including open textbook platforms. Based on this assumption, for a composite document  $d$  and its part  $d' \in parts(d)$ ,  $d'$  is called *child* of  $d$ , and  $d$  is called *parent* of  $d'$ , denoted as  $parent(d')$ . It is important to note that in our model the ordering of parts in a composite document is ignored because it is not relevant to our purposes. As we shall see shortly, deriving the description of a composite document from the descriptions of its parts does not depend on any ordering of the parts.

#### 3.2 Taxonomy and Description

Informally, descriptions in our model are just sets of terms taken from a taxonomy. We would like to start our explanation about descriptions from the formal definition of taxonomy in our model.

**Definition 3 (Taxonomy).** Let  $T$  be a set of keywords, or *terms*. A *taxonomy*  $\mathcal{T}$  defined over  $T$  is a tuple  $(T, \preceq)$  where  $\preceq$  is a reflexive and transitive binary relation over  $T$ , called *subsumption relation*.

Given two terms,  $s$  and  $t$ , if  $s \preceq t$  then we say that  $s$  is *subsumed* by  $t$ , or that  $t$  *subsumes*  $s$ . In our work, we assume that every taxonomy  $(T, \preceq)$  is a tree in which the nodes are the terms of  $T$  and where there is an arrow  $s \rightarrow t$  iff  $s$

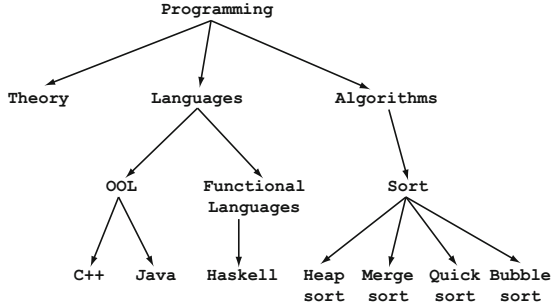


Fig. 1. A taxonomy

subsumes  $t$  in  $\preceq$ . Fig 1 shows the example taxonomy  $\mathcal{T}_p$  we use in this paper. In this example, the term **Sort** subsumes the term **Quick sort**, **OOL** subsumes **Java** and **C++**. Due to the transitivity of the subsumption relation, the term **Programming** subsumes all terms in the tree including itself. In the rest of this paper, we use a symbol  $tail(t)$  that stands for the set of all terms in the taxonomy strictly subsumed by  $t$ , i.e.,  $tail(t) = \{s \mid s \prec t\}$ .

In order to make a document sharable, a description of its content must be provided, so that users can judge whether the document in question matches their needs. Our model allows any sets of terms from a taxonomy as descriptions.

**Definition 4 (Description).** Given taxonomy  $(T, \preceq)$ , we call *description* in  $T$  any set of terms from  $T$ .

### 3.3 Inferred Descriptions

**Reduction of a Description.** A description can be redundant if some of the terms it contains are subsumed by other terms in the description. For instance, the description  $\{\text{Sort}, \text{Quick sort}, \text{java}\}$  is redundant, as **Sort** subsumes **Quick sort**. Redundant descriptions are sometimes undesirable as they can lead to redundant computations. Now we introduce the concept of non-redundant, or *reduced descriptions*, defined as follows:

**Definition 5 (Reduced description).** Given taxonomy  $(T, \preceq)$ , a set of terms  $D$  from  $T$  is called *reduced* if for any terms  $s$  and  $t$  in  $D$ ,  $s \not\preceq t$  and  $t \not\preceq s$ .

Following the above definition, we can make a description non-redundant by either removing all but its minimal terms, or by removing all but its maximal terms. We shall assume the former as it produces more accurate descriptions. This should be clear from our previous example, where the description  $\{\text{Quick sort}, \text{Java}\}$  is more accurate than  $\{\text{Sort}, \text{Java}\}$ . Hence the following definition:

**Definition 6 (Reduction).** Given a description  $D$  in taxonomy  $(T, \preceq)$ , we call *reduction* of  $D$ , denoted as  $reduce(D)$ , the set of minimal terms in  $D$  with respect to the subsumption  $\preceq$ .

The important point to note is that even if a description created by an author contains redundancy, our model respects the original form of the description and does not remove anything from the description. The choice of terms to include in a description is left entirely up to description authors. Reduction of descriptions and other concepts of inferred descriptions are made only internally, or for suggesting “hints” for users to create descriptions with less effort.

Therefore we distinguish between descriptions created by authors and descriptions inferred automatically by using algorithms. For a document  $d$ , the former type of description is called the *author description* of  $d$ , denoted as  $ADescr(d)$ : author descriptions are exactly the descriptions that authors create and a document repository stores. The latter type of description is what is generated internally, by machines, for helping authors in description authoring.

Additionally, we would like to introduce two more concepts of inferred descriptions.

**Cover of a Document.** To make a description of a composite document, it is sometimes useful to know all topics covered by the components of the document. Now we introduce the concept of *cover* of a document  $d$  that is an inferred description formed with a minimum set of terms and semantically covers all terms appearing in descriptions of  $d$ 's components. The cover of a document is formally defined as follows:

**Definition 7 (Cover of a document).** Given a document  $d$ , the *cover* of  $d$ , denoted as  $cover(d)$ , is a description recursively defined as follows:

- if  $d$  is atomic,  $cover(d) = reduce(ADescr(d))$ ,
- else, for  $d = d_1 + \dots + d_n$ ,  $cover(d) = reduce(cover(d_1) \cup \dots \cup cover(d_n))$ .

Informally, the cover of a document is the minimum but most accurate description of the document. To create the description of a document, authors should choose terms present in the cover of the document, or terms subsuming at least one term in the cover. Otherwise the created description might contain terms not related to any component of the described document.

**Summary of a Document.** On the other hand, sometimes we want to summarize topics of a big composite document. There are several possible approaches for summarization. One intuitive approach is to extract common topics shared by all components of a document. Suppose a composite document  $d = d_1 + d_2$  such that  $ADescr(d_1) = \{\text{Quick sort, Java}\}$  and  $ADescr(d_2) = \{\text{Bubble sort, C++}\}$ . In this case,  $D_{sum} = \{\text{Sort, OOL}\}$  is a possible summary of  $d_1$  and  $d_2$ . **Sort** subsumes both **Quick sort** and **Bubble sort**. **OOL** also subsumes both **Java** and **C++**. As the result,  $\{\text{Sort, OOL}\}$  represents what  $d_1$  and  $d_2$  have in common.

In this example,  $D'_{sum} = \{\text{Algorithms, Languages}\}$  is also a possible summary. However,  $D'_{sum}$  is less accurate than  $D_{sum}$ . The most extreme example is  $D^*_{sum} = \{\text{Programming}\}$ .  $D^*_{sum}$  summarizes any descriptions in  $T$  but with lowest accuracy. Usually such over-general summary is useless for document search.

Now we informally define the *summary of a document* as a description such that (1) it summarizes what all components of a document have in common in their descriptions and (2) it is minimal, in other words, has highest accuracy.

The examples of  $D'_{sum}$  and  $D^*_{sum}$  violate the second criterion because they have lower accuracy than  $D_{sum}$ .

In order to formalize this definition, we introduce the following *refinement relation on descriptions*.

**Definition 8 (Refinement relation).** Let  $D_1$  and  $D_2$  be two descriptions. We say that  $D_1$  is *finer* than  $D_2$ , denoted  $D_1 \sqsubseteq D_2$ , iff  $\forall t_2 \in D_2, \exists t_1 \in D_1 \wedge t_1 \preceq t_2$ .

For example,  $D_{sum}$  is finer than  $D'_{sum}$ , i.e.,  $D_{sum} \sqsubseteq D'_{sum}$  because for every term  $t$  in  $D'_{sum}$ , we can find a term in  $D_{sum}$  subsumed by  $t$  such as in our example, where  $\text{Sort} \preceq \text{Algorithms}$  and  $\text{OOL} \preceq \text{Languages}$ .

The refinement relation  $\sqsubseteq$  is clearly reflexive and transitive. Moreover, over reduced descriptions  $\sqsubseteq$  becomes antisymmetric. From these properties of  $\sqsubseteq$ , we can say that  $\sqsubseteq$  is a partial order over reduced descriptions, and a set of reduced descriptions has a least upper bound in  $\sqsubseteq$ . Here we omit the detail and just introduce the following proposition and theorem. For detailed discussion and proofs of them, see [2].

**Proposition 1.** The relation  $\sqsubseteq$  is a partial order over the set of all reduced descriptions.

**Theorem 1.** Let  $\mathcal{D} = \{D_1, \dots, D_n\}$  be any set of reduced descriptions. Let  $\mathcal{U}$  be the set of all reduced descriptions  $S$  such that  $D_i \sqsubset S, i = 1, \dots, n$ , i.e.,  $\mathcal{U} = \{S \mid D_i \sqsubset S, i = 1, \dots, n\}$ . Then  $\mathcal{U}$  has a least upper bound, that we shall denote as  $\text{lub}(\mathcal{D}, \sqsubseteq)$ .

The least upper bound (*lub*) of descriptions is the most accurate set of terms representing what the descriptions have in common. Therefore, by obtaining the *lub* of descriptions of documents, we can get the most accurate description that summarizes what the documents have in common. By using this theorem, we can now define the summary of a document as following:

**Definition 9 (Summary of a document).** Given a document  $d$ , the *summary* of  $d$ , denoted as  $\text{summary}(d)$ , is a description defined as follows:

- if  $d$  is atomic,  $\text{summary}(d) = \text{reduce}(A\text{Descr}(d))$ ,
- else, for  $d = d_1 + \dots + d_n$ , let  $\mathcal{D} = \{\text{summary}(d_1), \dots, \text{summary}(d_n)\}$ ,  $\text{summary}(d) = \text{lub}(\mathcal{D}, \sqsubseteq)$ .

The algorithm `summary` illustrated in Fig. 2 recursively computes the summary of a given document. We shall use these algorithms in the next section for helping authors to make descriptions of new documents.

## 4 Metadata-Aided Suggestion in Document Description Authoring

In this section, we would like to explain how we can use inferred descriptions of documents to help users to create and manage document descriptions. As we already mentioned, the author description of a document is left entirely up

**Algorithm summary**

```

Input a document  $d$ 
Output  $\text{summary}(d)$ 
if  $d$  is atomic then
  return  $\text{reduce}(ADescr(d))$ 
end if
for all  $d_i \in \text{parts}(d), i = 1, \dots, n$  do
   $D_i \leftarrow \text{summary}(d_i)$ 
end for
 $P \leftarrow D_1 \times D_2 \times \dots \times D_n$ 
for all  $L_k = [t_1^k, \dots, t_n^k] \in P, k = 1, \dots, l$  do
   $T_k \leftarrow \text{lub}_{\preceq}(t_1^k, \dots, t_n^k)$ 
end for
return  $\text{reduce}(T_1, \dots, T_l)$ 

```

$\text{lub}_{\preceq}(t_1, \dots, t_n)$  returns the least upper bound of the set of terms  $t_1, \dots, t_n$  with respect to  $\preceq$ .

**Algorithm cover**

```

Input a document  $d$ 
Output  $\text{cover}(d)$ , or false if  $d$  contains doc-
uments that have descriptions not satisfying
soundness.
if  $d$  is atomic then
  return  $\text{reduce}(ADescr(d))$ 
end if
 $C \leftarrow \emptyset$ 
for all  $d' \in \text{parts}(d)$  do
   $C' \leftarrow \text{cover}(d')$ 
  if  $C' = \text{false}$  then return false end if
   $C \leftarrow C \cup C'$ 
end for
for all  $t \in ADescr(d)$  do
  if there is no  $t' \in C'$  such that  $t' \preceq t$  then
    return false
  end if
end for
return  $\text{reduce}(C)$ 

```

**Fig. 2.** summary algorithm and cover algorithm

to description authors. Therefore, the algorithms explained in the previous section are not intended to *generate* descriptions of documents automatically. The purpose of the algorithms is to *suggest* inferred descriptions to avoid making descriptions from scratch. Before entering into details of our suggestion process, we would like to discuss two preliminary topics.

**Soundness of Descriptions.** While authors can freely choose any terms to make descriptions, are there any criteria that descriptions should satisfy? Our opinion is that every description should satisfy some kind of “soundness”. If a description contains a term  $t$ , the described document should contain something related to the term  $t$ . Now we formalize soundness of a description as follows:

**Definition 10 (Soundness of a document description).** A description  $D$  of document  $d$  is called *sound* if  $d$  and  $D$  satisfy the following condition:

- For every term  $t \in D$ , at least one term  $t' \in \text{cover}(d)$  is subsumed by  $t$ , or
- $d$  is atomic

We should comment on the last part of this definition. As we mentioned several times, our model does not deal with contents of documents. Consequently, our model has no way to determine whether an author description of an atomic document satisfies soundness or not with respect to the document content. Therefore, we firstly *believe* it. Our model depends on an assumption that all descriptions of atomic documents satisfy soundness.

In the rest of this paper, we would like to adopt the *soundness criterion* that requires every author description to be sound. We think it is a reasonable criterion for keeping integrity of a document repository. Without this criterion, a document repository might have an untrustworthy description that contains terms not related to any parts of a described document.

The soundness of a document description is defined over the cover of a document. Therefore, the algorithm `cover` illustrated in Fig. 2 can validate soundness of descriptions and compute cover of descriptions in parallel by using simple

depth-first search. We also use a symbol  $dtDom(d)$  that stands for the set of all terms in  $T$  such that we can use for describing a composite document  $d$ , i.e.,  $dtDom(d) = \{t | t \in T \wedge \exists t' \in cover(d) \wedge t' \preceq t\}$ .

**Types of Terms for Suggestion.** Briefly, suggestion is an activity to indicate a *suggestion list* of choices to users for allowing them to easily specify input values. In this paper, we do not deal with details of user interaction design. Here we would like to just classify terms for suggestion into the following three different types of term sets by how terms are initially selected in a suggestion list and what users can do on terms in a suggestion list.

- $D_{rec}$  (**recommended**): All terms in this set are selected as default and users can remove terms from the set
- $D_{opt}$  (**optional**): All terms in this set are not selected as default and users can add terms to the set
- $D_{obso}$  (**obsolete**): All terms in this set are not selected as default and users cannot select any of them.

$D_{rec}$  typically contains terms that affect summary of descriptions. By removing terms in  $D_{rec}$ , authors can *generalize* descriptions to give broader meanings.  $D_{opt}$  typically contains terms that do not affect the summary but can be used for descriptions. By selecting terms in  $D_{opt}$ , authors can *specialize* descriptions to give narrower meanings. In many cases, the size of  $D_{opt}$  is very big. Therefore, sometimes we partition  $D_{opt}$  into a sequence of disjoint subsets  $D_{opt1}, D_{opt2}, \dots$  by priorities of terms.  $D_{obso}$  contains terms that violate the soundness criterion of descriptions.  $D_{obso}$  is used for indicating what must be removed from descriptions to preserve soundness. Authors cannot change selections of terms in  $D_{obso}$ . Authors can only accept removing indicated terms from descriptions.

#### 4.1 Create New Atomic Documents

When an author creates a new atomic document as an independent one, the author needs to choose terms to define its description by taking into account the document content. However, if an author writes a new atomic document as a part of an existing composite document, we can suggest terms for its description by taking into account the descriptions of existing components.

Let  $d_p$  be a composite “parent” document. Suppose an author has created a new atomic document  $d$  as a child of  $d_p$  and now he is going to define an author description  $ADescr(d)$ . Firstly, if a parent document has its summary, the same summary should also be a “sound” description of a new child document. In this case, the author should define an author description  $ADescr(d)$  such that  $ADescr(d) \sqsubseteq summary(d_p)$ .

See the example illustrated in Fig. 3. In this example, the summary  $D_{sum}$  of  $d_p$  is  $\{\text{Sort, Java}\}$ . If the author defines  $ADescr(d)$  as  $\{\text{Sort, Java}\}$  or  $\{\text{Quick sort, Java}\}$ ,  $D_{sum}$  does not change. On the other hand, if the author defines  $ADescr(d)$  as  $\{\text{Sort}\}$ ,  $D_{sum}$  will be changed to  $\{\text{Sort}\}$  that has less accuracy. It is the absence of **Java** from  $ADescr(d)$  that causes such change of the document summary. Therefore, even if the author wants to remove this



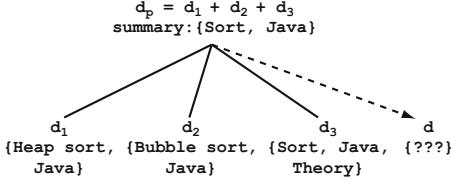


Fig. 3. Creation of an atomic document

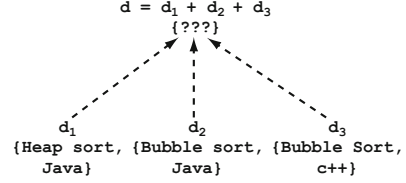


Fig. 4. Creation of a composite document

term from the description, it should be *intentionally* removed. Therefore, terms in  $D_{sum}$  should be suggested as members of  $D_{rec}$ . Because terms in  $D_{rec}$  are selected as default in a suggestion list, then users can *manually* remove them.

Secondly, a description of a new child document should be more specialized than the summary of its parent. We can perform such specialization (1) by specializing a part of the summary of a parent, or (2) by adding new terms in  $T$ . Regarding point (1), we can extract promising candidates of terms for specialization by comparing descriptions of *siblings*. Let's go back to the example in Fig. 3. In this example, siblings of  $d$  have {Heap sort, Java}, {Bubble sort, Java}, {Sort, Java, Theory} as their descriptions. Taking  $D_{sum} = \{\text{Sort, Java}\}$  into account, we can see that **Sort** is specialized in descriptions of some siblings but **Java** is not. Therefore, tails of **Sort**, for instance **Quick sort** and **Merge sort**, have higher priority for specialization than tails of **Java**. We use a symbol  $specializedIn(d)$  that stands for the set of all terms in  $summary(d)$  that are specialised in some descriptions of  $d$ 's parts, i.e.,  $specializedIn(d) = \{t | t \in summary(d) \wedge \exists d' \in part(d) \wedge \exists t' \in ADescr(d') \wedge t' \prec t\}$ .

Regarding point (2), any terms in  $T$  also can be candidates for specialization but have lower priority than the ones suggested in (1).

Summing up, for a new atomic document  $d$  which is a part of  $d_p$ , we can suggest the following sets of terms for a description of  $d$ .

- $D_{rec} = D_{sum} = summary(d_p)$
- $D_{opt1} = \{t | t_s \in specializedIn(d_p) \wedge t \in tail(t_s)\} / D_{rec}$
- $D_{opt2} = T / (D_{rec} \cup D_{opt1})$
- $D_{obso} = \emptyset$

In the above sets of terms,  $D_{rec} \cup D_{opt1} \cup D_{opt2}$  is equal to  $T$ . This means that authors can choose any sets of terms from  $T$  as a description of an atomic document even if the atomic document is a part of a composite document.

## 4.2 Create New Composite Documents

On the other hand, to define a description of a composite document, there is a strict criterion the description should satisfy, namely soundness. Suppose an author has made a composite document  $d$  with its components  $comp(d)$  and now he is going to define an author description  $ADescr(d)$ . In this case, any term in  $ADescr(d)$  should be a member of  $dtDom(d) = \{t | t \in T \wedge \exists t' \in cover(d) \wedge t' \preceq t\}$  to satisfy the soundness. Therefore we use  $dtDom(d)$  as  $D_{opt}$  for suggestion.

Regarding  $D_{rec}$ , there is no criterion for determining terms that a description of a composite document should have. However, to construct a description from an empty set of terms is a troublesome task. Therefore, here we would like to use a summary of a document as the starting point of description authoring.  $D_{sum} = \text{summary}(d)$  is suggested as  $D_{rec}$  for a description.

The important point to note is that, in this case, terms in  $D_{sum}$  are not mandatory for a description of  $d$ . The author can remove terms belonging to  $D_{sum}$  from  $ADescr(d)$  to hide some contents included in  $d$ . For instance, in the example illustrated in Fig. 4, the summary of the composite document is  $\{\text{Sort}, \text{OOL}\}$ . However, if the description author thinks that programming languages are not in important topic in the context of the composite document, he can drop  $\text{OOL}$  and keep only  $\{\text{Sort}\}$  as a description of the composite document.

As a consequence, for a new composite document  $d$ , we can suggest the following sets of terms for a description of  $d$ .

- $D_{rec} = \text{summary}(d)$
- $D_{opt} = dtDom(d)/D_{rec}$
- $D_{obso} = \emptyset$

Additionally, when an author makes a new composite document as a part of an existing document, we can give more accurate suggestion by comparing with siblings of the new document. See the example illustrated in Fig. 5. In this example, originally the summary of  $d = d_4 + d_5$  is just  $\{\text{Quick sort}\}$ . However, this description might be too brief because all siblings of  $d$  have terms related to programming languages in their descriptions. We can capture such topics shared by siblings as a summary of a parent document. In this example, the parent document  $d_p = d_1 + d_2 + d_3$  has its summary  $\{\text{Sort}, \text{Languages}\}$ . Therefore, terms in  $cover(d)$  related to, more precisely, subsumed by  $\text{Sort}$  or  $\text{Languages}$  also should be suggested as a part of a description of  $d$ . As a result, we get  $\text{C++}$  as an additional member of  $D_{rec}$ . Finally, we can suggest  $\{\text{Quick sort}, \text{C++}\}$  as an initial description of  $d$ .

To sum up, for a new composite document  $d$  which is a part of  $d_p$ , we can suggest the following sets of terms for a description of  $d$ .

- $D_{rec} = \text{reduce}(\text{summary}(d) \cup \{t \mid t \in cover(d) \wedge \exists t_s \in \text{summary}(d_p) \wedge t \preceq t_s\})$
- $D_{opt} = dtDom(d)/D_{rec}$
- $D_{obso} = \emptyset$

### 4.3 Removing Parts of Documents or Document Descriptions

When an author removes some parts of a composite document, or terms from descriptions of document components, such operation might change the cover of the document therefore it might affect soundness of the document description. To preserve soundness of descriptions, risk of soundness violation should be checked before applying operations and be appropriately notified with a list of terms that should be removed to keep soundness.

The algorithm `checkSoundness` in Fig. 6 takes a document  $d$  and a set of terms  $D$  to remove from  $ADescr(d)$ , or  $cover(comp(d))$  when the document  $d$

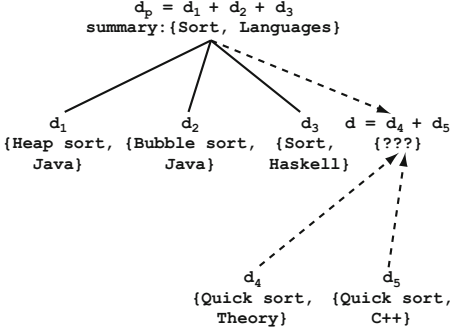


Fig. 5. Creation of a composite document as a part of an existing document

**Algorithm checkSoundness**

```

Input
A document d. A set of terms D to remove, or
D = cover(comp(d)) when d itself is removed.
Output
A set of mappings M_v = {m_v : d_v ↦ D_v} such
that d_v is a document having a set of terms D_v
that violate the soundness criterion.

if d has no parent then return ∅ end if
d_p ← parent(d)
D_r ← cover(d)/D
for all d' ∈ parts(d_p) such that d' ≠ d do
    D_r ← D_r ∪ cover(d')
end for
D ← reduce(D_r ∪ D)/D_r
if D = ∅ then return ∅ end if
M_v ← checkSoundness(d_p, D)
D_v ← {t | t ∈ ADescr(d_p) ∧ ∃t' ∈ D ∧ t ≼ t'}
if D_v ≠ ∅ then M_v ← M_v ∪ {d_p ↦ D_v} end if

return M_v
    
```

Fig. 6. checkSoundness algorithm

itself is removed. Then this algorithm returns a set of mappings  $M_v = \{m_v : d_v \mapsto D_v\}$  such that  $d_v$  is a document having a set of terms  $D_v$  that violate the soundness criterion. This algorithm recursively propagates terms in  $D$  from a child document to its parent. In each propagation step, terms compensated by descriptions of sibling documents are removed from  $D$ .

By using this algorithm, we can notify authors about document descriptions affected by removing operations, and indicate suggestion lists of terms for updating descriptions to keep soundness. Suppose an author intends to remove a set of terms  $D$  from a description of  $d$ . In this case, firstly we need to compute a set of mappings  $M_v = \text{checkSoundness}(d_p, D)$ . Then for each  $d_v$  with a term set  $D_v$  in  $M_v = \{m_v : d_v \mapsto D_v\}$ , we can suggest the following sets of terms:

- $D_{rec} = ADescr(d_v)/D_v$
- $D_{opt} = \emptyset$
- $D_{obso} = D_v$

**5 Concluding Remarks**

In this paper, we have presented a model for metadata of composite documents. Our model allows authors to freely choose terms from a taxonomy to make descriptions. However, once documents are placed in composite documents, we can infer various restrictions and suggestions on terms for descriptions by taking into account soundness of descriptions. We think soundness of descriptions is a simple but essential criterion for keeping integrity of a document repository.

In future work, an urgent task is prototyping for identifying matching points and mismatch between the model we have proposed here and problems in practice. As we have seen in this paper, the modeling part of this study is very

abstract. Currently we have a plan to prototype a document management system that uses this model in the metadata management part.

Regarding the model, firstly, we would like to extend the concept of document summary. In the current definition, a document summary summarizes topics shared by all *atomic documents*. However, summaries produced by using this definition are somehow too brief. Instead of summarizing at the level of atomic documents, we can summarize a document at a coarser granularity, for instance, direct children of a document to summarize, so that we can get more detailed summaries. In a future study we would like to introduce a *degree of summarization* to control the level of detail of summaries, and use it for assisting description authoring.

Finally, we have discussed only a case that a document has up to one parent. However, if a document is used as part of multiple composite documents, a document can have multiple parents. In this case, we can compare usage of the same document in different composite documents. We would like to find a way to use such comparison of document usage for improving term suggestion.

**Acknowledgement.** This work was partially supported by the European project ASSETS: Advanced Search Services and Enhanced Technological Solutions for the European Digital Library (CIP-ICT PSP-2009-3, Grant Agreement no 250527).

## References

1. Connexions web site, <http://cnx.org/>
2. Rigaux, P., Spyrtatos, N.: Metadata Inference for Document Retrieval in a Distributed Repository. In: Maher, M.J. (ed.) ASIAN 2004. LNCS, vol. 3321, pp. 418–436. Springer, Heidelberg (2004)
3. OWL 2 Web Ontology Language Document Overview, <http://www.w3.org/TR/owl2-overview/>
4. SKOS Simple Knowledge Organization System Reference, <http://www.w3.org/TR/skos-reference/>
5. The Gene Ontology Consortium Gene Ontology: tool for the unification of biology. Nature Genetics 25(1), 25–29 (2000)
6. AAT Web site, <http://www.getty.edu/research/tools/vocabularies/aat/>
7. The DBpedia Ontology, <http://wiki.dbpedia.org/Ontology>
8. Baeza-Yates, R., Ribeiro-Neto, B. (eds.): Modern Information Retrieval. Addison-Wesley (1999)
9. Erdmann, M., Maedche, A., Schnurr, H.-P., Staab, S.: From Manual to Semi-automatic Semantic Annotation: About Ontology-Based Text Annotation Tools. In: Proc. COLING Intl. Workshop on Semantic Annotation and Intelligent Context (2000)
10. Handschuh, S., Staab, S., Volz, R.: On deep annotation. In: Proc. Intl. World Wide Web Conference (WWW), pp. 431–438 (2003)
11. Pastorello Jr., G.Z., Daltio, J., Medeiros, C.B.: Multimedia Semantic Annotation Propagation. In: Proc. of IEEE International Symposium on Multimedia (ISM 2008), 509–514 (2008)
12. Leung, M.-K., Mandl, T., Lee, E.A., Latronico, E., Shelton, C., Tripakis, S., Lickly, B.: Scalable Semantic Annotation Using Lattice-Based Ontologies. In: Schürr, A., Selic, B. (eds.) MODELS 2009. LNCS, vol. 5795, pp. 393–407. Springer, Heidelberg (2009)

# A Multi-layer Digital Library for Mediaeval Legal Manuscripts

Monica Palmirani and Luca Cervone

University of Bologna, CIRSFD  
via Galliera 3, 40121 Bologna, IT  
{monica.palmirani, luca.cervone}@unibo.it

**Abstract.** This paper presents the results of the MOSAICO project, an Italian Government research project (2008–12) funded by the Italian Ministry of Education and Research, and carried out by an academic consortium.<sup>1</sup> The goal of the Mosaic project (<http://mosaico.cirsfid.unibo.it>) is to create a thematic and specialized digital library, relying on the Web 2.0 and the P5 TEI XML standard to manage heterogeneous descriptions of medieval codex images. The portal is designed for scholars of medieval legal history and emphasizes the intellectual path of the academic experts.

## 1 Introduction

The European Commission is currently devoting much attention to the digital library goal<sup>2</sup> as a complex method for favouring the access to rare materials, for guaranteeing the long-term preservation of the cultural heritage, and for sharing knowledge by overcoming physical limitations. In the domain of medieval manuscript digitalization we find outstanding projects by libraries, institutions, and universities (Manuscripta Medievalia, a German consortium;<sup>3</sup> e-codices virtual manuscript library of Switzerland;<sup>4</sup> the Max-Planck-Institut für europäische Rechtsgeschichte;<sup>5</sup> the Enrich project database;<sup>6</sup> the Europeana Regia project;<sup>7</sup> Shared Canvas, managed by the University of Stanford and Los Alamos National Laboratory;<sup>8</sup> etc.) that over time have digitalized the manuscripts for future generations. Even if these projects, among others, define a robust backbone of the digital library initiative, they are much too oriented toward bibliographic description and classification of the material based on librarian criteria and codicological best practices 1 2, rather than being focused on allowing scholars to annotate the precious manuscripts through their expertise.

---

<sup>1</sup> University of Bologna, CIRSFD (coordinator); University of Federico II, Naples; University of Roma Tre.

<sup>2</sup> See 6.

<sup>3</sup> <http://www.manuscripta-mediaevalia.de>

<sup>4</sup> <http://www.e-codices.unifr.ch/en>

<sup>5</sup> <http://dlib-pr.mpier.mpg.de/>

<sup>6</sup> <http://www.manuscriptorium.com/>

<sup>7</sup> <http://www.europeanaregia.eu/>

<sup>8</sup> <http://www.shared-canvas.org/>

The goal of the MOSAICO project is to provide scholars with a very rich and easy to consult platform of manuscripts and a quick way to write content and metadata related to each manuscript, piece of artwork, page, or page fragment as other works are done 5. The power reached by modern Web applications permits to us to create a full in-browser system having, as its first convenience, the characteristic of providing a unified and collaborative venue in which historical scholars from all over the world can work together to improve the catalogue. The MOSAICO platform aims to store and manage each digital resource (content, metadata, images, comments to the images, etc.) in a neutral way so as to allow a multiform access to them on three different historical points of view: Roman, medieval, and contemporary.

## 2 Functionalities and Requirements

The Mosaico environment includes the following functionalities arrived at through several interviews and focus-group meetings within the consortium:

1. Collecting different digital materials on medieval legal manuscripts using patterns and templates.
2. Permitting scholarly annotations by writing text and hypertext using multiple templates available through a special Web editor. This approach makes it possible to preserve the intellectual and original logical structure designed by the author. We want to go beyond the rigid architecture of the DMBS, which forces authors to organize their thought on the basis of the database's logic layer. We aim to provide a Web editor capable of marking up in XML the metadata in the hypertext template.
3. Managing a plurality of templates of historical works on the basis of the different products expected (descriptive schedules, critical editions, comparative editions, multi-layer presentation, etc.).
4. Annotating manuscripts in XML format, so as to better manage the embedded knowledge and share it with a network of libraries across Europe. Further, the metadata will be recursively annotated in P5 TEI Enrich format, making it possible to overlay comments onto other comments, either hierarchically or in multiple and simultaneous fashion.
5. Comparing different manuscripts related to the same topic, thus creating an environment for historians to build, with the support of technology, a comparative critical edition.
6. Searching each codex's incipit and explicit using the roots of the Latin vocabulary.
7. Pointing-and-clicking on any image to bring up information relative to the codex being viewed and to its history.
8. Zooming in and out of the manuscripts and isolating a portion so as to focus on it.
9. Connecting the resources stored using association expressed in RDF.
10. Using a special viewer that can manage high-resolution images and in the meantime protect them (through the pyramid processing method) from illegal processing.

11. Including in each fragment of the zoom process the original library's watermarking.
12. Exporting metadata into P5 TEI Enrich XML format, making it possible to share material by way of digital-library initiatives across the world.
13. Using thumbnails any time the text cites an image.
14. Dynamic creation of interconnection tables among the digital resources on the basis of RDF assertion.
15. Managing the glosses and "tracce d'uso."<sup>9</sup>

The platform can manage security and IPR issues, block access to images, and track illegal misuse. Before to accessing the digital library collection, the user has to accept the terms of use.

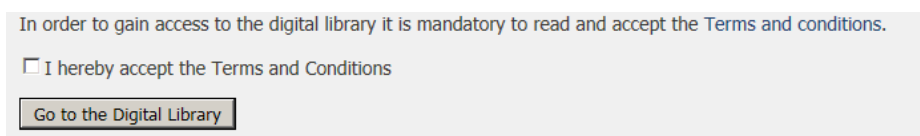


Fig. 1. Legal terms and conditions

### 3 Patterns and Templates

One of the most important features is to use patterns and templates for the content, so as to lead the author in organizing the material.

With the help of the consortium partners, we have identified five patterns:

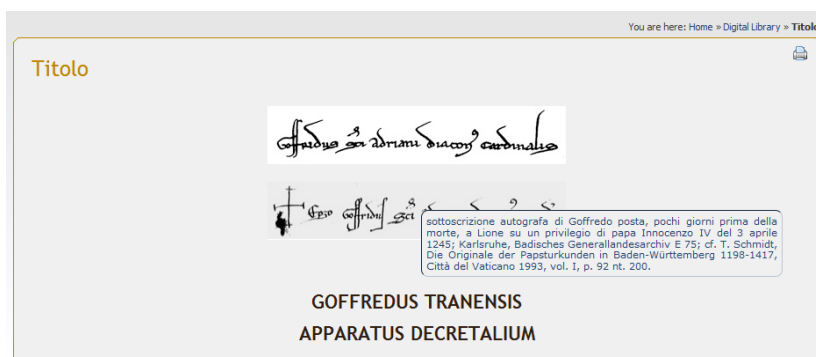


Fig. 2. Montecassino 266 description and mouse-over function

<sup>9</sup> *Tracce d'uso* ("traces of use") are annotations that students make on the code. They often record comments a professor has made during lecture. They are invaluable for scholars the medieval legal history, who can deduce the use and the interpretation of the code by the different schools of law.

### 1. Monographic and Hypertext Description of One Manuscript

In this template, the author describes the manuscript as a book and connects the images with the text using a hypertext model. The Montecassino 266 manuscript is an example of how Bertram described the images and connected them to the novellae.

Image thumbnails are included in all parts of the text.

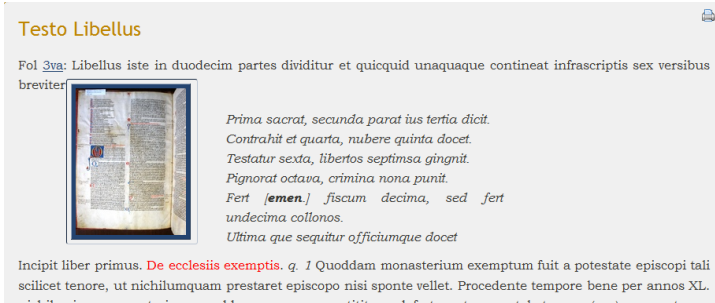


Fig. 3. Image thumbnail mentioned in the incipit the text

Clicking on the thumbnail will give access to the full page.

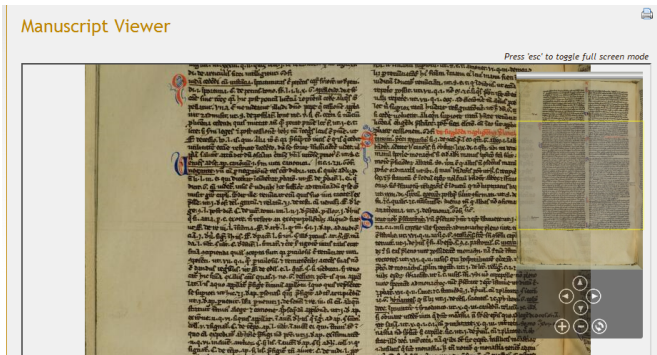


Fig. 4. Full page of Manuscript 266

Manuscripts

You are here: Home > Digital Library > Manuscripts

- Has previous descriptions
- Has related images
- Has bibliography
- Has related materials

Authenticum menu

- Introduction
- Manuscripts
- Studies Chronology

Num.	Manuscript	Comp. text	Int. lit. Voluntas	Sec.	VL. presaz.	Acc. Np. Nova	Orthographia absent	Chironomias present
1	Angers, BM, 333	x	composito	XII.2	**	100	nessuna	11*; 13*; 21*
2	Bamberg, SB, Jur_4	x	Inst.+Auth	XIII.1		96	63; 125	11 Jur. 4
3	Berlin, SBPK, lat. fol. 271	x	no	XII.2	**	97	105; 110; 125; 128	11*; 13*; 21*
4	Berlin, SBPK, lat. fol. 823	x	composito	XII.2	*	95	63; 95; 110; 125	11*; 13*
5	Bologna, BC, A.132	bc.	composito	XII.2	*	100	110	11*; 13*; 21*; 45*
6	Bruxelles, BR, 12084	x	composito	XIII	*	95	63; 110; 125	Ed. VIII

Fig. 5. Comparative table of different manuscripts related to the same subject (Authenticum)

### 2. Comparison of Different Manuscripts on the Same Subject

The Authenticum includes 28 descriptions of the same subject. In the table below it is possible to see the different manuscripts of the Authenticum that Loschiavo se-



lected and described. The legend indicates the presence of images, related material, other previous historical descriptions, and a bibliography related to the code.

The screenshot shows a web page titled "Leipzig, Hänel 5". It contains the following information:

11. Leipzig, Universitätsbibliothek, Hänel 5 (*antea* 3467) sigl8 Lei

59 ff. membr.; mm. \*\*\* x \*\*\*

**fascicolazione:** 12; 78

**età:** saec. XIIlex.-XIIIin.

**origine:** Italia

**provenienza:** G. Hänel; F.L. Keller; Warmkönig

**scrittura:** minuscola carolina

**testo:** unica mano

**ornamentazione:** molto curata: da segnare alcune eleganti figure zoomorfe (cane f. 1r inf.; cigno 2v; pavone 4r; draghi 6r inf., 14va inf., 17r inf.; oca f. 21v inf., 35vb) e frequenti motivi ornamentali (in particolare in occasione delle rubriche delle singole novelle)

**contenuto:**  
I. (ff. 1ra-59vb) *Authenticum* (mutlo) (sec. XIIlex./XIIIin.; Italia)

On the right side, there are three buttons: "View previous descriptions", "View manuscript's images", and "View bibliography".

**Fig. 6.** Description of a manuscript from the *Authenticum*

Using RDF relationships among the different digital resources (bibliography, other multimedia material, images, etc.) it is possible to dynamically create a chronological table of the different available material in the database related to a specific resource.

The screenshot shows a web page titled "Studies Chronology". It contains a table with the following data:

Wm.	Manuscript	HEIMBACH BIENER	SAVIGNY I	SAVIGNY HANEL 1830 II	PRODRONUS HANEL 1837	SAVIGNY 1816
1	Angers, BM 333	XOXXIII-XOXXIV n° 46				
2	Bamberg, Jur. 4	LXIII n° 101		V 16-17	37	
3	Berlin, lat. fol. 271	LXOIII-LXOVII n° 109				380-381
4	Berlin, lat. fol. 823					
5	Bologna, BC, A 132					
6	Bruxelles, 12084					
7	El Escorial, 11118					

**Fig. 7.** Chronological dynamic table of the resources in the database

### 3.1 List of Descriptions

This model (Fig. 8) is quite similar to applications that use relational databases. In this model users have a search mask in which they can see a list of tabs with metadata regarding several works. The user can do searches in the list in order to reach the needed work and can then click on the name of the work and read the tab. At this point you can also see the scans of the work using the reading tool described in the previous paragraph. This is a “multi editor” model in which there are several editors that write the metadata pertaining to the works.

### 3.2 Comparison of Different Transcription on the Same Subject

This model (Fig. 9) is used when there are several manuscripts that make up a unique “meta-manuscript.” Also, there are several editors that write the metadata related to the manuscripts. In this case, users need to read the different pages and data in a

“side-by-side” model. So they can open a page on the left side of the reading tool and another page, probably belonging to another manuscript, on the right side so as to compare them and read the different metadata written for each of the pages.

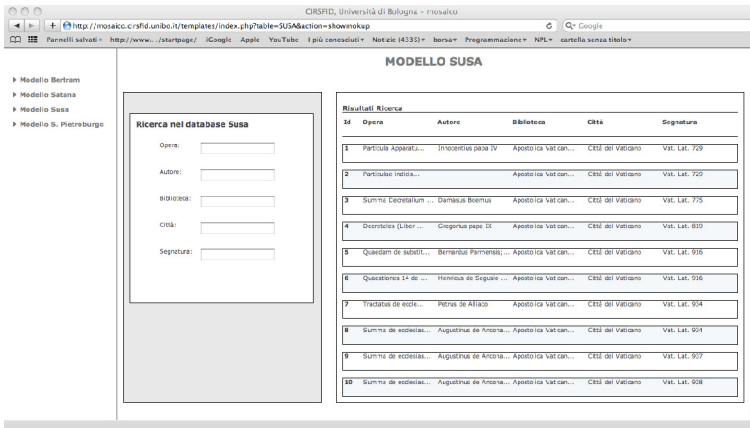


Fig. 8. List of descriptions

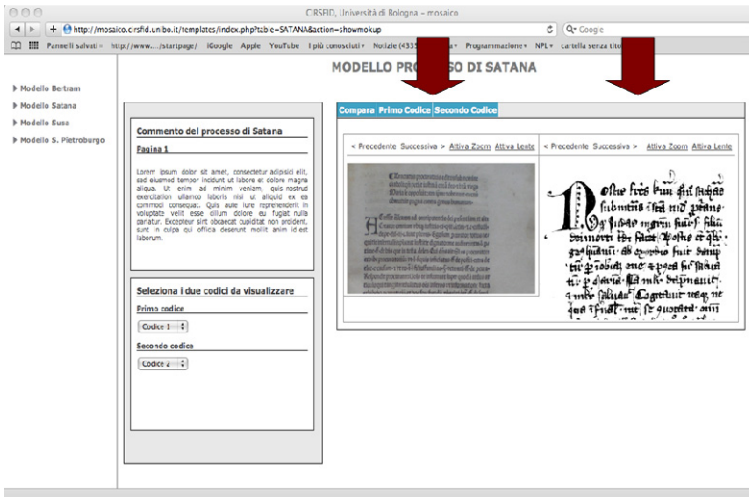


Fig. 9. Satana manuscripts: comparison windows

### 3.3 Temporal Sequence of Digital Material

The Saint Petersburg model” (Fig. 10) is used when there are different manuscripts and different metadata written by different editors over time. It is the most complex model for viewing metadata because there is a horizontal level of metadata and a vertical one. In this model the users can open a set of manuscripts and they can navigate them over time, so they can see the different metadata written over time navigating the

manuscript's vertical level, and they can see the different manuscripts navigating through the manuscripts' horizontal level.

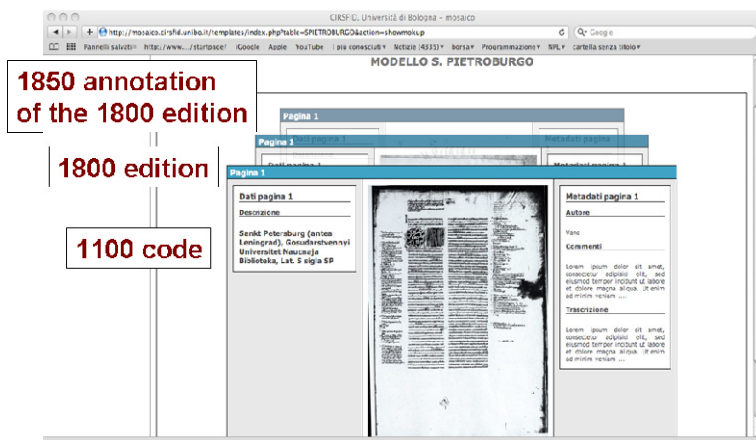


Fig. 10. Authenticum's Saint Petersburg Manuscript with related material by Biner

## 4 The MOSAICO Architecture

MOSAICO is a Web application comprising two main elements: the server-side component and the client-side component. The server-side component is charged with performing ordinary operations on the database (data retrieval, saving, updating, deleting), making all the computation requested, and displaying the final results to the client side. The client-side component is charged with accepting user requests (made in a human-readable manner), sending them to the server side, retrieving the results, transforming them in a human-readable format, and, finally, presenting them to the users. In the MOSAICO project both the client and the server sides are made of other macro components.

### 4.1 The Server-Side Component

The server side component of the MOSAICO project is actually composed of three layer for corresponding three servers.

There are two servers that host the MOSAICO data repositories: the first one contains the XML repository and the second one the image repository. The main sever contains the MOSAICO application core and the packages that are used to communicate with the repositories and with the application's client-side component.

The core application, hosted by the main server, uses the MOSAICO repository manager to access the repository manager via HTTP, and it provides the MOSAICO API, which can be invoked by the MOSAICO portal to do simple and specific operations on the XML documents and on the images in the repository. For instance, when a reader-tool requests a page, the CMS calls, typically using a REST query, and

selects the appropriate method of the MOSAICO API, which dispatches the request to the application core. The application core, passing through the MOSAICO repository manager, retrieves the images and the XML related to the requested page, packages them to make everything readable by the reading tool, and, always using HTTP, returns the results to the client, which may carry out other formatting and presentation operations as needed and will finally reply to the user's request by supplying a human readable version of the materials related to the requested page.

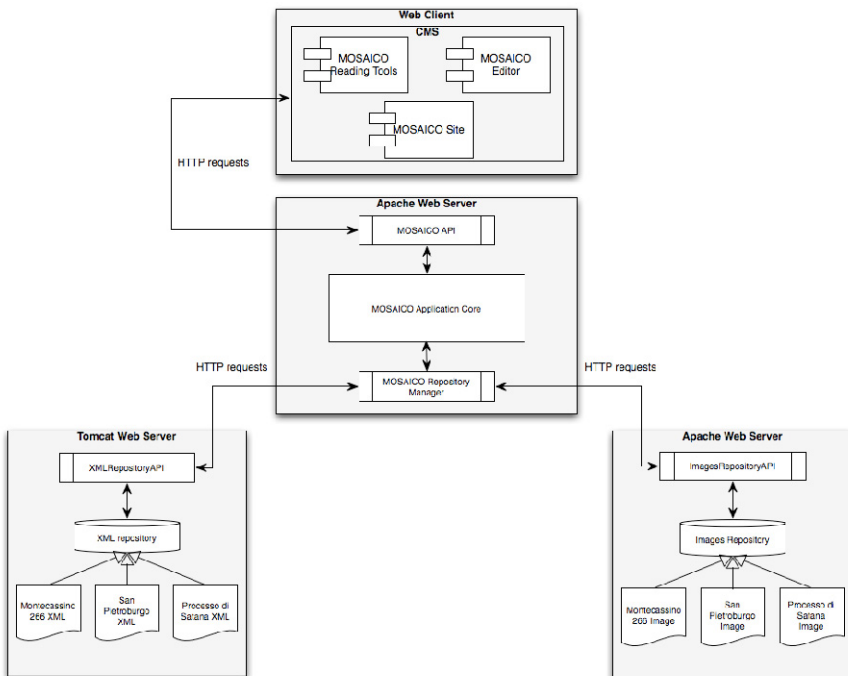


Fig. 11. MOSAICO architecture

## 4.2 The XML Repository

One of the main aims of this project is move past the idea of relational databases and start using an ML standard to markup the metadata relative to the manuscripts. This is because the descriptive tabs supplied by different historical experts can be formatted in very different ways. To store them in a relational database we need to identify any relevant partition, extrapolate it from the original document, and save it in an appropriate database table. But in this way we cannot preserve the structure of the original document, and that exponentially increases the risk of not being able to recreate the document. With XML we can mark up the relevant partitions preserving the original document format.

For this reason we need to use a native XML database that makes it possible to store the documents, and we also need to perform smart queries on the document collection. We choose the eXist database (<http://exist.sourceforge.net/>), currently the most widespread and supported native XML database. It comes with a built-in Tomcat Web server so it can be accessed through REST requests. In this way we can always keep separate the data (the document collection) from the application's other parts. So a service malfunction cannot corrupt the integrity of the document collection. Another point in favor of the system's security is that eXist is never queried by an external application but only by the MOSAICO repository manager in order to satisfy the application core requests. Another pillar of the MOSAICO project is the use of a permanent URI, after the FRBR<sup>10</sup> model (work, expression, manifestation, item: title/author/shelfmark), so as to have a permanent link for each resource independently of its physical storage in the image repository.

### 4.3 The Image Repository

The MOSAICO project collects a very large set of images that are essentially scans of the original manuscripts. These images are usually protected by copyright, so security policies are a main issue for the project. The best practice is to store them in a completely independent server accessed as a NAS (Network Attached Storage system) hosted on an APACHE Web server.

This is helpful in two respects:

- The images are protected from system failures. In other words, neither physical nor logical failure of the system can corrupt the integrity of the images.
- The images are protected from malicious attacks. The NAS is in a private LAN network, so the images can be accessed only by the application core (managing the repository API).

The reason for choosing to host the image repository in an APACHE Web server is because this makes it possible, even in this case, to use REST queries to send the application core's request (passing through the MOSAICO repository manager). So all communication between the project's components are sent using an homogeneous communication architecture.

### 4.4 The MOSAIO Repository Manager

Both the eXist database and the NAS have APIs for access to the database. eXist provides a REST API that can be used to retrieve documents, upload documents onto the database, and send simple queries to the documents collection. REST (Representational Transfer Rate) is a paradigm that makes it possible to manage resources using the HTTP protocol.

---

<sup>10</sup> Functional Requirements for Bibliographic Records,  
<http://www.ifla.org/en/publications/functional-requirements-for-bibliographic-records>

Also, the NAS can be accessed via REST interrogations.

It is important to note that the MOSAICO repository manager do not perform any computation on the resource retrieved or uploaded. Simply it has in charge to satisfy the complex requests that are performed by the application core.

## 5 The Client-Side Component

The client-side component is the interface that users use to request actions to the server-side component of the MOSAICO project. Essentially, it is made up of a CMS that contains three main objects:

- **The MOSAICO site.** This is simply MOSAICO project's institutional site, and the start point from which to access the reading tools and the editor and to do queries on the database.
- **The MOSAICO reading tools.** The reading tools are software that makes it possible to read the scans of the original manuscripts of the MOSAICO document set alongside the relative metadata. There are several types of manuscripts and metadata, and for this reason, and in order to provide users with a good reading experience, there are not one but several types of reading tools, and each manuscript uses the most appropriate one.
- **The MOSAICO editor.** The MOSAICO editor is used to write MOSAICO documents or to mark them up, using the MOSAICO XML P5 standard on documents already written with any other text editor. The editor is under construction.

### 5.1 The MOSAICO Web Site

The MOSAICO site is the institutional site of the MOSAICO project. It contains all the information about the project and the consortium. It is also the bridge the get access to the digital library. The site's content s created and updated using an open source CMS named Impress CMS (<http://www.impresscms.org>). In order to present the site, the client side of the MOSAICO project, simply requests the page to the server side API. The server side replies with the content of the requested page and then the client side applies to it a specific style sheet and present the page to the user.

### 5.2 The MOSAICO Reading Tools

The MOSAICO project permits to read the scans of the original manuscripts in several ways. This is because the manuscripts and the related metadata come from very different heterogeneous sources. In order to read the scans of a manuscript the user must access to the "digital library" section of the site and choose the manuscript to read or the model to use to see the images and the related metadata of a specific manuscript. When a user requests a manuscript, or a single page, the client side component of the MOSAICO system, dispatches the request to the server side API. As we have seen previously in this document, the server side component performs all the

action needed to retrieve the data related to the requested item and creates a MOSAICO package. When the package is returned to the client side component, the client side understands by the info contained in the package what is the model to use in order to read the objects contained in the package, instantiates and present to the user the appropriate reading tool for those objects. There are several models that are used in the MOSAICO project in order to give to the user the best reading experience. All the reading tools are, in a technical language, AJAX (Asynchronous Javascript And XML) applications that use Javascript and run time calls to the server side component and use the best Web 2.0 techniques in order to supply fast and powerful pages reading.

### 5.3 The MOSAICO Editor

The MOSAICO editor is a WYSIWYG in-browser editor used to mark up document using in the MOSAICO XML format (a customization of the TEI P5 format <http://www.tei-c.org/>). The editor is now under construction, so we present here its main features. It will be able to perform all the basic operations of common text editors, and it makes it possible to create a new metadata tab (as well as RDF triples) on manuscripts and also to import a previously written metadata tab and mark up it in the MOSAICO XML format. The editor will be able to perform the following actions:

- **Creating a new document.** The user can create a new blank document and can use the editor to write the document natively in the MOSAICO XML format. It is important to note that, at any time while editing, the user can access the photo catalogue and see the images in the same window of the editor, taking advantage of the “side-by-side” editing feature.
- **Importing a document.** A common behavior is that users write a document (a metadata tab related to a manuscript) using third-party software, for instance, desktop software, and then they needs to mark up the document in the MOSAICO XML format and upload it in the MOSAICO digital library system. Users can use the editor to import the document and can tag the different parts of the document assigning to each part the appropriate MOSAICO XML tag. However, this operation, is not completely manual, this because the editor also has a parser that, when a document is imported, tries to understand the relevant parts of the documents and tries to pre-mark up this parts.
- **Enriching documents stored in the XML repository.** Of course it is possible to save in the XML repository a document that has not completely been marked up and open it later to complete the markup. The MOSAICO editor also has a versioning system and makes it possible to manage RDF relationships. A group of users can cooperatively create a document, so when one of these users tries to open one of these documents, he or she can view the document’s latest version, in which all recent changes are highlighted. It is also possible to open a specific version of the document.
- **Saving a document.** A user can save a document at any time, when the markup is complete or when it is incomplete. If the user thinks the mark up is complete and tries to save it in the XML repository, the editor validates it to see if it belongs to

the MOSAIO XML schema. If the document is valid, the saving action is performed and the document is instantly made accessible in the digital library. If the document is incomplete, or if it is partially and voluntarily marked up, the editor returns to the user all the problems found in the validating operation so that the user can correct them until the markup is complete and valid.

- **Exporting a document.** All the documents stored in the XML repository can be exported in the most common third-party formats. When this operation is requested, the client-side component asks the server-side API to translate the document.

## 6 Conclusion

The MOSAICO project outstrips the current state of the art in digital library hosting manuscripts. Which is to say that the MOSAICO environment seeks to support scholars of mediaeval legal history in creating multimedia and hypertext content, thus preserving and even enriching the digital manuscripts heritage and the connected material. The idea is to not impose any rigid template to the authors but to provide a flexible environments using XML and RDF models that effectively manage the metadata, the semantic parts of the text, and the relationships among digital resources. In this way we make for new ways of using medieval legal history materials (e-books, critical editions) and can create new technical tools (comparison tables) for supporting research in this domain.

## References

1. Agosti, M., Mariani Canova, G., Orio, N., Ponchia, C.: A Case Study for the Development of Methods to Improve User Engagement with Digital Cultural Heritage Collections. In: Grana, C., Cucchiara, R. (eds.) MM4CH 2011. CCIS, vol. 247, pp. 166–175. Springer, Heidelberg (2012)
2. Baechler, M., Ingold, R.: Medieval Manuscript Layout Model. In: Procs. of ACM Symposium on Document Engineering, Manchester, UK, pp. 275–278 (September 2010)
3. Enrich Final Conference Proceedings, National Library of Spain, Madrid, November 5-6 (2009), [http://enrich.manuscriptorium.com/files/enrich/ENRICH\\_WP8\\_D8\\_5\\_Proceedings\\_Web.pdf](http://enrich.manuscriptorium.com/files/enrich/ENRICH_WP8_D8_5_Proceedings_Web.pdf)
4. Sanderson, R., Albritton, B., Schwemmer, R., Van de Sompel, H.: SharedCanvas: a collaborative model for medieval manuscript layout dissemination. In: JCDL 2011, pp. 175–184 (2011)
5. Second progress report on the digitization and online accessibility of cultural material and on digital preservation in the European Union (November 2010), [http://ec.europa.eu/information\\_society/activities/digital\\_libraries/doc/recommendation/reports\\_2010/2010%20Digitisation%20report%20overall.pdf](http://ec.europa.eu/information_society/activities/digital_libraries/doc/recommendation/reports_2010/2010%20Digitisation%20report%20overall.pdf)
6. TEI Consortium (eds.): TEI P5: Guidelines for Electronic Text Encoding and Interchange. TEI Consortium (November 2007), <http://www.tei-c.org/Guidelines/P5/> (accessed: January 31, 2012)



# Extracting Keyphrases from Web Pages

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.** Social tagging systems allow people to classify Web resources by using a set of freely chosen terms commonly called tags. However, by shifting the classification task from a set of experts to a larger and untrained set of people, the results of the classification are not accurate. The lack of control and guidelines generates noisy tags (i.e. tags without a clear semantic) which lower the precision of the user generated classifications. In order to face this limitation several tools have been proposed in the literature for suggesting to the users tags which properly describe a given resource. On the other hand we propose to suggest n-grams (named keyphrases) by following the idea that sequences of two/three terms can better face potential ambiguities. More specifically, in this work, we identify a set of features which characterize n-grams adequate for describing meaningful aspects reported in the Web pages. By means of these features, we developed a mechanism which can support people when classifying Web pages by automatically suggesting meaningful keyphrases.

## 1 Introduction

Jeff Howe defined social tagging systems as one of the main examples of *crowdsourcing* systems [8]. Coined by Howe in June 2006, the term crowdsourcing appeared the first time in the article ‘*The Rise of Crowdsourcing*’ [1] for defining the act of sourcing tasks traditionally performed by specific individuals (with specific competences) to an undefined large community of people (the crowd). According to Howe’s theory the technological advances can significantly reduce the gap between professionals and amateurs: people can use cheap technologies to execute complex tasks. In this way complex tasks, such as the classification of digital resources, can be executed by a large community of people by saving significant resources: this is clearly achieved at the cost of less accurate results. Money is not the only way to compensate the crowd for their work: prizes, services or the intellectual satisfaction can stimulate people to use their intelligence and talent into sophisticated tasks.

The large population of the users of social tagging systems are the crowd used to classify Web resources on behalf of knowledge engineers and domain experts. Social tagging systems do not provide a monetary compensation to the taggers, but people are compensated with:

- The services provided by the social tagging system. Social collaboration is a good mean for retrieving meaningful information since each user can enjoy the classification produced by other peers. Moreover, by tagging resources, people can easily retrieve resources they classified in the past.
- Intellectual satisfaction. Users can be interested in using social systems to propagate their ideas, to influence other people, and to help people with similar information needs.

Obviously, by shifting the classification task from a set of experts to a larger set of untrained people, the results of the classification cannot be rigorous. In fact, due to the lack of control and guidelines, the precision of the returned classification produced is lowered by noisy tags (i.e. tags without a clear semantic).

How can we reduce the gap between experts and Web users? The answer to this question is still in Howe’s ideas of filling the gap between people with specific expertise and not experts with proper technologies: according to this theory we can reduce the gap between knowledge engineers and users of social tagging systems by introducing tools able to simplify the classification task.

In order to reach this aim, we can support people with mechanisms able to suggest significant and appropriate tags which can be used to classify Web resources in a adequate way. In this work we propose to suggest to the user multi-terms, i.e. n-grams named keyphrases, as a support for classification. The main motivation to suggest keyphrases is that many concepts are reported as multi-terms (for instance the concept ‘*Unified Modeling Language*’). In these cases, keywords (i.e. uni-grams) do not properly represent the concepts which should be used to label/classify digital document. Following this idea, we propose in this paper the DIKpEW (Domain Independent Keyphrase Extraction for Web pages) mechanism which is aimed at supporting people classifying Web pages by extracting potentially relevant and significant n-grams from the content of the specific considered HTML page. Obviously the proposed system cannot substitute the work of experts, but it is a tool usefull to normalize the user classifications by reducing the number of ambiguous/misleading classifications.

The paper is organized as follows: in Section 1.1 we survey the keyphrase extraction task; the proposed approach to extract keyphrases from Web pages is illustrated in Section 2; Section 3 describes the evaluation settings and the results; final considerations conclude the paper in Section 4.

## 1.1 Keyphrases Extraction

Keyphrase extraction methods have been successfully used for executing relevant tasks in the field of digital libraries, such as: indexing document collections [7], classifying resources [14], providing automatic tagging [19], and filtering resources [6,16]. The task of extracting keyphrases from textual resources is usually implemented in two steps: the *candidate identification* phase and the *selection* phase. The candidate identification phase is exploited in order to identify an initial set of possible keyphrases for a given document. This initial set of keyphrases (referred as ‘*candidate keyphrases*’) is then analyzed in the selection phase for

selecting only the most meaningful ones, i.e. the candidates keyphrases which better summarize the textual resource. Existing methods for keyphrase extraction can be divided into supervised and unsupervised approaches.

A *supervised approach* builds a model by using training documents that have already keyphrases assigned by humans. This model is trained to learn features of the relevant keyphrases (the keyphrases assigned by humans to the training documents) and then it is exploited in order to select keyphrases from previously unseen documents. *KEA* [22] is a notable supervised approach which uses a Bayesian classifier. *KEA* analyzes training documents by taking into account orthographic boundaries (such as punctuation marks and newlines) in order to find candidate phrases. In *KEA* two specific features are exploited:  $\text{tf} \times \text{idf}$  (term frequency  $\times$  inverse document frequency) and the position of the first occurrence of the term. Hulth [9] introduces linguistic knowledge (i.e., *POS*, *Part-Of-Speech tags*) in determining candidate sets: 56 potential *pos-patterns* are used for identifying candidate phrases in the text. The experimentation carried out by Hulth has shown that, using a *POS tag* as a feature in candidate selection, a significant improvement of the keyphrase extraction results can be achieved. Another system that relies on linguistic features is *LAKE* (Learning Algorithm for Keyphrase Extraction) [5]: it exploits linguistic knowledge for candidate identification and it applies a Naive Bayes classifier in the final keyphrase selection. All the above systems need training data (in a larger or smaller extent) in order to construct an extraction system. However, acquiring training data with known keyphrases is not always feasible and human assignment is time-consuming. Furthermore, a model that is trained on a specific domain, does not always produce adequate classification results in other domains.

The *unsupervised approach* eliminates the need of training data. It selects a general set of candidate phrases from the given document, and it uses some ranking strategy to select the most important candidates as keyphrases for the document. Barker and Cornacchia [2] extract noun phrases from a document and ranks them by using simple heuristics, based on their length, frequency, and the frequency of their head noun. In [3], Bracewell et al. extract noun phrases from a document, and then cluster the terms which share the same noun term. The clusters are ranked based on term and noun phrase frequencies. Finally, the top- $n$  ranked clusters are selected as keyphrases for the document. The authors of [17] and [15] proposed unsupervised approaches based on a graph representation of documents. Such approaches use ranking strategies (similar to the PageRank algorithm [4]) to assign scores to each term. Keyphrase extraction systems that are developed by following unsupervised approaches are in general domain independent since they are not constrained by specific training documents.

## 2 Extracting Keyphrases from Web Pages

In [20] we proposed an approach for extracting keyphrases from scientific papers showing also that it outperforms other state of the art mechanisms. The approach we proposed in [20] works under two main assumptions:

1. **A large part of scientific papers is usually written in English.** This simplifies the analysis of the textual content since we have to take into account only the characteristics of the English language.
2. **Scientific papers organize their contributions according to a well-defined schema.** The abstract, the introduction and the conclusion are the sections where the authors usually summarize the goals, the issues and the findings of the work. For this reason, we assign a score to each keyphrase by evaluating the position of the keyphrase in the text: it is plausible that keyphrases in the first part and in the last section of the paper better describe the resource.

These two assumptions are not always true when we want to extract keyphrases from Web pages. In fact, Web pages can be written in languages different from English and, moreover, Web pages do not follow the structure normally adopted by scientific papers. The main aim of this work is to extend, to modify, and to improve the approach we proposed in [20] in order to extract keyphrases from Web pages.

The workflow of DIKpEW, the mechanism proposed in this paper, is shown in Figure 1. By following the traditional schema adopted by keyphrase extraction mechanisms we split the workflow in two parts focused respectively on candidate phrase extraction and on phrase selection phase, described in the following two subsections.

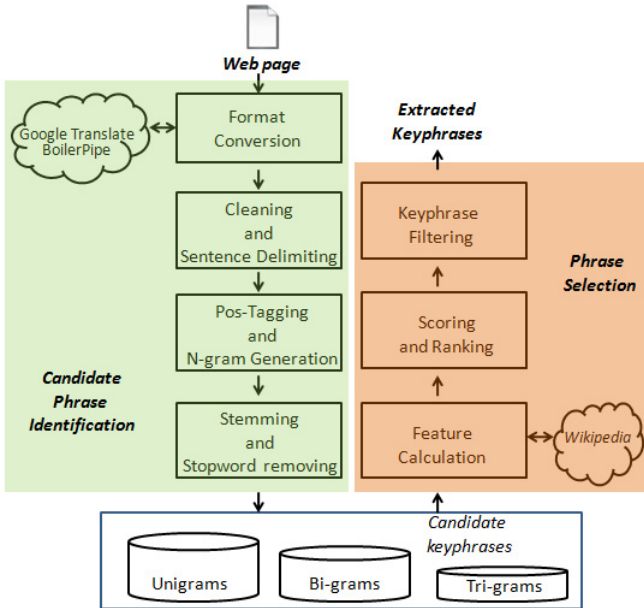


Fig. 1. The workflow used for extracting keyphrases from Web pages

## 2.1 DIKpEW: Candidate Phrase Identification

Given an HTML page, a format conversion step is exploited for extracting the meaningful textual corpus from the document, i.e the textual parts which contain the relevant facts reported in the resource. More specifically, the format conversion is aimed at:

- removing the irrelevant parts from the document. Unfortunately, the main contents of Web pages are often mixed with other textual parts (typically in the headers, the footers, etc.) which are completely irrelevant. In order to discard these useless and noisy parts from the Web page we use an open source Web service called Boilerpipe<sup>1</sup>. The Boilerpipe service, developed by researchers from the L3S Research Center of Hannover, can remove the ‘*surplus*’ text from a Web page. Given a Web page, Boilerpipe returns the main text in the Web page by discarding other information (banner, footers, advertisement, etc.).
- Extracting metadata included by the authors of the Web page. HTML pages are often enriched by their authors with some labels and summaries. These metadata are stored by using tags of the HTML language (*KEYWORDS*, *DESCRIPTION*, and *TITLE* tags).
- Translating the text into the English language. We cannot assume that Web pages are always written into English. In order to re-use the POS-Tagger as well as the POS-Patterns adopted in [20], we translate the text extracted by the Boilerpipe service into English. Currently, we use the Google Translate Api in order to recognize the input language and to translate the text in English.

The output of the format conversion phase is a text in English constituted by the title of the Web page, followed by the metadata extracted from the HTML tags, and concluded by the text extracted by the Boilerpipe service.

This text is analyzed in the cleaning and sentence delimiting step in order to delimit sentences, following the assumption that a keyphrase cannot be located simultaneously in two distinct sentences.

In the POS-tagging and n-gram extraction step we assign a POS tag (noun, adjective, verb, etc.) to each token in the cleaned text by using the Stanford log-linear part-of-speech tagger<sup>2</sup> and then we extract all possible subsequences of phrases including up to 3 words (uni-grams, bi-grams, and tri-grams).

A pruning process is exploited in the stemming and stopword removing step in order to discard keyphrases which do not have a very significant meaning. To this aim, we remove the phrases that start and/or end with a stopword and the phrases containing a sentence delimiter. Partial stemming (i.e., unifying the plural forms and singular forms which refer essentially to the same concept) is performed using the first step of Porter stemmer algorithm [18]. We do not exploit the other steps of the Porter stemmer since they are not appropriate for

<sup>1</sup> <http://code.google.com/p/boilerpipe/>

<sup>2</sup> <http://nlp.stanford.edu/software/tagger.shtml>

keyphrase extraction (consider, for example, the removal of the ‘*ing*’ suffix in the bi-gram ‘*software engineering*’). To further reduce the size of the candidate phrase set, we filter out some candidate phrases by using POS tagging information: Uni-grams that are not labeled as noun, adjective, or verb are filtered out. For bi-grams and tri-grams, only the POS-patterns defined by Justeson and Katz [13] and other patterns that include adjective and verb forms are considered.

Generally, in a document, uni-grams are more frequent than bi-grams, and bi-grams are more frequent than tri-grams, and so on. For taking into account this phenomenon, we build three lists, containing uni-grams, bi-grams, or tri-grams respectively. This allows to treat them separately, without any bias towards uni-grams with respect to bi-grams and tri-grams.

## 2.2 DIKpEW: Phrase Selection

As in [20], some characteristics of the candidate keyphrases are assessed in the feature calculation step for identifying the most relevant keyphrases. The evaluated characteristics have been identified by taking into account how usually Web pages store meaningful information. The considered features are listed and described in following.

1. **Phrase frequency**: this feature is the classical term frequency (TF) metric, exploited in many state of the art keyphrase extraction systems [21, 9, 10]. In our work, the TF value is normalized with respect to the specific n-gram list. More specifically, given the phrase  $P$  in the list  $L$  (the list of unigrams, bi-grams or tri-grams) we define

$$frequency(P, L) = \frac{freq(P, L)}{size(L)}$$

where  $freq(P, L)$  is the number of times  $P$  occurs in  $L$  and  $size(L)$  is the total number of phrases included in  $L$ .

2. **POS value**: as observed in [9] and [2], most author-assigned keyphrases for a document turn out to be noun phrases. For this reason, in our approach, we stress the presence of nouns in candidate phrases by computing POS value as the ratio of the number of nouns in the keyphrase by the total number of terms in the keyphrase.
3. **Phrase depth**: this feature reflects the belief that very frequently Web pages report the most relevant facts at the very beginning of the document: some statistics identify the initial 25% of the text as the part where all main concepts and information are usually reported [12]. In order to highlight such phrases we compute the phrase depth value for phrase  $P$  in a document  $D$  as:

$$depth(P, D) = 1 - \left[ \frac{first\_index(P)}{size(D)} \right]$$

where  $first\_index(P)$  is the number of words preceding the phrase’s first occurrence and  $size(D)$  is the total number of words in  $D$ . The result is

a value in  $[0, 1]$  and highest values are assigned to phrases reported in the initial part of the document.

4. **Wikipedia.** The Wikipedia feature is used to identify more coherent and recognized phrases by following the idea that keyphrases associated to articles in the Wikipedia encyclopedia are more likely associated to well-defined concepts/meaning. The Wikipedia feature is then set to 1 if Wikipedia has a page for describing the keyphrase, 0 otherwise.
5. **Title.** It highlights keyphrases that are included in the title of the Web page (if known). We followed the hypothesis that the title summarizes meaningful concepts which are more deeply discussed in the rest of the text. For each keyphrase, we compute a boolean feature which is set to 1 if the keyphrase is in the title of the Web page, 0 otherwise.
6. **Description.** Authors of Web pages often add a short description of the main contents of the Web page by using the *DESCRIPTION* HTML tag. According to the idea that the summary provided by the author may contain very meaningful information we compute this boolean feature for each keyphrase: the feature is set to 1 if the keyphrase is in the description, 0 otherwise.
7. **Keyword.** Even if authors of Web pages are not required to classify their published resources, they usually add some keywords in order to be properly indexed by search engines. Since these terms are labels generated by the authors themselves, we consider these terms as meaningful keyphrases. The keyword feature is then computed as a boolean value which is set to 1 if the keyphrase is one of the keywords proposed by the author of the Web page, 0 otherwise.

In the scoring and ranking step, all the above features are used in order to compute a score (named *keyphraseness*) for each candidate keyphrase. The keyphraseness is a weighted combination of the evaluated features, and in particular, given a candidate keyphrase  $p$ , the keyphraseness is computed as

$$\text{keyphraseness}(p) = \sum_i w_i * f_i(p)$$

where:  $f_i(p)$  is the value of the  $i$ -th feature for  $p$  and  $w_i$  is the weight assigned to the  $i$ -th feature.

A preliminary experimentation was carried out for identifying a proper set of weights for the features: a first prototype was implemented for collecting the opinions of a restricted set of subjects about the accuracy of the extracted keyphrases. By using this feedback, we identified the weights currently assigned to the features, which are the same for uni-grams, bi-grams, and tri-grams. However, future work will also investigate the idea of using different weights for uni-grams, bi-grams, and tri-grams since they have different characteristics. For example, unigrams extracted from a Web page are more frequent than bi-grams and trigrams. This preliminary experimentation allowed us to identify the weights of the features reported in Table [II](#).

**Table 1.** The weights assigned to the features

Feature Name	Weight
phrase frequency	0.5
POS value	0.5
phrase depth	0.6
wikipedia	0.9
title	0.9
description	0.6
keyword	0.6

The weights shown in Table 1 are used to compute the keyphraseness of the candidate phrases extracted from Web pages and then, the obtained lists of unigrams, bi-grams, and tri-grams, are ranked according to their keyphraseness.

Finally, the keyphrases associated with higher scores (higher keyphraseness) are recommended in the final keyphrase filtering step. We decided to extract the two top scored unigrams, the five top scored bi-grams, and the three top scored tri-grams since this setting generated the best results during a preliminary analysis. The reader can also notice that we use keyphraseness only for ordering the keyphrases and for this reason we do not need to normalize the keyphraseness in  $[0, 1]$ .

### 3 Evaluation

Web pages are usually not classified with keyphrases by their authors and this lack had a strong impact on our evaluation procedure. In fact there are not freely available datasets which can be used to execute an automatic evaluation of the described mechanism. For this reason we decided to exploit a live evaluation involving a set of volunteers which had the task of judging the accuracy of the results returned by our approach. Moreover, due to the lack of keyphrases associated to Web pages, we could not use KEA for comparing our results to one of the state of the art mechanisms. In fact, the KEA mechanism needs to be trained by using a corpus of annotated documents. This is a strong limitation since, at the best of our knowledge, there are not freely available APIs for extracting ranked keyphrases from Web pages. In order to face this issue we decided to use as baseline approach a system where keyphrases are scored and ranked according only to their frequencies. This choice seems reasonable since, as our approach does, the baseline approach takes into account only the information available in a specific document (without considering the characteristics of the documents in a specific collection). This baseline mechanism is still domain independent and the results are not biased by the characteristics of a specific corpus. More specifically, the baseline mechanism assigns a score to the set of candidate keyphrases according to their frequency: the most frequent keyphrases obtain an higher score. By using the score assigned to keyphrases, the baseline mechanism can extract the two top scored uni-grams, the five top



scored bi-grams, and the three top scored tri-grams. The final set of keyphrases is then built by these 10 filtered keyphrases.

The results returned by both our mechanism and the baseline approach were evaluated by using a Web application where a set of volunteers judged the accuracy of the results. Since our approach is mainly aimed at supporting the users of social tagging systems, we built a Web based application which simulates the interaction of a user with a social tagging system. By using this application, the volunteers could submit an URL and then the evaluation framework returned to the users a list of suggested keyphrases for the specific Web page. The list of returned keyphrases was built by merging the results produced by both the proposed approach and the baseline mechanism. However, the two sets of keyphrases were presented to the evaluators in a random order.

By merging the keyphrases without a specific order we avoided to bias the human evaluators since they were not able to recognize the keyphrases returned by one of the two compared approaches.

The evaluators had to vote each returned keyphrase by using the following 5-Likert scale: **Excellent** - The keyphrase is very meaningful, it reports relevant facts, people, topics or other elements which characterize the Web page; **Good** - The keyphrase is still significant for classifying the document, but it is not the best: the keyphrase reports facts, people, topics or other elements which characterize the Web page, but are more weakly connected to the main content of the page; **Neutral** - You are not sure about the significance of the keyphrase for the document; **Poor** - The keyphrase does not properly describe the contents; **Very Poor** - The keyphrase does not make sense.

The evaluation involved 26 volunteers (20 men and 6 women) who worked for two weeks. The volunteers were students and workers. The oldest participant was 63 years old, the youngest was 22 years old and the average age was 37 years. The volunteers evaluated the keyphrases generated for 209 Web pages written in Italian and in English.

We used the Normalized Discounted Cumulative Gain (NDCG) metric [11] to evaluate the experimental results. The NDCG metric is commonly used in Information Retrieval in order to evaluate the accuracy of ranking mechanisms. This measure is specifically used in scenarios where the ranked results are associated to different relevance levels, since it takes into account both the position and the usefulness (or gain) of the results. In other words, the NDCG metric evaluates a ranking mechanism according to its capability of placing the most relevant resources in the higher positions of the generated ranking. Technically, given a ranked list of resources returned by the evaluated mechanism, where the resource (in our case the keyphrase) in position  $i$  is associated to a relevance level  $rel_i$  (in our case the position is defined by our algorithm and the relevance by one of the evaluators), the NDCG computes the gain for this list as follows

$$DCG = rel_1 + \sum_{i=2}^n \frac{rel_i}{\log_2 i}$$

where  $n$  is the number of results in the ranked list and in our specific case  $n$  is equal to 10. In our evaluation the graded relevance scale is defined by the following relevance levels: Excellent = 4; Good = 3; Neutral = 2; Poor = 1; Very poor = 0. The DCG is then used to quantify the accuracy of a response generated by a ranking mechanism according to both a fixed relevance scale and the opinions of an evaluator.

By computing the DCG over each evaluation provided by our evaluators, we obtained an assessment of the accuracy for each evaluated Web page. These DCGs are then normalized with respect to the ideal rankings (i.e., the DCGs of the rankings generated by placing the most relevant results in the higher positions) to compute the NDCG and a higher NDCG corresponds to a more accurate approach.

Table 2 reports the 8 different NDCG values computed for evaluating and comparing the accuracy of the top 5 and top 10 keyphrases extracted by: (i) our approach from Web pages written in Italian (DIKpEW\_Ita); (ii) the baseline system from Web pages written in Italian (Base\_Ita); (iii) our approach from Web pages written in English (DIKpEW\_Eng); (iv) the baseline system from Web pages written in English (Base\_Eng).

**Table 2.** Performance of DIKpEW compared to the baseline mechanism

	NDCG@5	NDCG@10
<b>Base_Ita</b>	0.484	0.437
<b>DIKpEW_Ita</b>	0.558	0.614
<b>Base_Eng</b>	0.485	0.576
<b>DIKpEW_Eng</b>	0.523	0.686

According to the results showed in the table our approach outperforms the baseline mechanism. Moreover, the accuracy of the results computed for the Web pages in Italian are comparable to the accuracy for the Web pages in English. This means that the noise introduced by the translation in English does not significantly lowers the accuracy of the results. This can be justified in two ways: (i) the weight of the keyphrase depends on a set of statistical features which discard possible incorrect translations; (ii) the Wikipedia feature allows us to throw out (or at least to assign to lower positions) the bi-grams and tri-grams which have not a clear meaning.

## 4 Conclusion

In this work we presented an approach which is aimed at supporting the users of social tagging systems in classifying Web pages. In particular, the proposed approach identifies n-grams from a Web document for suggesting meaningful labels for the specific resource. An experimental evaluation showed that the proposed approach is plausible and future analysis will investigate if the proposed

approach can produce better results for specific topics or specific sets of Web pages (blogs, newspapers, etc.).

The proposed approach can extract keyphrases which appear already in a given document. Future work will focus on overcoming this limitation by navigating other knowledge sources such as Wikipedia, Wordnet or a specific domain ontology. In such a way it is possible to produce meaningful tags constituted by uni-grams, bi-grams, and tri-grams which are not contained in the text, and that are the result of a domain reasoning activity. A future work will investigate the problem of identifying a suitable threshold in the value of keyphraseness above/below which to accept/reject a candidate keyphrase.

## References

1. The rise of crowdsourcing. Website, <http://www.wired.com/wired/archive/14.06/crowds.html>
2. Barker, K., Cornacchia, N.: Using Noun Phrase Heads to Extract Document Keyphrases. In: Hamilton, H.J. (ed.) Canadian AI 2000. LNCS (LNAI), vol. 1822, pp. 40–52. Springer, Heidelberg (2000)
3. Bracewell, D.B., Ren, F., Kuroiwa, S.: Multilingual single document keyword extraction for information retrieval. In: Proceedings of the 2005 IEEE International Conference on Natural Language Processing and Knowledge Engineering, Wuhan, pp. 517–522 (2005)
4. Brin, S., Page, L.: The anatomy of a large-scale hypertextual web search engine. *Computer Networks* 30(1-7), 107–117 (1998)
5. D'Avanzo, E., Magnini, B., Vallin, A.: Keyphrase extraction for summarization purposes: the lake system at duc2004. In: DUC Workshop, Human Language Technology Conference/North American Chapter of the Association for Computational Linguistics Annual Meeting, Boston, USA (2004)
6. Ferrara, F., Pudota, N., Tasso, C.: A Keyphrase-Based Paper Recommender System. In: Agosti, M., Esposito, F., Meghini, C., Orto, N. (eds.) IRCDL 2011. CCIS, vol. 249, pp. 14–25. Springer, Heidelberg (2011)
7. Frank, E., Paynter, G.W., Witten, I.H., Gutwin, C., Nevill-Manning, C.G.: Domain-specific keyphrase extraction. In: Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence, pp. 668–673. Morgan Kaufmann Publishers, San Francisco (1999)
8. Howe, J.: *Crowdsourcing: Why the Power of the Crowd Is Driving the Future of Business*, 1st edn. Crown Publishing Group, New York (2008)
9. Hulth, A.: Improved automatic keyword extraction given more linguistic knowledge. In: Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing, pp. 216–223. Association for Computational Linguistics, Morristown (2003)
10. Hulth, A., Megyesi, B.B.: A study on automatically extracted keywords in text categorization. In: ACL-44: Proc. of the 21st Int. Conf. on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics, vol. 44, pp. 537–544. ACL, Morristown (2006)
11. Järvelin, K., Kekäläinen, J.: Cumulated gain-based evaluation of ir techniques. *ACM Transaction on Information Systems* 20(4), 422–446 (2002)
12. Jones, S., Paynter, G.W.: An evaluation of document keyphrase sets. *Journal of Digital Information* 4(1) (2003)

13. Justeson, J., Katz, S.: Technical terminology: some linguistic properties and an algorithm for identification in text. *Natural Language Engineering* 1, 9–27 (1995)
14. Krulwich, B., Burkey, C.: Learning user information interests through the extraction of semantically significant phrases. In: Hearst, M., Hirsh, H. (eds.) *AAAI 1996 Spring Symposium on Machine Learning in Information Access*, pp. 110–112. AAAI Press, California (1996)
15. Litvak, M., Last, M.: Graph-based keyword extraction for single-document summarization. In: *Proceedings of the Workshop on Multi-Source Multilingual Information Extraction and Summarization*, pp. 17–24. ACL, Morristown (2008)
16. Micarelli, A., Gasparetti, F., Biancalana, C.: Intelligent Search on the Internet. In: Stock, O., Schaerf, M. (eds.) *Reasoning, Action and Interaction in AI Theories and Systems. LNCS (LNAI)*, vol. 4155, pp. 247–264. Springer, Heidelberg (2006)
17. Mihalcea, R., Tarau, P.: TextRank: Bringing order into texts. In: Dekang, L., Dekai, W. (eds.) *Proc. of Empirical Methods in Natural Language Processing*, pp. 404–411. Association for Computational Linguistics, Barcelona (2004)
18. Porter, M.F.: An algorithm for suffix stripping. In: *Readings in Information Retrieval*, pp. 313–316 (1997)
19. Pudota, N., Dattolo, A., Baruzzo, A., Ferrara, F., Tasso, C.: Automatic keyphrase extraction and ontology mining for content-based tag recommendation. *International Journal of Intelligent Systems, Special Issue: New Trends for Ontology-Based Knowledge Discovery* 25, 1158–1186 (2010)
20. Pudota, N., Dattolo, A., Baruzzo, A., Tasso, C.: A New Domain Independent Keyphrase Extraction System. In: Agosti, M., Esposito, F., Thanos, C. (eds.) *IR-CDL 2010. CCIS*, vol. 91, pp. 67–78. Springer, Heidelberg (2010)
21. Turney, P.: Learning to extract keyphrases from text. Technical Report ERB-1057, National Research Council, Institute for Information Technology (1999)
22. Witten, I.H., Paynter, G.W., Frank, E., Gutwin, C., Nevill-Manning, C.G.: Kea: practical automatic keyphrase extraction. In: *Proceedings of the Fourth ACM Conference on Digital Libraries*, pp. 254–255. ACM, New York (1999)

# Learning to Recognize Critical Cells in Document Tables

Nicola Di Mauro<sup>1,2</sup>, Stefano Ferilli<sup>1,2</sup>, and Floriana Esposito<sup>1,2</sup>

<sup>1</sup> Dipartimento di Informatica, LACAM Laboratory  
Università degli Studi di Bari “Aldo Moro”  
{ndm,ferilli,esposito}@di.uniba.it

<sup>2</sup> Centro Interdipartimentale per la Logica e sue Applicazioni  
Università degli Studi di Bari “Aldo Moro”

**Abstract.** Tables are among the most informative components of documents, because they are exploited to compactly and intuitively represent data, typically for understandability purposes. The needs are to identify and extract tables from documents, and, on the other hand, to be able to extract the data they contain. The latter task involves the understanding of a table structure. Due to the variability in style, size, and aims of tables, algorithmic approaches to this task can be insufficient, and the exploitation of machine learning systems may represent an effective solution. This paper proposes the exploitation of a first-order logic representation, that is able to capture the complex spatial relationships involved in a table structure, and of a learning system that can mix the power of this representation with the flexibility of statistical approaches. The obtained encouraging results suggest further investigation and refinement of the proposal.

## 1 Introduction

The main motivation underlying the birth and spread of libraries consisted in collecting large quantities of documents, usually in paper format, with preservation and access objectives. Each library was typically characterized by a specific focus-of-interest, that established the direction and limits according to which the collections were developed, thus helping users to have in one place a complete landscape of the information they were interested in. As a technological counterpart, digital libraries have the same aims and objectives, but dealing with documents in digital format. This change of medium has a dramatic impact on the management of the collections, and on their exploitation by end-users. Huge quantities of documents can be easily collected and spread all over the world using the Internet, however, the users may experience difficulties in properly retrieving the data they are interested in. Information Retrieval (IR) and Machine Learning (ML) technologies can provide automatic tools to support such activities.

The identification of relevant documents that can satisfy the users' query is the subject of the IR research field. Of course, to provide more effective results the

automatic techniques should move from the purely lexical aspects of document to those concerning their semantics, which is still an open research issue. But having a thorough understanding of the document content is not only useful to support search and retrieval. It is also a fundamental requirement to be able to correlate the documents, and in particular the data and information they carry. In particular, a component of documents that is usually very informative and information-dense are tables. Authors use tables to compactly represent many important data in a small space, to draw more attention from readers, or for information comparison [10]. Thus, the availability of automatic components that can identify tables in documents, and that are able to understand the table structure, would be a precious support to extract the knowledge they contain, represent it formally (e.g., using a relational Database) and make it available to people and/or other software (e.g., using semantic technologies that are being developed nowadays). This paper proposes a set of intelligent techniques that cover the steps going from a document in digital format up to the identification of tables and the understanding of their structure, and particularly focuses on the exploitation of Machine Learning methodologies and systems for understanding the table structure.

## 2 Preliminaries

Many works are present in the literature concerning tables and their analysis, focusing on different objectives, aspects and problems. Some concern theoretical contributions, such as the distinction between genuine tables (tables aimed at representing and organizing meaningful information) and non-genuine ones (tables just aimed at obtaining a spatial partition of the page, as in most Web documents) [16]. This is a relevant issue, since, according to [2], only 1% Web tables are genuine. Others face more practical problems, such as table boundary identification [13] and table structure decomposition [9], or the classification of tables according to their type of content and intended exploitation. In addition to table data extraction, table functionality analysis (aimed at understanding table types, functions, and purposes) is another crucial task for table understanding, table data sharing and reuse [10]. Yet other researchers focus on applications such as table search [12] or table classification [16]. Indeed, accurately extracting tables from document repositories is a challenging problem, but also selecting interesting tables from the set of collected tables is an open issue.

As concerns the table identification step, we considered the DOMINUS framework [5] for document processing and management, and extended it with suitable techniques for table recognition. DOMINUS provides an integrated and general framework to manage digital libraries in which most knowledge-intensive tasks are carried out using intelligent techniques, among which Expert System and symbolic Machine Learning technologies play a predominant role. After submission, documents in different digital formats are processed to identify their layout structure, then to identify the kind of document and its relevant components, to extract the content from selected components and to exploit such a content

for indexing and information extraction purposes. Hence, while the layout analysis phase is involved in table recognition, the information extraction step is concerned with table structure identification and subsequent content extraction.

As to table recognition, DOMINUS deals with two kinds of digital documents: born-digital ones and digitized ones (typically obtained by scanning legacy paper sources). This distinction is relevant because the basic document components (text, images, geometric shapes—and specifically lines) are explicitly represented in born-digital documents (such as PDF ones), but not in digitized ones (usually coming in the form of raster images). Thus, in the latter case, suitable image processing techniques must be applied to identify them. In particular, horizontal and vertical lines are fundamental components for table recognition, although not sufficient (some tables do not show a perfect grid for visually highlighting their cell organization). Thus, in the case of document images, a variation of the Hough transform, specifically focused on horizontal/vertical lines, and on the identification of line segments, was developed and integrated in the DOMINUS framework. Then, the set of lines and other content blocks in a page were fed as an input to an expert module in charge of identifying and collecting the subsets of elements that together make up a table. Expert Systems technology was exploited because there is no standard representation for tables, and the many different styles used by different authors can vary significantly as regards the alignment of the content of rows and columns, the use of horizontal/vertical lines to separate portions of the table, and the position of the table in the page. Moreover, some tables are particularly tricky due to the presence of blank cells, or of cells that span several rows and/or columns. The expert component, whose detailed description is outside the scope of this paper, was able to identify and extract most tables in different kinds of documents, with some difficulties on very small tables and on multi-column documents.

Caption<sub>1</sub>

...

Caption<sub>m</sub>

Stub	Column heading
Row heading	Data

Note<sub>1</sub>

...

Note<sub>n</sub>

**Fig. 1.** Table structure according to Nagy et al.

As to table structure identification, our work specifically stems from a research stream carried on by Nagy et al. [15], specifically concerned with the extraction of the table structure and with its formal manipulation aimed at transposing the content into a relational representation that can be integrated in a typical database. They presented [15] a method based on *header paths* for efficient and complete extraction of labeled data from tables meant for humans. Header paths are a purely syntactic representation of visual tables that can be transformed (*factored*) into existing representations of structured data such as category trees, relational tables, and RDF triples.

A table contains a rectangular configuration of data cells, each of which can be uniquely referred by a row and a column index. As reported in [15], a table contains a set of *content-cells* that can be identified by a *column-header path* and a *row-header-path*. The table segmentation process aims at identifying four *critical cells* useful to partition the table into *stub*, *row header*, *column header*, and *delta* regions. In particular, this setting is concerned with six kinds of table-related elements, as shown in Figure 1.

**Caption.** A text placed above the table, aimed at explaining it;

**Data.** The set of cells containing the actual information carried by the table;

**Column Heading.** The table cells placed above the table data, aimed at explaining part of the dimensions according to which the data are organized;

**Row Heading.** The table cells placed to the left of the table data, aimed at explaining the remaining part of the dimensions according to which the data are organized;

**Stub.** One or more cells that correspond to the intersection between the horizontal projection of the row heading and the vertical projection of the column heading;

**Notes.** One or more text lines following the table, aimed at explaining portions of its content.

Some details should be pointed out. First of all, the notes are optional, and hence might be missing in some tables. The row and column headings may be quite complex, when the table is intended to represent data that are conceptually distributed along more than two dimensions (as a side effect, this event typically causes the presence of cell content that spans over many rows or columns). The stub can be made up of just one cell (if both the row headings consist of a single column, and the column headings consist of a single row) or of many cells (in case of composite row and/or column headings); it may be empty, but it often contains a meta-header aimed at explaining the row and/or column headings.

Thus, although the mutual position of these elements is known and fixed, identifying the specific boundaries of each of them may become very complex. To do this Nagy et al. [15] adopt an algorithmic approach, leveraging typical patterns. High accuracy should be required if the table data in the available documents are to be extensively and precisely extracted. Due to the many different kinds of tables that can be found in documents, and to the many different ways in which information can be organized in tables, we believe that a significant high accuracy cannot be reached by hand-written rules, but the characterizing essence of



the above elements can be captured only using automatic techniques provided by Machine Learning.

### 3 Proposed Approach

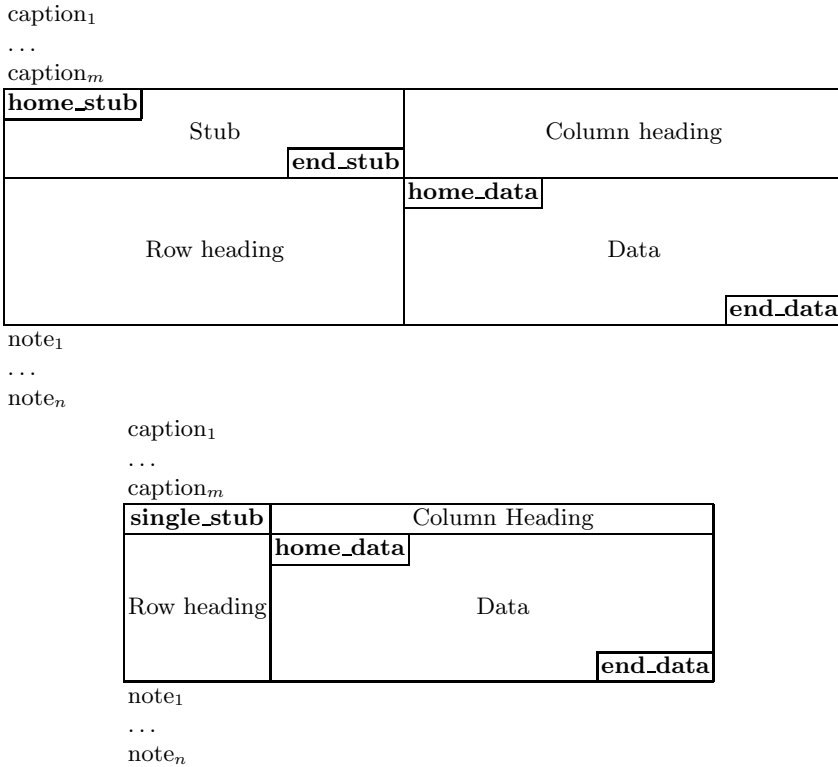
The first question to answer for applying Machine Learning to table structure recognition is the type of approach to be used. To answer this question, several aspects must be considered. A fundamental one concerns the kind of representation to be exploited. In this respect, it is quite clear that the most important feature to understand a table lies in its spacial structure, which in turn is made up of several relationships among the cells (both spacial and content ones). Indeed, it is self-evident that, when trying to understand a table, and specifically its components as described above, these are the parameters that any human considers. As a consequence, propositional techniques don't have a sufficient representational power to handle this kind of complexity, and first-order logic approaches must be considered. In particular, the following features/predicates were deemed as profitable for table description:

- Table boundaries
- Columns and Rows, and adjacency between them
- Cells and their belonging to a given row and column
- Cell content type

It should be noted that, in the proposed setting, the whole set of elements (stub, table cells and headings, caption, notes) associated to a table is represented in a Comma Separated Values (CSV) file, and hence in this file not only the actual table elements, but also caption and notes (if any) are represented as cells. Thus, there is no structural hint in the CSV file to distinguish different kinds of elements. In particular, caption and notes are considered as belonging to a single cell (typically in the first column), and the content of multi-row or multi-column cells is assumed to be placed in the (top-left)-most cell.

Another issue is the choice of classes to be learned. A straightforward possibility would be learning, for each cell, the type of table component to which it belongs. However, this would cause a significant growth in the number of examples, which would affect computational costs, and would be more difficult to handle in the subsequent classification phase (because each cell would be classified independently of the others (so that, for example, a data cell might be identified in the heading section). To solve the former problem, and to reduce the impact of the latter, a different solution was adopted. Four classes were defined as shown in Figure 2, that are non-redundant and are sufficient, alone, to univoquely determine the whole table structure:

- home\_stub.** The top-left cell in the stub;
- end\_stub.** The bottom-right cell in the stub;
- home\_data.** The top-left cell in the data;
- end\_data.** The bottom-right cell in the data.



**Fig. 2.** Classes for the table structure learning problem

In fact, either `end_stub` or `home_data` is redundant, because the `home_data` cell is always assumed to be placed just one column to the right, and one row below, the `end_stub`. Conversely, if classes are to be considered mutually exclusive, an additional class must be included, to specifically identify the case of a stub in which `home_stub` and `end_stub` coincide:

**single\_stub.** The stub cell, in the case of a single-cell stub.

Indeed, the captions can be identified as the content cell above the `home_stub` row, and the notes as the content cells below the `end_data` row; the column heading cells are those in the columns to the right of the `end_stub` column and in the rows between the `home_stub` row and the `end_stub` row; the row heading cells are those in the rows below the `end_stub` row and in the columns between the `home_stub` column and the `end_stub` column.

The last question concerns how rigid the learned models should be. Due to the problem being very multi-faceted, and to the lack of stable criteria to identify the table components, it is desirable that the learned models are quite flexible, with a preference for statistical approaches over purely logical ones.

### 3.1 Lynx: A Statistical Relational Learning Approach

The SRL approach **Lynx** [3] was used here to tackle the specific problem of critical cells identification in tables. **Lynx** implements a probabilistic query-based classifier, using first-order logic as a representation language. A first-order *alphabet* consists of a set of *constants*, a set of *variables*, a set of *function symbols*, and a non-empty set of *predicate symbols*. Both function symbols and predicate symbols have a natural number (its *arity*) assigned. A *term* is a constant symbol, a variable symbol, or an  $n$ -ary function symbol  $f$  applied to  $n$  terms  $t_1, t_2, \dots, t_n$ . An atom  $p(t_1, \dots, t_n)$  is a predicate symbol  $p$  of arity  $n$  applied to  $n$  terms  $t_i$ . An atom  $l$  and its negation  $\bar{l}$  are said to be (resp., positive and negative) *literals*. **Lynx** adopts the relational framework, and the corresponding query mining algorithm, reported in [4].

**Feature Construction via Query Mining.** The first step of **Lynx** carries out a feature construction process by mining frequent queries with an approach similar to that reported in [11]. The algorithm for frequent relational query mining is based on the same idea as the generic level-wise search method, known in data mining from the Apriori algorithm [1]. The algorithm starts with the most general queries. Then, at each step it tries to specialize all the candidate frequent queries, discarding the non-frequent queries and storing those whose length is equal to the user specified input parameter `maxsize`. For each new refined query, semantically equivalent queries are detected (using the  $\theta_{OI}$ -subsumption relation [7]) and discarded. The algorithm uses a background knowledge  $\mathcal{B}$  containing the examples and a set of constraints that must be satisfied by the generated queries, among which: `maxsize(M)`, maximal query length; `type(p)` and `mode(p)`, denote, respectively, the type and the input/output mode of the predicate's arguments  $\mathbf{p}$ , used to specify a language bias; `key([p1, p2, ..., pn])` specifies that each query must have one of the predicates  $p_1, p_2, \dots, p_n$  as a starting literal. Given a set of relational examples  $D$  defined over a set of classes  $C$ , the *frequency* of a query  $p$ ,  $\text{freq}(p, D)$ , corresponds to the number of examples  $s \in D$  such that  $p$  subsumes  $s$ .

**Query-Based Classification.** After identifying the set of frequent queries, the next question is how to use them as features in order to correctly classify unseen examples. Let  $\mathcal{X}$  be the input space of relational examples, and  $\mathcal{Y} = \{1, 2, \dots, Q\}$  denote the finite set of possible class labels. Given a training set  $D = \{(X_i, Y_i) | 1 \leq i \leq m\}$ , where  $X_i \in \mathcal{X}$  is a single relational example and  $Y_i \in \mathcal{Y}$  is the label associated to  $X_i$ , the goal is to learn a function  $h : \mathcal{X} \rightarrow \mathcal{Y}$  from  $D$  that predicts the label for each unseen instance. Let  $\mathcal{P}$ , with  $|\mathcal{P}| = d$ , be the set of constructed features obtained in the first step of the **Lynx** system (the queries mined from  $D$ ), as previously reported. For each example  $X_k \in \mathcal{X}$  we can build a  $d$ -component vector-valued random variable  $\mathbf{x} = (x_1, x_2, \dots, x_d)$  where each  $x_i \in \mathbf{x}$  is 1 if the query  $p_i \in \mathcal{P}$  subsumes example  $X_k$ , and 0 otherwise, for each  $1 \leq i \leq d$ .

Using Bayes' theorem, if  $p(Y_j)$  describes the prior probability of class  $Y_j$ , then the posterior probability  $p(Y_j | \mathbf{x})$  can be computed from  $p(\mathbf{x} | Y_j)$  as

$p(Y_j|\mathbf{x}) = \frac{p(\mathbf{x}|Y_j)p(Y_j)}{\sum_{i=1}^Q p(\mathbf{x}|Y_i)p(Y_i)}$ . Given a set of discriminant functions  $g_i(\mathbf{x})$ ,  $i = 1, \dots, Q$ , a classifier is said to assign vector  $\mathbf{x}$  to class  $Y_j$  if  $g_j(\mathbf{x}) > g_i(\mathbf{x})$  for all  $j \neq i$ . Taking  $g_i(\mathbf{x}) = P(Y_i|\mathbf{x})$ , the maximum discriminant function corresponds to the *maximum a posteriori* (MAP) probability. For minimum error rate classification, the following discriminant function will be used

$$g_i(\mathbf{x}) = \ln p(\mathbf{x}|Y_i) + \ln P(Y_i). \quad (1)$$

We are considering a multi-class classification problem involving discrete features. In this problem the components of vector  $\mathbf{x}$  are binary-valued and conditionally independent. In particular, let the component of vector  $\mathbf{x} = (x_1, \dots, x_d)$  be binary valued (0 or 1). We define  $p_{ij} = \text{Prob}(x_i = 1|Y_j)_{\substack{i=1,\dots,d \\ j=1,\dots,Q}}$  with the components of  $\mathbf{x}$  being statistically independent for all  $x_i \in \mathbf{x}$ . The factors  $p_{ij}$  can be estimated by frequency counts on the training examples, as  $p_{ij} = \text{support}_{Y_j}(p_i)$ .

By assuming conditional independence we can write  $P(\mathbf{x}|Y_i)$  as a product of the probabilities of the components of  $\mathbf{x}$ . Given this assumption, a particularly convenient way of writing the class-conditional probabilities is as follows:  $P(\mathbf{x}|Y_j) = \prod_{i=1}^d (p_{ij})^{x_i} (1-p_{ij})^{1-x_i}$ . Hence, Eq. 1 yields the discriminant function

$$g_j(\mathbf{x}) = \ln p(\mathbf{x}|Y_j) + \ln p(Y_j) = \sum_{i=1}^d x_i \ln \frac{p_{ij}}{1-p_{ij}} + \sum_{i=1}^d \ln(1-p_{ij}) + \ln p(Y_j). \quad (2)$$

The factor corresponding to the prior probability for class  $Y_j$  can be estimated from the training set as  $p(Y_i) = \frac{|\{(X,Y) \in D \text{ s.t. } Y=Y_i\}|}{|D|}$ ,  $1 \leq i \leq Q$ . The minimum probability of error is achieved by the following decision rule: decide  $Y_k$ ,  $1 \leq k \leq Q$ , if  $\forall j, 1 \leq j \leq Q \wedge j \neq k : g_k(\mathbf{x}) \geq g_j(\mathbf{x})$ , where  $g_i(\cdot)$  is defined as in Eq. 2.

**Feature Selection with Stochastic Local Search.** After constructing a set of features, and presenting a method to use those features to classify unseen examples, the problem is how to find a subset of these features that optimizes prediction accuracy. The optimization problem of selecting a subset of features with a superior classification performance may be formulated as follows. Let  $\mathcal{P}$  be the constructed original set of queries, and let  $f : 2^{|\mathcal{P}|} \rightarrow \mathbb{R}$  be a function scoring a selected subset  $X \subseteq \mathcal{P}$ . The problem of feature selection is to find a subset  $\hat{X} \subseteq \mathcal{P}$  such that  $f(\hat{X}) = \max_{Z \subseteq \mathcal{P}} f(Z)$ . An exhaustive approach to this problem would require examining all  $2^{|\mathcal{P}|}$  possible subsets of the feature set  $\mathcal{P}$ , making it impractical for even small values of  $|\mathcal{P}|$ . The use of a stochastic local search procedure [8] allows to obtain *good* solutions without having to explore the whole solution space.

Given a subset  $P \subseteq \mathcal{P}$ , for each example  $X_j \in \mathcal{X}$  we let the classifier find the MAP hypothesis  $\hat{h}_P(X_j) = \arg \max_i g_i(\mathbf{x}_j)$  by adopting the discriminant function reported in Eq. 1, where  $\mathbf{x}_j$  is the feature based representation of example  $X_j$  obtained using queries in  $P$ . Hence the initial optimization problem corresponds to minimize the expectation  $E[\mathbf{1}_{\hat{h}_P(X_j) \neq Y_j}]$  where  $\mathbf{1}_{\hat{h}_P(X_j) \neq Y_j}$  is the characteristic function of training example  $X_j$  returning 1 if  $\hat{h}_P(X_j) \neq Y_j$ ,

and 0 otherwise. Finally, given  $D$  the training set with  $|D| = m$  and  $P$  a set of features, the number of classification errors made by the Bayesian model is  $err_D(P) = mE[\mathbf{1}_{\hat{h}_P(X_j) \neq Y_j}]$ .

Consider a *combinatorial optimization* problem, where one is given a discrete set  $X$  of solutions and an objective function  $f : X \rightarrow \mathbb{R}$  to be minimized, and seek a solution  $x^* \in X$  such that  $\forall x \in X : f(x^*) \leq f(x)$ . A method to find high-quality solutions for a combinatorial problem consists of a two-step approach made up of a greedy construction phase followed by a perturbative local search [8]. GRASP [6] solves the problem of the limited number of different candidate solutions generated by a greedy construction search method by randomizing the construction method. GRASP is an iterative process combining at each iteration a construction and a local search phase. In the construction phase a feasible solution is built, and then its neighborhood is explored by the local search. The Lynx system includes an implementation of the GRASP procedure in order to perform the feature selection task, as reported in [3].

## 4 Problem Characterization and Validation

Lynx has been applied to a dataset consisting of 100 table descriptions. The dataset [1] is a collection of tables randomly selected from ten large statistical web sites [14]. HTML tables are represented in Comma Separated Value (CSV) files. Information about each table cell (its contained value and its absolute position in terms of row and column index) are used to provide its relational representation to be exploited by Lynx. The goal is to correctly predict the label of the critical cells belonging to each table. In particular, each table cell has been labeled to belong to one of the following classes: `caption`, `note`, `home_data`, `end_data`, `home_stub`, `end_stub`, and `single_stub`.

Figure 3 reports a sample table description adopting the relational language we used. In particular, predicate `label/2` indicates the corresponding class label of a cell; `row/3` (resp., `col/3`) define the position and the identifier of a row (resp., column) belonging to the table; `next_row/2` (resp., `next_col/2`) denote the spatial relationship between two adjacent rows (resp., columns); and finally, `cell/4` specifies the type of a cell. Given a table (also including caption and notes, if any), a `row/3` (resp., `col/3`) atom is introduced for each row (resp., column) of the CSV file, reporting as arguments the table identifier, the row (resp., column) identifier, and its index. Then, suitable `next_row/2` (resp., `next_col/2`) atoms are introduced to link adjacent rows (resp., columns) to each other in the proper sequence. Finally, for each cell a `cell/4` atom is added, reporting as arguments the corresponding identifier, the associated row and column identifiers, and the type of content.

Given a training set made up of the relational descriptions of critical cells belonging to each table, Lynx is applied in order to construct the relevant relational features maximizing the likelihood on the training data, as reported in

<sup>1</sup> The DocLab Dataset for Evaluating Table Interpretation Methods available at [http://www.iapr-tc11.org/mediawiki/index.php/Datasets\\_List](http://www.iapr-tc11.org/mediawiki/index.php/Datasets_List)

```

doc_table(c10076).
label(c10076_2_1, single_stub).
label(c10076_1_1, caption).
label(c10076_3_2, home_data).
...
row(c10076, c10076_r_1, 1).
next_row(c10076_r_1, c10076_r_2).
row(c10076, c10076_r_2, 2).
next_row(c10076_r_2, c10076_r_3).
row(c10076, c10076_r_3, 3).
...
col(c10076, c10076_c_1, 1).
next_col(c10076_c_1, c10076_c_2).
col(c10076, c10076_c_2, 2).
next_col(c10076_c_2, c10076_c_3).
col(c10076, c10076_c_3, 3).
...
cell(c10076_1_1, c10076_r_1, c10076_c_1, alphanumeric).
cell(c10076_2_1, c10076_r_2, c10076_c_1, empty).
cell(c10076_2_2, c10076_r_2, c10076_c_2, integer).
cell(c10076_3_2, c10076_r_3, c10076_c_2, numericSymbol).
...

```

**Fig. 3.** An example of a relational table description

Section 3.1. After this first step the system build a model composed of probabilistic query such as `label(A)`, `cell(B,A,C,D)`, `next_row(C,E)`, `cell(B,-,E,D)`, whose corresponding class probabilities are  $p(q|\text{note}) = 0.507$ ,  $p(q|\text{home\_data}) = 0.944$ ,  $p(q|\text{caption}) = 0.497$ ,  $p(q|\text{single\_stub}) = 0.628$ ,  $p(q|\text{end\_data}) = 0.000$ ,  $p(q|\text{end\_stub}) = 0.714$ , and  $p(q|\text{home\_stub}) = 0.548$ . These probabilistic queries are then used to predict critical cells belonging to testing tables.

Table 1 reports the results obtained with `Lynx` with a 10-fold cross validation in terms of accuracy, Conditional Log Likelihood (CLL), and areas under the Receiver operating characteristic (ROC) and Precision Recall (PR) curve.

**Table 1.** Accuracy, CLL, AUC of ROC and PR with a 10-fold cross validation

		AUC-ROC	AUC-PR
	caption	0,984	0,951
	note	0,986	0,978
	home stub	0,987	0,925
	end stub	0,983	0,825
	single stub	0,989	0,810
	home data	0,991	0,968
	end data	1,000	0,998
	<b>Mean</b>	0,989	0,922
	<b>Dev.St.</b>	0,006	0,075

		Accuracy	CLL
Mean	90,69	-4,89	
Dev.St.	0,017	2,38	

As we can see from the table, the results are very promising. The two labels on which the system obtains best results are those regarding the data region. While, it has some difficulties in correctly classifying the labels `single_stub` and `end_stub`. The next step towards improving these results is to use some collective classification techniques.

## 5 Conclusions

Tables are very informative components of documents, that compactly represent many inter-related data. It would be desirable to extract these data in order to make them available also outside the document. This requires to understand a table structure. Machine learning solutions may help to deal with the extreme variability in table styles and structures. We propose to exploit a first-order logic representation to capture the complex spatial relationships involved in a table structure, and a learning system that can mix the power of this representation with the flexibility of statistical approaches. On a dataset including different kinds of tables, encouraging results were obtained.

As a future work we are trying to combine the prediction of single critical cell labels in order to improve the accuracy of the segmentation process. We will adopt a collective classification procedure whose aim should be to improve the likelihood of a prediction for a given label knowing the probability of the prediction made on the neighbor labeled cells with an iterative approach. The iterative procedure will combine expectation steps, predicting labels on the known distribution, and maximization steps, improving the probability distribution.

**Acknowledgment.** The research leading to this paper has been partially funded by the MIUR PRIN 2009 project “A multi-relational approach to spatial and spatio-temporal data mining”.

## References

1. Agrawal, R., Srikant, R.: Mining sequential patterns. In: Proceedings of the International Conference on Data Engineering, pp. 3–14 (1995)
2. Cafarella, M., Halevy, A., Wang, Z., Wu, E., Zhang, Y.: Webtables: Exploring the power of tables on the web. In: Proceedings of VLDB (2008)
3. Di Mauro, N., Basile, T.M.A., Ferilli, S., Esposito, F.: Optimizing Probabilistic Models for Relational Sequence Learning. In: Kryszkiewicz, M., Rybinski, H., Skowron, A., Raś, Z.W. (eds.) ISMIS 2011. LNCS, vol. 6804, pp. 240–249. Springer, Heidelberg (2011)
4. Esposito, F., Di Mauro, N., Basile, T., Ferilli, S.: Multi-dimensional relational sequence mining. *Fundamenta Informaticae* 89(1), 23–43 (2008)
5. Esposito, F., Ferilli, S., Basile, T.M., Di Mauro, N.: Machine learning for digital document processing: From layout analysis to metadata extraction. In: Marinai, S., Fujisawa, H. (eds.) *Machine Learning in Document Analysis and Recognition*. SCI, vol. 90, pp. 105–138. Springer, Heidelberg (2008)

6. Feo, T., Resende, M.: Greedy randomized adaptive search procedures. *Journal of Global Optimization* 6, 109–133 (1995)
7. Ferilli, S., Di Mauro, N., Basile, T.M.A., Esposito, F.:  $\theta$ -Subsumption and Resolution: A New Algorithm. In: Zhong, N., Raś, Z.W., Tsumoto, S., Suzuki, E. (eds.) *ISMIS 2003. LNCS (LNAI)*, vol. 2871, pp. 384–391. Springer, Heidelberg (2003)
8. Hoos, H., Stützle, T.: *Stochastic Local Search: Foundations & Applications*. Morgan Kaufmann Publishers Inc., San Francisco (2004)
9. Kieninger, T.: Table structure recognition based on robust block segmentation. In: *Proc. Document Recognition V*, vol. 3305, pp. 22–32. SPIE (1998)
10. Kim, S., Liu, Y.: Functional-based table category identification in digital library. In: *International Conference on Document Analysis and Recognition*, pp. 1364–1368 (2011)
11. Kramer, S., De Raedt, L.: Feature construction with version spaces for biochemical applications. In: *Proceedings of the 18th International Conference on Machine Learning*, pp. 258–265. Morgan Kaufmann Publishers Inc. (2001)
12. Liu, Y., Bai, K., Mitra, P., Giles, C.L.: Tableseer: Automatic table metadata extraction and searching in digital libraries categories and subject descriptors. In: *Proceedings of JCDL 2007*, pp. 91–100 (2007)
13. Liu, Y., Mitra, P., Giles, C.: Identifying table boundaries in digital documents via sparse line detection. In: *Proceedings of CIKM 2008* (2008)
14. Nagy, G., Padmanabhan, R., Jandhyala, R.C., Silversmith, W., Krishnamoorthy, M.S.: Table metadata: Headers, augmentations and aggregates. In: *Ninth IAPR International Workshop on Document Analysis Systems* (2010)
15. Nagy, G., Seth, S.C., Jin, D., Embley, D.W., Machado, S., Krishnamoorthy, M.S.: Data extraction from web tables: The devil is in the details. In: *International Conference on Document Analysis and Recognition*, pp. 242–246 (2011)
16. Wang, Y., Hu, J.: A machine learning based approach for table detection on the web. In: *Proceedings of WWW*, pp. 242–250 (2002)



# Document Image Understanding through Iterative Transductive Learning

Michelangelo Ceci, Corrado Loglisci, Lucrezia Macchia,  
Donato Malerba, and Luciano Quercia

Dipartimento di Informatica, Università degli Studi di Bari “Aldo Moro”  
{ceci,loglisci}@di.uniba.it,  
{donato.malerba,lucrezia.macchia}@uniba.it, luciano.quercia@gmail.com

**Abstract.** In Document Image Understanding, one of the fundamental tasks is that of recognizing semantically relevant components in the layout extracted from a document image. This process can be automatized by learning classifiers able to automatically label such components. However, the learning process assumes the availability of a huge set of documents whose layout components have been previously manually labeled. Indeed, this contrasts with the more common situation in which we have only few labeled documents and abundance of unlabeled ones. In addition, labeling layout documents introduces further complexity aspects due to multi-modal nature of the components (textual and spatial information may coexist). In this work, we investigate the application of a relational classifier that works in the transductive setting. The relational setting is justified by the multi-modal nature of the data we are dealing with, while transduction is justified by the possibility of exploiting the large amount of information conveyed in the unlabeled layout components. The classifier bootstraps the labeling process in an iterative way: reliable classifications are used in subsequent iterative steps as training examples. The proposed computational solution has been evaluated on document images of scientific literature.

## 1 Introduction

The recognition of semantically relevant components in the layout extracted from a document image is based on domain-specific knowledge, which is represented in very different forms (e.g. formal grammars or production rules). Several prototypical document image understanding systems have been developed by manually encoding the required knowledge (e.g., DeLoS [14]). However, the layout of documents, even for the same publisher, may change considerably. To prevent obsolescence of the developed systems, it is necessary to continuously update the required knowledge, which is unfeasible if based only on manual encoding.

In order to guarantee *versatility* of Document Image Analysis Systems [1], that is, guarantee competence over a broad and precisely specified class of document images, the application of machine learning methods has been investigated for almost two decades [10] [5]. Operatively, a human operator provides a document image analysis system with images of documents and then detects and labels

semantically relevant layout components from which document structures are induced. This *supervised* learning approach, though providing some flexibility, still does not ensure the key requirement of versatility. Indeed, to acquire the necessary knowledge on a really broad class of documents, supervised learning methods may require a large set of labeled documents. This contrasts with the common situation in which only few labeled training documents are available due to the significant cost of manual annotation. Therefore, it is important to exploit the large amount of information potentially conveyed by unlabeled documents.

Two main settings have been proposed in the literature to exploit information contained in both labeled and unlabeled data: the *semi-supervised* setting and the *transductive* setting [15]. The former is a type of inductive learning: the learned function is used to make predictions on any possible example. The latter is only interested in making predictions for the given set of unlabeled data. When the set of documents to label is known a priori, the transductive setting is more suitable, since it is an easier problem than (semi-supervised) induction. In this paper, we propose a transductive approach where unlabeled documents are used to reprioritize models learned from labeled documents alone. Indeed, while discriminative learning methods base their decisions on the posterior probability  $p(y|x)$ , the transductive learning method uses unlabeled documents to improve the estimate of the prior probability  $p(x)$ , and hence correct the posterior probability  $p(y|x)$  by assuming some form of dependence with  $p(x)$ .

The proposed learning method follows a logic-based approach in which models are represented by a set of rules expressed in relational logic and documents are represented as facts in the same formalism. So, to “understand” the layout structure of an unlabeled document, rules are matched against the relational description of the document layout. The relational representation of document layout and rules is motivated by the fact that layout objects can be related by a number of spatial relationships, such as distance, directional or topological relationships. The study of relational learning in a transductive setting has received little attention (see [4], [11], [13]) while the application of transductive relational learning to bootstrap the labeling process of document image collections remains still unexplored. This work extends the research reported in [6], by introducing an iterative bootstrapping framework and by extending empirical evaluation to additional datasets. In the iterative bootstrapping framework, at each iteration, the algorithm expands the training set by including (originally unlabeled) examples for which the classification is considered to be reliable.

The paper is organized as follows. In Section 2, we define the problem to be solved. Sections 3 and 4 are devoted to the presentation of the method. Finally, experimental results are reported in Section 5 and some conclusions are drawn.

## 2 Motivations and Problem Definition

The recognition of semantically relevant layout components in document images is part of a complex transformation process of document images into a structured symbolic form. This transformation is articulated into several steps. Initial

processing steps include binarization, skew detection, and noise filtering. Then, the document image is segmented into several layout components, such as text lines, half-tone images, line drawings or graphics (this step is called layout analysis). The interpretation or understanding of document images follows layout analysis. It aims to associate a logical label (e.g. title, abstract of a scientific paper, picture of a newspaper) to semantically relevant layout components, as well as to extract relevant relationships between logical components (e.g., reading order). Document image understanding is typically based on layout information, such as the relative positioning of layout components or the size of layout components, as well as on content information (e.g., textual, graphical). This is the case of the work reported in this paper, where the association of logical labels to layout components is based on both layout information and textual information. However, the novelty here is mainly in the strategy applied to learn a classifier which can be used to recognize semantically relevant components.

In this work we investigate this issue and propose a transductive method for learning classifiers from training data represented in relational formalism. In a formal way, the problem is defined as follows:

*Given:*

- a database schema  $SC$  which consists of a set of  $h$  relational tables  $\{T_0, \dots, T_{h-1}\}$ , a set PK of primary keys on the tables in  $SC$ , and a set FK of foreign key constraints on the tables in  $SC$ ,
- a target relation  $T \in SC$  (that represents layout components) and a target discrete attribute  $Y$  in  $T$ , different from the primary key of  $T$ , whose domain is the finite set  $\{C_1, C_2, \dots, C_L\}$  (Logical label),
- the projection  $T'$  of  $T$  on all attributes of  $T$  except  $Y$ ,
- a training (working) set that is an instance  $TS$  ( $WS$ ) of the database schema  $SC$  with known (unknown) values for  $Y$ ;

*Find:* the most accurate classification of  $Y$  for examples in  $WS$ .

In this work, the classification of  $Y$  is based on an approach that exploits both the relational data mining setting and the classical Naïve Bayesian framework.

More precisely, given an object  $E \in WS$  to be classified, a classical naïve Bayes classifier assigns  $E$  to the class  $C_i$  that maximizes the *posterior probability*  $P(C_i|E)$ . By applying the Bayes theorem,  $P(C_i|E)$  is expressed as follows:

$$P(C_i|E) = P(C_i) \cdot P(E|C_i) / P(E). \quad (1)$$

In fact, the decision on the class that maximizes the posterior probability can be made only on the basis of the numerator, that is  $P(C_i) \cdot P(E|C_i)$ , since  $P(E)$  is independent of the class  $C_i$ . The probability  $P(C_i|E)$  can then be used to identify examples  $E$  for which the classification is reliable. This property can be used to iteratively extend the training data by propagating the most reliable decisions when bootstrapping the labeling process.

In [\[1\]](#), the main problem is in the computation of  $P(E|C_i)$ . By following the main intuition in [\[2\]](#), it is possible to consider a set  $\mathfrak{R}$  of association rules to define a suitable decomposition of the likelihood  $P(E|C_i)$  à la naïve Bayes in

order to simplify the probability estimation problem. In particular, if  $\mathfrak{R}(E) \subseteq \mathfrak{R}$  is the set of first order association rules whose antecedent covers  $E$ ,  $P(E|C_i)$  is:

$$P(E|C_i) = P\left(\bigwedge_{R_j \in \mathfrak{R}(E)} \text{antecedent}(R_j)|C_i\right). \quad (2)$$

The straightforward application of the naïve Bayes independence assumption to all literals in  $\bigwedge_{R_j \in \mathfrak{R}(E)} \text{antecedent}(R_j)$  is not correct, since it may lead to underestimating  $P(E|C_i)$  when several similar clauses in  $\mathfrak{R}(E)$  are considered for the class  $C_i$ . To prevent this problem the authors resort to the logical notion of factorization. Details are reported in [2].

Although this approach would potentially be used in this application, two main limitations could prevent its actual applicability: *i*) It does not exploit the transductive learning setting. *ii*) As in most associative classifiers, extracted association rules do not permit to adequately characterize classes.

To overcome these limitations, in this paper, we use Emerging Patterns (EPs) instead of association rules in order to discover a characterization of classes and we use this characterization in a transductive classifier. In fact, emerging patterns discovery is a descriptive data mining task which aims at detecting significant differences between objects of distinct classes. EPs are introduced in [8] as a particular kind of patterns (or multi-variate features) whose support significantly changes from one data class to another: the larger the difference of pattern support, the more interesting the pattern. Change in pattern support is estimated in terms of the support ratio (or *growth rate*). EPs with sharp change in support (high growth rate) can be used to characterize classes.

### 3 Mining Emerging Patterns with SPADA

Data mining research has provided several solutions (e.g. [8]) for the task of emerging patterns discovery but only one attempt [3] has been done to deal with relational data. In this work, we exploit the system SPADA [12], originally designed for *relational* frequent patterns discovery, for mining emerging patterns.

SPADA represents relational data *à la* Datalog, a logic programming language with no function symbols specifically designed to implement deductive databases. SPADA distinguishes between the set  $S$  of *reference* (or *target*) *objects*, which are the main subject of analysis, and the sets  $R_k$ ,  $1 \leq k \leq m$ , of *task-relevant* (or *non-target*) objects, which are related to the former and can contribute to account for the variation. From a database viewpoint,  $S$  corresponds to the target table  $T \in SC$  and each  $R_k$  corresponds to a different relational table  $T_i \in SC$ . A unit of analysis corresponds to a tuple in  $t \in T$  and to all tuples in the database related to  $t$  according to foreign key constraints.

In the following sub-sections, the document description and the learning strategy are described, as it has been modified to mine emerging patterns.

**Document Description.** In the logic framework adopted by SPADA, a relational database is boiled down into a deductive database where properties of

**Table 1.** The complete list of used predicates

Layout structure	Locational features	$x\_pos\_center/2$
		$y\_pos\_center/2$
	Geometrical features	$height/2$
		$width/2$
	Topological features	$on\_top/2$
		$to\_right/2$
	Aspatial feature	$type\_of/2$
Logical structure	Logical features	application dependent (e.g., $abstract/1$ )
Text	Textual features	application dependent (e.g., $text\_in\_abstract/2$ )

both reference objects (which are the main subject of the analysis) and task-relevant objects (which are relevant for the task at hand, but not necessarily the main subjects of the analysis) are represented in the extensional part  $D_E$ , while the domain knowledge is expressed as a normal logic program which defines the intensional part  $D_I$ . As an example, we report a fragment of the extensional part of a deductive database  $D$  which describes multimodal information which can be extracted from any document image:

*block(b1). block(b2). ... height(b2,[11..54]). width(b1,[7..82]). ...  
 on\_top(b2,b1). ... on\_top(b2,b3). ... part\_of(b1,p1). part\_of(b2,p1). page\_first(p1).  
 ... abstract(b1). title(b2). ... text\_in\_abstract(b1,'base'). text\_in\_title(b2,'model')...*

In this example,  $b1$  and  $b2$  are two constants which denote as many distinct layout components (reference objects), while  $p1$  denotes a document page (task-relevant object). Predicate *block* defines a layout component, *part\_of* associates a block to a document page, *height* and *width* describe geometrical properties of layout components, *on\_top* expresses a topological relationship between layout components, *page\_first(p1)* refers to the position of the page in the document, *abstract* and *title* associate  $b1$  and  $b2$  with a logical label, *text\_in\_abstract* and *text\_in\_title* describe the textual content of the logical components.

The complete list of predicates is reported in Table 1. The aspatial feature *type\_of* specifies the content type of a layout component (e.g. image, text, horizontal line). Logical features are used to associate a logical label to a layout object and depend on the specific domain. In the case of scientific papers (considered in this work), possible logical labels are: *affiliation*, *page\_number*, *figure*, *caption*, *index\_term*, *running\_head*, *author*, *title*, *abstract*, *formulae*, *subsection\_title*, *section\_title*, *biography*, *references*, *paragraph*, *table*. Textual content is represented by means of another class of predicates, which are true when the term reported as second argument occurs in the layout component denoted by the first argument. Terms are automatically extracted by means of a text-processing module [7].

**The Mining Step.** The original algorithm of SPADA mines frequent patterns at multiple levels  $l$  of granularity in order to properly deal with hierarchies  $H_k$  of objects. When these are available, it is important to take them into account since patterns involving more abstract objects are better supported (although less precise). SPADA operates in two steps for each granularity level: i) pattern generation; ii) pattern evaluation. It takes advantage of statistics computed at granularity level  $l$  when computing the supports of patterns at the granularity

level  $l + 1$ . To discover emerging patterns, SPADA has been modified to mine patterns which characterize classes by detecting significant differences between the objects of these classes. This problem requires the following formulation:

*Given:*

- a set  $S$  of *reference objects*,
- a label value  $y \in Y = \{C_1, C_2, \dots, C_L\}$  associated to each reference object,
- some sets  $R_k$ ,  $1 \leq k \leq m$ , of *task-relevant objects*,
- a background knowledge  $BK$  including hierarchies  $H_k$  on objects in  $R_k$ ,
- $M$  granularity levels in the descriptions,
- a set of granularity assignments  $\Psi_k$  which associate each object in  $H_k$  with a granularity level,
- a couple of sets of thresholds  $minSup[l]$ ,  $minGR[l]$  for each granularity level,
- a language bias LB that constrains the search space;

*Find:* A set of multilevel emerging patterns  $\{F | supp_{C_i}(F) \geq minSup[l], GR_{C_i}(F) \geq minGR[l]\}$ .

In this formulation,  $supp_{C_i}(F)$  represents the support of the pattern  $F$  in the subset of reference objects labeled with  $C_i$  while the growth rate  $GR_{C_i}(F)$  is defined as:  $GR_{C_i}(F) = \frac{supp_{C_i}(F)}{supp_{-C_i}(F)}$  where  $supp_{-C_i}(F)$  is the support of the pattern  $F$  in the subset of reference objects labeled with  $c \in \{C_1, \dots, C_{i-1}, C_{i+1}, \dots, C_L\}$ .

To efficiently mine frequent patterns, SPADA prunes the search space by exploiting the monotonicity of the support. Let  $F'$  be a refinement of a pattern  $F$  (i.e.  $F'$  is more specific than  $F$ ). If  $F$  is an infrequent pattern for the class  $C_i$  (i.e.  $supp_{C_i}(F) < minSup$ ), then also  $supp_{C_i}(F') < minSup$ . This means that  $F'$  cannot be an emerging pattern that distinguishes  $C_i$  from  $-C_i$ . Hence, SPADA does not refine patterns which are infrequent in  $C_i$ .

Unluckily, the monotonicity property does not hold for the growth rate: a refinement of an emerging pattern whose growth rate is lower than the threshold  $minGR$  may or may not be an EP. However, also in this case, it is possible to prune the search space. According to [16], we modified the mining algorithm originally developed in SPADA in order to avoid to generate the refinements of a pattern  $F$  in the case that  $GR_{C_i}(F) = \infty$  (i.e.,  $supp_{C_i}(F) > 0$  and  $supp_{-C_i}(F) = 0$ ). Indeed, due to the monotonicity of support, for each pattern  $F'$  obtained as refinement of  $F$ :  $supp_{C_i}(F) \geq supp_{C_i}(F')$  then  $supp_{C_i}(F') = 0$ . Thereby,  $GR_{C_i}(F') = 0$  in the case that  $supp_{C_i}(F') = 0$ , while  $GR_{C_i}(F') = \infty$  in the case that  $supp_{C_i}(F') > 0$ . In the former case,  $F'$  is not worth to be considered. In the latter case, we prefer  $F$  to  $F'$  based on the Occams razor principle, according to which all things being equal, the simplest solution tends to be the best one ( $F$  has the same discriminating ability than  $F'$ ).

In our application domain, reference objects are all the logical components for which a logical label is specified. Task relevant objects are all the logical components (including undefined components) as well as pages and documents. The  $BK$  is used to specify the hierarchy of logical components (Figure 1). The  $BK$  also allows us to automatically associate information on page order to layout components, since the presence of some logical components may depend on the page order (e.g. author is in the first page).

```

article
+ -- heading
| + -- identification
| | + -- (title, author, affiliation)
| + -- synopsis
|   + -- (abstract, index_term)
+ -- content
| + -- final components
| | + -- (biography, references)
| + -- body
|   + -- (section_title, subsect_title, paragraph, caption, figure, formulae, table)
+ -- page_component
| + -- running_head
| + -- page_number
+ -- undefined

```

**Fig. 1.** Hierarchy of logical components

---

**Algorithm 1.** The iterative transductive learning algorithm.

---

**Input:**  $TS$  training data,  $WS$  working data.

**Output:**  $H$  working objects associated with labels

```

1:  $H \leftarrow \emptyset$ ;  $W' \leftarrow WS$ ;
2: while  $WS \neq \emptyset$  do
3:   Compute the score matrix  $\Xi = [score_{TS \cup H}(o_j, C_i)]_{o_j \in W', C_i \in \mathcal{Y}}$ ;
4:   Sort the objects in  $o_j \in W'$  according to  $\max_{C_i} (score_{TS \cup H}(o_j, C_i))$ ;
5:   Add all  $\langle o_j, \arg \max_{C_i} (score_{TS \cup H}(o_j, C_i)) \rangle$  to  $H$ , where  $o_j$  is one of the top  $\lfloor |WS|/k \rfloor$  objects
      in  $W'$ ;
6:   Remove the top  $\lfloor |WS|/k \rfloor$  objects from  $W'$ ;
7: end while

```

---

## 4 Transductive Classification

The transductive classifier implemented in our proposal is described in Algorithm 1, where at each iteration of the cycle at line 3, the algorithm labels objects belonging to the working set  $WS$  and uses a subset of them of size  $\lfloor |WS|/k \rfloor$  as training objects in the subsequent iteration, where  $k$  is a user defined parameter. The subset is created according to the function  $score_{TS \cup H}(o_j, C_i)$  which represents a membership score of an object  $o_j$  to the class  $C_i$ . This score is a growth rate based function which is estimated on the current training set  $TS \cup H$  and is computed by adapting the EP-based classifier CAEP [9] to the relational setting. The largest score determines the object's class.

In our case, it is computed on the basis of the subset of relational emerging patterns that cover the object to be classified. Formally, let  $o_j$  be the description of the object to be classified (an object is represented by a tuple in the target table and all the tuples related to it according to foreign key constraints),  $\mathfrak{R}(o_j) = \{F \in \mathfrak{R} \mid \exists \theta F \theta \subseteq o_j\}$  is the set of emerging patterns that cover the object  $o_j$ .

---

<sup>1</sup> This means that there are, at most,  $k + 1$  iterations.

The score of  $o_j$  on the class  $C_i$  is computed as follows:

$$score_{TS \cup H}(o_j, C_i) = \sum_{F \in \mathfrak{R}(o_j)} \frac{GR_{C_i}(F)}{GR_{C_i}(F) + 1} sup_{C_i}(F) \quad (3)$$

where  $GR_{C_i}(F)$  and  $sup_{C_i}(F)$  are computed on the current training set  $TS \cup H$ .

This measure may result in an inaccurate classifier in the case of unbalanced datasets that is, when training objects are not uniformly distributed over the classes. In order to mitigate this problem the authors in [9] proposed to normalize this score on the basis of the median of the scores obtained from training objects belonging to  $C_i$ . This results in the following classification function:

$$class_{TS \cup H}(o_j) = \arg \max_{C_i} \frac{score_{TS \cup H}(o_j, C_i)}{median_{ro \in TS \cup H}(score_{TS \cup H}(ro, C_i))} \quad (4)$$

where  $TS \cup H$  represents the training set.

However, in our case, the main problem comes from the different number of EPs that are extracted from different classes. This means that, in our case a different normalization that weights the number of EPs is necessary:

$$score_{TS \cup H}(o_j, C_i) = \frac{1}{|\mathfrak{R}(o_j)|} \sum_{F \in \mathfrak{R}(o_j)} \frac{GR_{C_i}(F)}{GR_{C_i}(F) + 1} sup_{C_i}(F) \quad (5)$$

Since  $sup_{C_i}(F)$  represents the probability that a reference object belonging to class  $C_i$  is covered by  $F$ , Equation (5) can be transformed as follows:

$$score_{TS \cup H}(o_j, C_i) = \frac{1}{|\mathfrak{R}(o_j)|} \sum_{F \in \mathfrak{R}(o_j)} \frac{GR_{C_i}(F)}{GR_{C_i}(F) + 1} P(F|C_i) \quad (6)$$

By applying the Bayes theorem:

$$score_{TS \cup H}(o_j, C_i) = \frac{1}{|\mathfrak{R}(o_j)|} \sum_{F \in \mathfrak{R}(o_j)} \frac{GR_{C_i}(F)}{GR_{C_i}(F) + 1} \frac{P(C_i|F)}{P(C_i)} \times P(F) \quad (7)$$

where  $P(C_i|F)$  can be estimated as the percentage of objects covering  $F$  in  $TS \cup H$  that belong to  $C_i$ .  $P(C_i)$  can be estimated as the percentage of objects in  $TS \cup H$  that belong to  $C_i$ . Finally,  $P(F)$  is the percentage of objects covering  $F$ . According to the transductive learning setting, this factor is estimated by considering the whole set of objects ( $TS \cup WS$ ). This would provide a more reliable estimation of  $P(F)$  (since obtained from a larger population of objects potentially coming from the same distribution).

$$P(F) = \frac{\#\{ro|ro \in TS \cup WS, \exists \theta F \theta \subseteq ro\}}{\#\{ro|ro \in TS \cup WS\}} \quad (8)$$



## 5 Experiments

The proposed approach has been applied to three different real-world datasets consisting of articles published in two international journals, namely IEEE TPAMI and Behavior Genetics (BG), and in the proceedings of the International Conference on Machine Learning (ICML). More precisely, the dataset TPAMI includes twenty-four multi-page papers corresponding to 217 document images from which we consider *abstract*, *affiliation*, *author*, *biography*, *caption*, *figure*, *formulae*, *index term*, *page number*, *paragraph*, *references*, *running head*, *section title*, *subsection title*, *table*, *title* as possible layout components. The dataset BG includes twenty-four single-page papers from which we consider *abstract*, *author*, *index term*, *page number*, *paragraph*, *references*, *running head*, *section title*, *title* as possible layout components. The dataset ICML includes thirty multi-page papers, corresponding to 240 document images from which we consider *abstract*, *affiliation*, *author*, *body*, *figure*, *index term*, *paragraph*, *section title*, *subsection title*, *table*, *title* as possible layout components.

The iterative transductive classification algorithm is evaluated considering the following experimental setups: 4-fold cross-validation in the case of TPAMI, 6-fold cross-validation in the case of BG and 5-fold cross-validation for ICML. Unlike the standard cross-validation, here one fold at a time is set aside to be used as the *training set* (and not as the *test set*). Small training set sizes allow us to validate the transductive approach, but may result in high error rates.

In the step of mining emerging patterns, three experimental schemes of the thresholds  $minGR$ ,  $minSup$  have been set: in the case of TPAMI  $minGR = \{1, 2, 8, 64\}$  and  $minSup = \{30\%, 40\%, 50\%\}$ , in the case of BG  $minGR = \{1, 2, 8, 64\}$  and  $minSup = \{10\%, 20\%, 30\%\}$ , while in the case of ICML  $minGR = \{1, 2, 8, 64\}$  and  $minSup = \{10\%, 20\%, 30\%\}$ . In Table 2 the average number of emerging patterns mined with different parameter values is reported. As expected, by increasing  $minSup$  and  $minGR$  values, the total number of EPs (sum of the number of EPs in the folds) is reduced. In particular, the number of EPs is more drastically reduced when increasing  $minSup$  than when increasing  $minGR$ . This means that there are several patterns which characterize a class (a specific layout component) and therefore present a high discriminative power with respect to components belonging to other classes.

Another consideration can be done on the number of EPs mined for each specific class (Table 3). We note that the layout components, for which the descriptions are more heterogeneous or which can be misclassified, are characterized by an higher number of EPs. Indeed, the components which present strong regularities (e.g., described with the same set of features) are those which can be more easily identified and which therefore generate a smaller set of EPs for the classification. Differently, the components which present low regularities can be erroneously labeled and therefore require an higher number of EPs to be discriminated from the others<sup>2</sup>. For instance, a figure can be more easily identified than an abstract layout component.

<sup>2</sup> The risk is that in these cases we can have overfitting problems.

**Table 2.** Total number of emerging patterns mined from TPAMI, BG and ICML

TPAMI	minSup (%)			BG	minSup (%)			ICML	minSup (%)		
<i>minGR</i>	30	40	50	<i>minGR</i>	10	20	30	<i>minGR</i>	10	20	30
1	528032	344798	254805	1	128327	88684	58603	1	386996	176407	114492
2	523274	341534	252355	2	126840	87644	58091	2	382639	173372	112476
8	516958	336733	248658	8	122591	84208	55718	8	376645	169406	109814
64	513503	334292	246843	64	121363	82980	54490	64	374736	167742	108595

**Table 3.** Minimum and maximum number of emerging patterns mined per class

TPAMI	minSup (%)		
<i>minGR</i>	30	40	50
1	min:11470(references) max:89008(index_term)	min:5450(figure) max:43422(abstract)	min:3319(figure) max:37475(abstract)
2	min:11394(references) max:88158(index_term)	min:5436(figure) max:42908(abstract)	min:3310(figure) max:37035(abstract)
8	min:11309(references) max:87124(index_term)	min:5364(figure) max:42085(abstract)	min:3271(figure) max:36304(abstract)
64	min:11276(references) max:86426(index_term)	min:5321(figure) max:41880(abstract)	min:3240(figure) max:36112(abstract)
BG	minSup (%)		
<i>minGR</i>	10	20	30
1	min:4380(references) max:45671(abstract)	min:4380(references) max:27342(author)	min:4380(references) max:15923(abstract)
2	min:4380(references) max:45179(abstract)	min:4380(references) max:26820(author)	min:4380(references) max:15825(abstract)
8	min:4218(references) max:43555(abstract)	min:4218(references) max:25713(author)	min:4218(references) max:15148(abstract)
64	min:4075(references) max:43171(abstract)	min:4075(references) max:25437(author)	min:4075(references) max:14764(abstract)
ICML	minSup (%)		
<i>minGR</i>	10	20	30
1	min:13923(body) max:169787(author)	min:5131(body) max:39728(abstract)	min:2780(body) max:27849(abstract)
2	min:13905(body) max:168468(author)	min:5120(body) max:38886(abstract)	min:2769(body) max:27213(abstract)
8	min:13843(body) max:166879(author)	min:5089(body) max:37828(abstract)	min:2756(body) max:26453(abstract)
64	min:13814(body) max:166671(author)	min:5065(body) max:37408(abstract)	min:2741(body) max:26152(abstract)

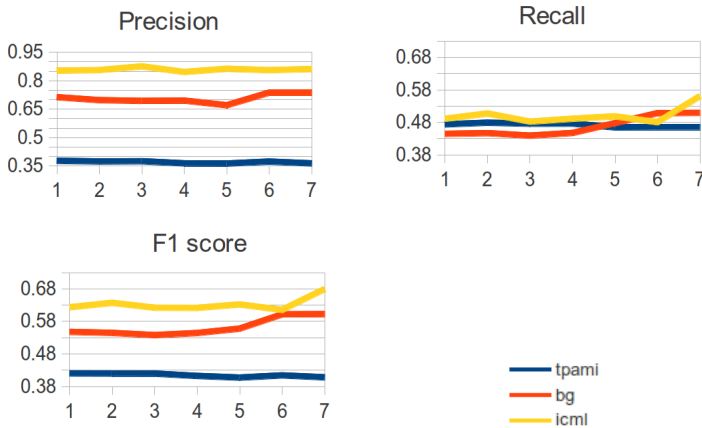
In Table 4, the macro average F1-score values are reported. Results are collected for different values of *minGR* and *minSup*. As we can see, better results are obtained when increasing *minGR* and/or when decreasing *minSup*. Indeed, higher values of *minGR* lead to exclude EPs with low discriminative capabilities and consider those with higher growth rate values with the result of (slightly) higher accuracy. While, when increasing *minSup* the number of EPs decreases and this leads to exclude models which, being infrequent, can characterize each class, with the result that the system has no enough information to discriminate among classes.

In Figure 2, precision, recall and *F1* values are plotted by varying the value of *k* (which regulates the number of iterations). While, by increasing the number of iterations there is no improvement in terms of precision, results in terms of recall show benefits coming from the iterative transductive (bootstrapping) approach. This means that, with the iterative transduction, the system is able to associate

**Table 4.** Macro average  $F1$ -score values on TPAMI, BG and ICML with  $k = 1$  and by varying  $minGR$  and  $minSup$ 

TPAMI		minSup (%)			BG		minSup (%)			ICML		minSup (%)		
$minGR$		30	40	50	$minGR$		10	20	30	$minGR$		10	20	30
1	0.2906	0.2949	0.2555		1	0.6359	0.6323	0.6199		1	0.3247	0.2791	0.2493	
2	0.3258	0.2694	0.2509		2	0.6548	0.6287	0.6091		2	0.3118	0.2762	0.2686	
8	0.3264	0.2689	0.2511		8	0.6566	0.6341	0.6135		8	0.3052	0.2988	0.1987	
64	0.3072	0.2684	0.2502		64	0.6411	0.6295	0.6142		64	0.4028	0.2969	0.1976	

to the correct class components that, otherwise, would remain unclassified. An exception is represented by TPAMI, where the system, due to the high number of components and to highly unbalanced data, is not able to reach good values of precision/recall. Obviously, a bad initial classification, negatively affects results of the iterative transductive approach.

**Fig. 2.** Macro average precision, recall and  $F1$ -score on TPAMI, BG and ICML by varying the value of  $k$ . Results for TPAMI are obtained with  $minGR = 8$  and  $minSup = 30$  while results for BG and ICML are obtained with  $minGR = 8$  and  $minSup = 10$ .

## 6 Conclusions

In this work, the induction of a classifier for the automated recognition of relevant layout components has been investigated. In particular, we have investigated the combination of transductive inference with principled relational classification in order to face the challenges posed by the application domain, characterized by complex and heterogeneous data, which are naturally modeled as several tables of a relational database, and characterized by the availability of a small (large) set of labeled (unlabeled) data. On the basis of an iterative bootstrapping approach, we exploit reliable classifications to classify other working examples in subsequent iterative steps. Interesting results on three real-world datasets are reported. They show that the iterative bootstrapping approach is able to increase recall of the obtained classifications.

## References

1. Baird, H.S., Casey, M.R.: Towards Versatile Document Analysis Systems. In: Bunke, H., Spitz, A.L. (eds.) DAS 2006. LNCS, vol. 3872, pp. 280–290. Springer, Heidelberg (2006)
2. Ceci, M., Appice, A.: Spatial associative classification: Propositional vs structural approach. *Journal of Intelligent Information Systems* 27(3), 191–213 (2006)
3. Ceci, M., Appice, A., Malerba, D.: Discovering Emerging Patterns in Spatial Databases: A Multi-relational Approach. In: Kok, J.N., Koronacki, J., Lopez de Mantaras, R., Matwin, S., Mladenić, D., Skowron, A. (eds.) PKDD 2007. LNCS (LNAI), vol. 4702, pp. 390–397. Springer, Heidelberg (2007)
4. Ceci, M., Appice, A., Malerba, D.: Transductive Learning for Spatial Data Classification. In: Koronacki, J., Raś, Z.W., Wierzchoń, S.T., Kacprzyk, J. (eds.) *Advances in Machine Learning I*. SCI, vol. 262, pp. 189–207. Springer, Heidelberg (2010)
5. Ceci, M., Berardi, M., Malerba, D.: Relational Data Mining and ILP for Document Image Understanding. *Applied Artificial Intelligence* 21(4-5), 317–342 (2007)
6. Ceci, M., Loglisci, C., Malerba, D.: Transductive Learning of Logical Structures from Document Images. In: Biba, M., Xhafa, F. (eds.) *Learning Structure and Schemas from Documents*. SCI, vol. 375, pp. 121–142. Springer, Heidelberg (2011)
7. Ceci, M., Malerba, D.: Classifying web documents in a hierarchy of categories: a comprehensive study. *Journal of Intelligent Information Systems* 28(1), 37–78 (2007)
8. Dong, G., Li, J.: Efficient mining of emerging patterns: Discovering trends and differences. In: *International Conference on Knowledge Discovery and Data Mining*, pp. 43–52. ACM Press (1999)
9. Dong, G., Zhang, X., Wong, L., Li, J.: CAEP: Classification by Aggregating Emerging Patterns. In: Arikawa, S., Nakata, I. (eds.) *DS 1999*. LNCS (LNAI), vol. 1721, pp. 30–42. Springer, Heidelberg (1999)
10. Esposito, F., Malerba, D., Semeraro, G.: Multistrategy learning for document recognition. *Applied Artificial Intelligence* 8(1), 33–84 (1994)
11. Krogel, M.-A., Scheffer, T.: Multi-relational learning, text mining, and semi-supervised learning for functional genomics. *Mach. Lear.* 57(1-2), 61–81 (2004)
12. Lisi, F.A., Malerba, D.: Inducing multi-level association rules from multiple relations. *Machine Learning* 55(2), 175–210 (2004)
13. Malerba, D., Ceci, M., Appice, A.: A relational approach to probabilistic classification in a transductive setting. *Engineering Applications of Artificial Intelligence* 22(1), 109–116 (2009)
14. Niyogi, D., Srihari, S.N.: Knowledge-based derivation of document logical structure. In: *ICDAR 1995: Proceedings of the Third International Conference on Document Analysis and Recognition*, vol. 1, p. 472. IEEE Computer Society, Washington, DC (1995)
15. Seeger, M.: Learning with labeled and unlabeled data. Technical report, Institute for Adaptive and Neural Computation. University of Edinburgh (2001)
16. Zhang, X., Dong, G., Ramamohanarao, K.: Exploring constraints to efficiently mine emerging patterns from large high-dimensional datasets. In: *Knowledge Discovery and Data Mining*, pp. 310–314 (2000)

# A Domain Based Approach to Information Retrieval in Digital Libraries

Fulvio Rotella<sup>1</sup>, Stefano Ferilli<sup>1,2</sup>, and Fabio Leuzzi<sup>1</sup>

<sup>1</sup> Dipartimento di Informatica – Università di Bari

{fulvio.rotella, stefano.ferilli, fabio.leuzzi}@uniba.it

<sup>2</sup> Centro Interdipartimentale per la Logica e sue Applicazioni – Università di Bari

**Abstract.** The current abundance of electronic documents requires automatic techniques that support the users in understanding their content and extracting useful information. To this aim, improving the retrieval performance must necessarily go beyond simple lexical interpretation of the user queries, and pass through an understanding of their semantic content and aims. It goes without saying that any digital library would take enormous advantage from the availability of effective Information Retrieval techniques to provide to their users. This paper proposes an approach to Information Retrieval based on a correspondence of the domain of discourse between the query and the documents in the repository. Such an association is based on standard general-purpose linguistic resources (WordNet and WordNet Domains) and on a novel similarity assessment technique. Although the work is at a preliminary stage, interesting initial results suggest to go on extending and improving the approach.

## 1 Introduction

The easy and cheap production of documents using computer technologies, plus the extensive digitization of legacy documents, have caused a significant flourishing of documents in electronic format, and the spread of Digital Libraries (DLs) aimed at collecting and making them available to the public, removing time and space barriers to distribution and fruition that are typical of paper material. In turn, the fact that anybody can produce and distribute documents (without even the cost of printing them) may negatively affect the average quality of their content. Although, as a particular kind of library, a DL has the mission of gathering a collection of documents which meets the quality standards chosen by the institution that maintains it, some repositories may adopt looser quality enforcing policies, and leave this responsibility to the authors, also due to the difficulty in manually checking and validating such a huge amount of material. In these cases, the effectiveness of document retrieval might be significantly tampered, affecting the fruition of the material in the repository as a consequence. Under both these attacks, anyone who is searching for information about a given topic is often overwhelmed by documents that only apparently are suitable for satisfying his information needs. In fact, most information in these documents is redundant, partial, sometimes even wrong or just unsuitable for the user's aims.

A possible way out consists in automatic instruments that (efficiently) return significant documents as an answer to user queries, that is the branch of interest of Information Retrieval (IR).

IR aims at providing the users with techniques for finding interesting documents in a repository, based on some kind of query. Although multimedia digital libraries are starting to gain more and more attention, the vast majority of the content of current digital document repositories is still in textual form. Accordingly, user queries are typically expressed in the form of natural language sentences, or sets of terms, based on which the documents are retrieved and ranked. This is clearly a tricky setting, due to the inherent ambiguity of natural language. Numerical/statistical manipulation of (key)words has been widely explored in the literature, but in its several variants seems unable to fully solve the problem. Achieving better retrieval performance requires to go beyond simple lexical interpretation of the user queries, and pass through an understanding of their semantic content and aims.

This work focuses on improving fruition of a DL content, by means of advanced techniques for document retrieval that try to overcome the aforementioned ambiguity of natural language. For this reason, we looked at the typical behavior of humans, when they take into account the possible meanings underlying the most prominent words that make up a text, and select the most appropriate one according to the context of the discourse. To carry out this approach, we used a well-known lexical taxonomy, and its extension to deal with domain categories, as a background knowledge.

The rest of this paper is organized as follows. After a brief recall of previous work on Information Retrieval, with a particular attention to techniques aimed at overcoming lexical limitations, toward semantic aspects, Section 3 introduces a new proposal for semantic information retrieval based on taxonomic information. Then, Section 4 proposes an experimental evaluation of the proposed technique, with associated discussion and evaluation. Lastly, Section 5 concludes the paper and outlines open issues and future work directions.

## 2 Related Work

Many works, aimed at building systems that tackle the Information Retrieval problem, exist in the literature. Most of such works are based on the ideas in [17], a milestone in this field. This approach, called *Vector Space Model* (VSM), represents a corpus of documents  $D$ , and the set of terms  $T$  appearing in those documents, as a  $T \times D$  matrix, in which the  $(i, j)$ -th cell contains a weight representing the importance of the  $i$ -th term in the  $j$ -th document (usually computed according to the number and distribution of its occurrences both in that document and in the whole collection). This allows to compute the degree of similarity of a user query to any document in the collection, simply using any geometrical distance measure on that space. Much research has been spent on developing effective similarity measures and weighting schemes, and on variations of their implementations to enhance retrieval performance. Most similarity

approaches [8, 16, 15] and weighting schemes [14, 13, 18] are based on inner product and cosine measure. Motivations came, on one hand, from the growth of the Web, and, on the other, from the success of some implementations in Web search engines. 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: e.g., synonymy (different words having similar meaning), and polysemy (words having many different meanings).

More recently, techniques based on dimensionality reduction have been explored for capturing the concepts present in the collection. The main idea behind these techniques is mapping both the documents in the corpus and the queries into a lower dimensional space that explicitly takes into account the dependencies between terms. Then, the associations provided by the low-dimensional representation can be used to improve the retrieval or categorization performance. Among these techniques, Latent Semantic Indexing (LSI) [3] and Concept Indexing (CI) [9] can be considered relevant. The former is a statistical method that is capable of retrieving texts based on the concepts they contain, not just by matching specific keywords, as in previous approaches. It starts from a classical VSM approach, and applies Singular Value Decomposition (SVD) to identify latent concepts underlying the collection, and the relationships between these concepts and the terms/documents. Since the concepts are weighted by relevance, dimensionality reduction can be carried out by filtering out less relevant concepts, and the associated relationships. In this way new associations emerge between terms that occur in similar contexts, and hence query results may include documents that are conceptually similar in meaning to the query even if they don't contain the same words as the query. The latter approach, CI, carries out an indexing of terms using concept decomposition (CD) [4] instead of SVD (as in the LSI). It represents a collection of documents in  $k$ -dimensions by first clustering the documents in  $k$  groups using a variant of the  $k$ -means algorithm [11], and considering each group as potentially representing a different concept in the collection. Then, the cluster centroids are taken as the axes of the reduced  $k$ -dimensional space. Although LSI and CI have had much success (e.g., LSI was implemented by Google) for their ability to reduce noise, redundancy, and ambiguity, they still pose some questions. First of all, their high computational requirements prevent exploitation in many digital libraries. Moreover, since they rely on purely numerical and automatic procedures, the noisy and redundant semantic information must be associated with a numerical quantity that must be reduced or minimized by the algorithms. Last but not least, a central issue is the choice of the matrix dimension [2].

### 3 A Domain-Based Approach

This section describes a proposal for a domain-based approach to information retrieval in digital libraries. In order to get rid of the constraints imposed by the syntactic level, we switch from the terms in the collection to their meaning by choosing a semantic surrogate for each word, relying on the support of

external resources. At the moment, we exploit WordNet [5], and its extension WordNet Domains [12], as readily available general-purpose resources, although the proposed technique applies to any other taxonomy.

The first step consists in off-line preprocessing the digital library in order to obtain, for each document, a list of representative keywords, to each of which the corresponding meaning will be associated later on. Using a system based on the DOMINUS framework [6], each document in the digital library is progressively split into paragraphs, sentences, and single words. In particular, the Stanford Parser [10] is used to obtain the syntactic structure of sentences, and the lemmas of the involved words. In this proposal, only nouns are considered and used to build a classical VSM weighted according to the TF\*IDF scheme. In addition to stopwords, typically filtered out by all term-based approaches, we ignore adverbs, verbs and adjectives as well, because their representation in WordNet is different than that of nouns (e.g., verbs are organized in a separate taxonomy), and so different strategies must be defined for exploiting these lexical categories, which will be the subject of future work. More specifically, only those nouns that are identified as keywords for the given documents, according to the techniques embedded in DOMINUS, are considered. In order to be noise-tolerant and to limit the possibility of including non-discriminative and very general words (i.e., common words that are present in all domains) in the semantic representation of a document, it can be useful to rank each document keyword list by decreasing TF\*IDF weight and to keep only the top items (say, 15) of each list.

The next step consists in mapping each keyword in the document to a corresponding synset (i.e., its semantic representative) in WordNet. Since this task is far from being trivial, due to the typical polysemy of many words, we adopt the one-domain-per-discourse (ODD) assumption as a simple criterion for Word Sense Disambiguation (WSD): the meanings of close words in a text tend to refer to the same domain, and such a domain is probably the dominant one among the words in that portion of text. Hence, to obtain such synsets, we need to compute for each document the prevalent domain. First we take from WordNet all the synsets of each word, then, for each synset, we select all the associated domains in WordNet Domains. Then, each domain is weighted according to the density function presented in [1], depending on the number of domains to which each synset belongs, on the number of synsets associated to each word, and on the number of words that make up the sentence. Thus, each domain takes as weight the sum of all the weights of synsets associated to it, which results in a ranking of domains by decreasing weight. This allows to perform the WSD phase, that associates a single synset to each term by solving possible ambiguities using the domain of discourse (as described in Algorithm 1). Now, each document is represented by means of WordNet synsets instead of terms.

The output of the previous step, for each document, is a list of pairs, made up of keywords and their associated synsets. All these synsets are partitioned into different groups using pairwise clustering, as shown in Algorithm 2: initially each synset makes up a different singleton cluster; then, the procedure works by iteratively finding the next pair of clusters to merge according to the *complete*



---

**Algorithm 1.** Find “best synset” for a word

---

**Input:** word  $t$ , list of domains with weights.

**Output:** best synset for word  $t$ .

```

bestSynset ← empty
bestDomain ← empty
for all synset( $s_t$ ) do
  maxWeight ←  $-\infty$ 
  optimalDomain ← empty
  for all domains( $d_s$ ) do
    if weight( $d_s$ ) > maxWeight then
      maxWeight ← weight( $d_s$ )
      optimalDomain ←  $d_s$ 
    end if
  end for
  if maxWeight > weight(bestDomain) then
    bestSynset ←  $s_t$ 
    bestDomain ← optimalDomain
  end if
end for

```

---

*link* strategy (shown in Algorithm 3), based on the similarity function proposed in [7]:

$$sf(i', i'') = sf(n, l, m) = \alpha \frac{l+1}{l+n+2} + (1-\alpha) \frac{l+1}{l+m+2}$$

where:

- $i'$  and  $i''$  are the two items (synsets in this case) under comparison;
- $n$  represents the information carried by  $i'$  but not by  $i''$ ;
- $l$  is the common information between  $i'$  and  $i''$ ;
- $m$  is the information carried by  $i''$  but not by  $i'$ ;
- $\alpha$  is a weight that determines the importance of  $i'$  with respect to  $i''$  (0.5 means equal importance).

In particular, we adopt a global approach based on all the information provided by WordNet on the two synsets, rather than on just one of their subsumers as in other measures in the literature. Indeed, we compute the distance between each pair ( $i', i''$ ) by summing up three applications of this formula, using different parameters  $n$ ,  $m$  and  $l$ . The first component works in depth, and obtains the parameters by counting the number of common and different hypernyms between  $i'$  and  $i''$ . The second one works in breadth, and considers all the synsets with which  $i'$  and  $i''$  are directly connected by any relationship in WordNet, and then takes the number of common related synsets as parameter  $l$ , and the rest of synsets, related to only  $i'$  or  $i''$ , as parameters  $n$  and  $m$ . Lastly, the third component is similar to the second one, but it considers the inverse relationships (incoming links) in the computation. The considered relationships in the last two measures are:

- *member meronymy*: the latter synset is a member meronym of the former;
- *substance meronymy*: the latter synset is a substance meronym of the former;
- *part meronymy*: the latter synset is a part meronym of the former;
- *similarity*: the latter synset is similar in meaning to the former;
- *antonymy*: specifies antonymous word;
- *attribute*: defines the attribute relation between noun and adjective synset pairs in which the adjective is a value of the noun;
- *additional information*: additional information about the first word can be obtained by seeing the second word;
- *part of speech based*: specifies two different relations based on the parts of speech involved;
- *participle*: the adjective first word is a participle of the verb second word;
- *hyperonymy*: the latter synset is a hypernym of the former.

*Example 1.* To give an idea of the breadth-distance between  $S_1$  and  $S_2$ , let us consider the following hypothetical facts in WordNet:

$$\begin{array}{ccc} rel_1(S_1, S_3) & rel_2(S_1, S_4) & rel_3(S_1, S_5) \\ & rel_4(S_2, S_5) & rel_5(S_2, S_6) \end{array}$$

for the direct component, and

$$\begin{array}{ccc} rel_1(S_7, S_1) & rel_2(S_8, S_1) & rel_3(S_9, S_1) \\ rel_4(S_9, S_2) & rel_5(S_3, S_2) & rel_2(S_8, S_2) \end{array}$$

for the inverse component, where  $rel_i$  represents one of the relationships listed above. In the former list, the set of synsets linked to  $S_1$  is  $\{S_3, S_4, S_5\}$  and the set of synsets linked to  $S_2$  is  $\{S_5, S_6\}$ . Their intersection is  $\{S_5\}$ , hence we have  $n = 2$ ,  $l = 1$ ,  $m = 1$  as parameters for the similarity formula. In the latter list, the set of synsets linked to  $S_1$  is  $\{S_7, S_8, S_9\}$  and the set of synsets linked to  $S_2$  is  $\{S_9, S_3, S_8\}$ , yielding  $n = 1$ ,  $l = 2$ ,  $m = 1$  as parameters for the similarity formula. The depth-distance component considers only hypernyms, and collects the whole sets of ancestors of  $S_1$  and  $S_2$ .

Now, each document is considered in turn, and each of its keywords votes for the cluster to which the associated synset has been assigned (as shown in Algorithm 4). The contribution of such a vote is equal to the TF\*IDF value established in the keyword extraction phase normalized on the sum of the weights of the chosen keywords. However, associating each document to only one cluster as its descriptor would be probably too strong an assumption. To smooth this, clusters are ranked in descending order according to the votes they obtained, and the document is associated to the first three clusters in this ranking. This closes the off-line preprocessing macro-phase, aimed at suitably partitioning the whole document collection according to different sub-domains. In our opinion, the pervasive exploitation of domains in this phase justifies the claim that the proposed approach is *domain-based*. Indeed, we wanted to find sets of similar synsets that might be usefully exploited as a kind of ‘glue’ binding together a sub-collection of documents that are consistent with each other. In this perspective, the obtained clusters can be interpreted as intensional representations of

---

**Algorithm 2.** Pairwise clustering of all detected synsets

---

**Input:**  $S$ : list of all synsets detected in WSD phase applied to the keywords;  $C$ : an empty set of clusters.

**Output:** set of clusters.

```

for all  $s_i \in S$  do
   $c_i \leftarrow s_i \mid c_i \in C$ 
end for
for all  $pair(s_i, s_j) \mid i \neq j$  do
  if  $completeLink(s_i, s_j)$  then
     $clustersAgglomeration(s_i, s_j)$ 
  end if
end for

```

---



---

**Algorithm 3.** Complete link between two clusters

---

**Input:**  $C1$ : former cluster;  $C2$ : latter cluster;  $T$ : the threshold for Ferilli et al. similarity measure.

**Output:** check outcome.

```

for all  $c_i \in C1$  do
  for all  $k_j \in C2$  do
    if  $similarityScore(c_i, k_j) < T$  then
       $return \rightarrow false$ 
    end if
  end for
end for
 $return \rightarrow true$ 

```

---

specific domains, and thus they can be exploited to retrieve the sub-collection they are associated to. Note that a cluster might correspond to an empty set of documents (when it was not in the 3 most similar clusters of any document in the collection).

The previous steps pave the way for the subsequent on-line phase, in which information retrieval is actually carried out. This phase starts with a user's query in natural language. The query undergoes the same grammatical preprocessing as in the off-line phase, yielding a set of words that are potentially useful to detect the best subset of documents to be presented as a result. For consistency with the off-line phase, only nouns are chosen among the words in the query. However, since the query is usually very short, keyword extraction is not performed, and all nouns are retained for the next operations. For each word, all corresponding synsets are taken from WordNet. Since WSD applied to the query would not be reliable (because it might be too short to identify a significant domain), we decided to keep all synsets for each word, and to derive from a single lexical query many semantic queries (one for each combination of synsets, one from each word). Specifically, given an  $n$ -term query, where the  $i$ -th term has associated

---

**Algorithm 4.** Association of documents to clusters

---

**Input:**  $D$ : the list of documents;  $W$ : the list of words of each document;  $S$ : the list of synsets of each document;  $C$ : the set of clusters.

**Output:** set of clusters with the assigned documents.

$V$ : vector of votes, one for cluster. Starting value: 0.

```

for all  $d_i \in D$  do
  for all  $w_i \in W$  do
     $s \leftarrow \text{getSynset}(w_i)$ 
     $c \leftarrow \text{getClusterOfSynset}(s)$ 
     $V.\text{getVoteOfCluster}(c, s.\text{getCluster}())$ 
  end for
   $\text{rankedList} \leftarrow \text{descendingOrdering}(V)$ 
  for all  $v_j \in V \mid 0 \leq j < 3$  do
     $\text{associateDocumentToCluster}(d_i, v_j.\text{getCluster}())$ 
  end for
end for

```

---

$n_i$  synsets,  $\prod_{i=1}^n n_i$  semantic queries are obtained, each of which represents a candidate disambiguation.

For each such query, a similarity evaluation is performed against each cluster that has at least one associated document, using the same complex similarity function as for clustering, that takes as input two sets of synsets (those in the query and those associated to the cluster), computes the distance between each possible pair of synsets taken from such sets, and then returns the maximum distance between all such pairs. This evaluation has a twofold objective: finding the combination of synsets that represents the best word sense disambiguation, and obtaining the cluster to which the involved words are most similar. The main motivation for which this phase considers only clusters that have at least one associated document is that, as already stated, clusters can be interpreted a set of descriptors for document subsets, and hence it makes sense keeping only those descriptors that are useful to identify the best set of documents according to the user's search. At this point, the best combination is used to obtain the list of clusters ranked by descending relevance, that can be used as an answer to the user's search. It should be pointed out that the ranked list is exploited, instead of taking just the best cluster, to avoid the omission of potentially useful results contained in positions following the top, this way losing information.

## 4 Evaluation

To understanding the contribution of each step in the overall result, we used a collection made up of 200 documents obtained by randomly drawing 50 documents from 4 Wikipedia top-categories (general science, music, politics, religion). A structured version of the Wikipedia dump was obtained exploiting the Java Wikipedia Library [19]. A selection of queries, with a corresponding performance

**Table 1.** Performance evaluation

#	Query	Outcomes	$P$	$P'$
1	creation of the mankind	[1 to 5] religion [6 to 10] science [+3] science	0.5	1.0
2	traditions and folks	[1 to 8] music [9 to 10] religion [+3] religion	0.8	1.0
3	ornaments and melodies	[1 to 8] music [9] science [10] religion	0.8	0.9
4	capitalism vs communism	[1 to 2] religion [3 to 10] politics [+4] politics	0.8	0.8
5	markets and new economy	[1 to 10] politics [+1] politics	1.0	1.0
6	gene structure and function	[1 to 2] science [3] religion [4] politics [5 to 10] science [+2] science	0.8	0.8

evaluation, is summarized in Table 1. For each query, the ranked list of most similar clusters was considered, and the top 10 documents were exploited for evaluating two performance measures: classical Precision  $P$ , expressing how many retrieved documents belong to the intended category of the query, and a looser version thereof  $P'$ , considering as good outcomes also documents in categories that are compatible with the query, even if that was not in the user intention. A first consideration is that the decision to take several clusters (not just the top-ranked one) improved the result for all queries as regards true positives. In addition to the best 10 documents used for computing  $P$  and  $P'$ , we have also reported (preceded by a '+' symbol) the number of immediately following documents that were nevertheless relevant for the query, which shows that good performance is not limited to top items only. Going beyond the purely numerical figures expressing the above measures, also a deeper insight into the specific cases reveals interesting aspects. For instance all results for query # 1 can be accepted as good, taking into account that a scientific perspective might correctly satisfy the user's search about the creation of the mankind, as well. Also for query # 2, it is quite agreeable that both traditions and folks are strictly related to religion as well as popular music. This motivated further analysis of some specific queries. In the following, for the sake of readability, when dealing with concepts both the synset code, and the set of associated terms, along with the corresponding gloss, will be reported. We will focus specifically on two sample queries purposely selected to help the reader understand the corresponding behavior.

The former is *ornaments and melodies*. Only 2 combinations were found, among which the best one was:

- *synset*: 103169390; *lemmas*: decoration, ornament and ornamentation; *gloss*: something used to beautify;
- *synset*: 107028373; *lemmas*: air, line, melodic line, melodic phrase, melody, strain and tune; *gloss*: a succession of notes forming a distinctive sequence.

This combination was recognized by the technique to be most similar to the following cluster:

- *synset*: 107044760; *lemmas*: symphonic music, symphony; *gloss*: a long and complex sonata for symphony orchestra;
- *synset*: 107033753; *lemmas*: mass; *gloss*: a musical setting for a Mass;
- *synset*: 107026352; *lemmas*: opera; *gloss*: a drama set to music, consists of singing with orchestral accompaniment and an orchestral overture and interludes;
- *synset*: 107071942; *lemmas*: genre, music genre, musical genre and musical style; *gloss*: an expressive style of music;
- *synset*: 107064715; *lemmas*: rock, rock 'n' roll, rock and roll, rock music, rock'n'roll and rock-and-roll; *gloss*: a genre of popular music originating in the 1950s, a blend of black rhythm-and-blues with white country-and-western;
- *synset*: 107043275; *lemmas*: concerto; *gloss*: a composition for orchestra and a soloist.

It's easy to note that this cluster contains elements that are consistent with each other, a positive result that we may trace back to the decision of using a complete link pair-wise clustering, which is more restrictive in grouping items. In particular, this cluster represents an intensional description of 8 documents returned as first (or more relevant) outcomes, all talking about music. Furthermore, it is noteworthy that this query result satisfies the initial aim, of retrieving query-related documents that do not necessarily contain the terms that are present in the query. Thus, the technique is actually able to go beyond simple lexical interpretation of the user queries, retrieving documents in which no occurrence of the words forming the query are present, even in cases in which those words are not present at all in the entire collection. The latter sample is *market and new economy*. It is made up of 2 nouns, yielding a total of 20 combinations to be analyzed, of which the system recognized as the best one the following:

- *synset*: 108424951; *lemmas*: market; *gloss*: the customers for a particular product or service;
- *synset*: 100192613; *lemmas*: economy, saving; *gloss*: an act of economizing; reduction in cost.

The most similar cluster was:

- *synset*: 108166552; *lemmas*: country, land, nation; *gloss*: the people who live in a nation or country;

- *synset*: 108179689; *lemmas*: populace, public, world; *gloss*: people in general considered as a whole;
- *synset*: 107965937; *lemmas*: domain, world; *gloss*: people in general, especially a distinctive group of people with some shared interest.

Here we obtained 8 main results talking about politics. As in the former case, we can appreciate both the benefits of returning as a result the ranked list of clusters instead just the best one, and the consistency of the cluster elements. Again, it should be noted that, although very simple, the WSD technique based on the one-domain-per-discourse assumption was able to select a strongly consistent solution.

## 5 Conclusions

This work proposed an approach to extract information from digital libraries trying to go beyond simple lexical matching, toward the semantic content underlying the actual aims of user queries. For all the documents in the corpus, after a keyword extraction phase, all keywords are disambiguated with a simple domain-driven WSD approach. The synsets obtained in this way are clustered, and each document is assigned to the cluster which contains more synsets related to its keywords. Then, given a user query, due to the typically low number of words in a query, that would affect the reliability of the WSD technique, all possible combinations of word meanings are considered, and the one that is most similar to a cluster is chosen. The outcome of the query presents the set of retrieved documents ranked by decreasing similarity of the associated cluster with such a combination. Preliminary experiments show that the approach can be viable, although extensions and refinements are needed to improve its effectiveness. In particular, the substitution of the ODD assumption with a more elaborated strategy for WSD might produce better results. Another issue regards incrementality: the current version of the approach requires a pre-processing, due to the underlying techniques for keyword extraction and clustering; this might be limiting when new documents are progressively included in the collection, a case that is very important in some digital libraries. Moreover, it might be interesting to evaluate the inclusion of adverbs, verbs and adjectives in order to improve the quality of the semantic representatives of the documents, and to explore other approaches to choose better intensional descriptions of each document.

## References

- [1] Angioni, M., Demontis, R., Tuveri, F.: A semantic approach for resource cataloguing and query resolution. *Communications of SIWN. Special Issue on Distributed Agent-based Retrieval Tools* 5, 62–66 (2008)
- [2] Bradford, R.B.: An empirical study of required dimensionality for large-scale latent semantic indexing applications. In: *Proceedings of the 17th ACM Conference on Information and Knowledge Management, CIKM 2008*, pp. 153–162. ACM, New York (2008)

- [3] 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), Atlanta, Georgia, vol. 25. American Society for Information Science (October 1988)
- [4] Dhillon, I.S., Modha, D.S.: Concept decompositions for large sparse text data using clustering. In: Machine Learning, pp. 143–175 (2001)
- [5] Fellbaum, C. (ed.): WordNet: An Electronic Lexical Database. MIT Press, Cambridge (1998)
- [6] Ferilli, S.: Automatic Digital Document Processing and Management: Problems, Algorithms and Techniques, 1st edn. Springer Publishing Company, Incorporated (2011)
- [7] 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 (LNAI), vol. 5883, pp. 131–140. Springer, Heidelberg (2009)
- [8] 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)
- [9] 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)
- [10] 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)
- [11] MacQueen, J.B.: Some methods for classification and analysis of multivariate observations. In: Le Cam, L.M., Neyman, J. (eds.) Proc. of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, vol. 1, pp. 281–297. University of California Press (1967)
- [12] Magnini, B., Cavaglià, G.: Integrating subject field codes into wordnet, pp. 1413–1418 (2000)
- [13] Robertson, S.E., Walker, S., Jones, S., Hancock-Beaulieu, M., Gatford, M.: Okapi at trec-3. In: TREC (1994)
- [14] Salton, G.: The SMART Retrieval System—Experiments in Automatic Document Processing. Prentice-Hall, Inc., Upper Saddle River (1971)
- [15] Salton, G.: Automatic term class construction using relevance—a summary of work in automatic pseudoclassification. *Inf. Process. Manage.* 16(1), 1–15 (1980)
- [16] Salton, G., McGill, M.: Introduction to Modern Information Retrieval. McGraw-Hill Book Company (1984)
- [17] Salton, G., Wong, A., Yang, C.S.: A vector space model for automatic indexing. *Commun. ACM* 18, 613–620 (1975)
- [18] Singhal, A., Buckley, C., Mitra, M., Mitra, A.: Pivoted document length normalization, pp. 21–29. ACM Press (1996)
- [19] Zesch, T., Müller, C., Gurevych, I.: Extracting lexical semantic knowledge from wikipedia and wiktionary. In: Proceedings of the 6th International Conference on Language Resources and Evaluation, Marrakech, Morocco, electronic proceedings (May 2008)



# Uncertain (Multi)Graphs for Personalization Services in Digital Libraries

Claudio Taranto, Nicola Di Mauro, and Floriana Esposito

Department of Computer Science, University of Bari "Aldo Moro"  
via E. Orabona, 4 - 70125, Bari, Italy  
{claudio.taranto,ndm,esposito}@di.uniba.it

**Abstract.** Digital Libraries organized collections of multimedia objects in a computer processable form. They also comprise services and infrastructures to manage, store, retrieve and share objects. Among these services, personalization services represent an active and broad area of digital library research. A popular way to realize personalization is by using information filtering techniques aiming to remove redundant or unwanted information from data. In this paper we propose to use a probabilistic framework based on uncertain graphs in order to deal with information filtering problems. Users, items and their relationships are encoded in a probabilistic graph that can be used to infer the probability of existence of a link between entities involved in the graph. The goal of the paper is to extend uncertain graphs definition to multigraphs and to study whether uncertain graphs could be used as a valuable tool for information filtering problems. The performance of the proposed probabilistic framework is reported when applied to a real-world domain.

## 1 Introduction

Over the past years the information content have undergone a profound change in terms of information representation and services for the use of the contents. In particular, the information content has become heterogeneous, representing different information sources such as texts, images, audio and videos. The large number of these multimedia objects and their inherent complexity has led to the need of specific services for their management and interrogation. Digital Libraries organized digital collections of multimedia objects available online in computer processable form [4]. These libraries also comprise services and infrastructures to manage, store, retrieve and share objects. In [12] the authors identify many core topics focusing on Digital Libraries research area. These topics refers to the creation of digital libraries, applications (e-learning, health care, mobile learning), preservation of data and information organization and research.

In this paper we have decided to address the problem of information organizing and finding, focusing on the personalization services, representing an active and broad area of Digital Library research [19,12,10,6]. Information filtering is a popular way to realize personalization, which can be classified into content-based filtering and collaborative filtering. The goal of this paper is to show how the use

of *uncertain graphs*, an increasingly important research topic [14,22,9], is useful to manage and solve some information filtering problems. In particular, we will see how relationships among users and among multimedia objects with their corresponding likelihood could be easily encoded adopting an uncertain graph. This probabilistic knowledge can then be used to infer the probability of existence of links between an user and an object involved in the graph. Predicting possible relationships between an user and a multimedia object can help to find useful information and to suggest multimedia objects that user could be interested in. The proposed probabilistic framework will be evaluated along its ability to represent multimedia objects, users and their relationships and to predict new relationships among the involved entities. The basic definition of uncertain graph will be extended to that of multigraphs in order to deal with multiple connection types between nodes. In particular, we will study the behavior of the system by varying the considered neighborhood of the nodes, by studying the inference accuracy, and by considering contextual information. Experimental results on real world data show that the proposed approach is promising.

## 2 Related Works

Digital Libraries organizing digital collections of multimedia objects, are one of many examples of information overload problem. Information filtering systems, more broadly, aim at removing redundant or unwanted information. They aim at presenting relevant information and reducing the information overload, while improving the signal-to-noise ratio at the semantic level. Personalization services for Digital Library are a key component for the fruition of the contents. Their implementation is very close to the problem of recommender systems [1] and link prediction [7], since they share the same objective to filter relevant contents for the user. A recommender system performs information filtering to bring information items such as movies, music, books, news, images, web pages, tools to a user. This information is filtered so that it is likely to interest the user. It is possible to categorize a recommender system into five groups depending on the required knowledge as follows.

**Content-Based Systems.** These systems analyse user preferences in order to create a profile. Using the user profile and a description of the multimedia objects, the system can identify one or more objects that are relevant to the user profile and therefore interesting for the user. The limitation of these systems is that they assume to have a significant number of preferences for each user in order to create the profile, a problem known as the cold-start problem [3].

**Collaborative Filtering Systems.** Collaborative filtering systems are based on collecting and analysing a large amount of information about users behaviour and preferences, and predicting what users will like based on their similarity to other users. These systems ignore the representation of multimedia objects. The suggestion of objects can be done in three ways: *user-based* where user preferences are compared with those of other most similar users;

*item-based* using objects similar to those that the user has seen, and *hybrid* combination of the two approaches. These approaches are called *memory-based*. Collaborative filtering methods centred on computing the relationships between multimedia objects or between users. This approach may be viewed as computing a measure of proximity or a similarity between user and objects. A similar problem is the *link prediction* problem that wants to infer missing links from an observed network: in a number of domains, one constructs a network of interactions based on observable data and then tries to infer additional links that, while not directly visible, are likely to exist [8,13,17].

**Demographic Systems.** These systems create a user profile based on demographic information. The suggested new multimedia objects is retrieved by considering the user demographic information and ignoring information about the description of the objects.

**Knowledge-Based Systems.** These systems use a user profile that the user has previously filled. In this profile the user explicitly indicates his preferences in order to guide the suggestions of the system.

**Hybrid Systems.** These hybrid recommendation systems combine the results of multiple recommendation systems in order to obtain a more accurate recommendation. They could be divided into *homogeneous recommendation systems* which combine the output from different versions of the same recommender system and *heterogeneous recommendation systems* which combines the output from different recommender systems.

Over the last few years uncertain graphs have become an important research topic [14,20,21]. In these graphs each edge is associated with an existence probability that quantifies the likelihood that the edge exists in the graphs. Using this representation it is possible to adopt the *possible world* semantics to model it. One of the main issues in uncertain graphs is how to compute the connectivity of the network. The network reliability problem [5] is a generalization of the pairwise reachability, in which the goal is to determine the probability that all pairs of nodes are reachable from one another. Unlike a deterministic graph in which the reachability function is a binary function indicating whether or not there is a path connecting two nodes, in the case of uncertain graphs the function assumes probabilistic values. In [14], the authors provide a list of alternative shortest path distance measures for uncertain graphs in order to discover the  $k$  closest vertices to a given one. Another work [11] try to deal with the concept of  $x - y$  distance constraint reachability problem. In particular, given two vertices  $x$  and  $y$ , they try to solve the problem of computing the probability that the distance from  $x$  to  $y$  is less than or equal to a user-defined threshold. In order to solve this problem, they proposed an exact algorithm and two reachability estimators based on probability sampling.

In this paper the idea is to use the expressive power of uncertain graphs formalism to address the information filtering problem and to allow the user to find objects of interest. The approach proposed in this paper takes advantage of the encouraging results obtained in [16], where the uncertain graph formalism

has been applied to solve the problem of collaborative filtering. In this paper we extended that framework to *uncertain multigraph*, allowing us to represent many heterogeneous connections among the involved entities. In order to test the multigraph extension we applied the system on an extension of the Movielens dataset used in [16] and its performances have been tested adopting different metrics. Furthermore, the behavior of the system has been studied by varying the neighborhood of the nodes during the creation of the uncertain graph, and by varying the inference accuracy. Finally, we will introduce contextual information in the form of probabilistic edges to study whether it contributes to the improvement in the inference step.

### 3 Uncertain Multi-graphs

Let  $G = (V, E)$ , be a graph where  $V$  is a collection of nodes and  $E \subseteq V \times V$  is the set of edges, or relationships, between the nodes.

**Definition 1 (Uncertain multi-graph).** *A uncertain multi-graph is a system  $G = (V, E, \Sigma, l_V, l_E, s, t, p_e)$ , where  $(V, E)$  is an directed graph,  $V$  is the set of nodes,  $E$  is the set of ordered pairs of nodes where  $e = (s, t)$ ,  $\Sigma$  is a set of labels,  $l_V : V \rightarrow \Sigma$  is a function assigning labels to nodes,  $l_E : E \rightarrow \Sigma$  is a function assigning labels to the edges,  $s : E \rightarrow V$  is a function indicating the source node of an edge,  $t : E \rightarrow V$  is a function indicating the target node of an edge, and  $p_e : E \rightarrow [0, 1]$  is a function assigning existence probability values to the edges.*

Each edge  $a = (u, v) \in E$  has a probability called *existence probability*  $p_e(a)$  which expresses the probability that the edge  $a$ , between  $u$  and  $v$ , can exist in the graph. A particular case of uncertain graph is the *discrete graph*<sup>1</sup>, where binary edges between nodes represent the presence or absence of a relationship between them, i.e., the existence probability value on all observed edges is 1.0. The semantic of an uncertain graph is the *possible world semantics* where we can imagine an uncertain graph  $G$  as a sampler of worlds, where each world is an instance of  $G$ . An instance of  $G$  is a discrete graph  $G'$  obtained by sampling from an uncertain graph  $G$  according to the probability distribution  $P_e$ , denoted as  $G' \sqsubseteq G$ , when each edge  $a \in E$  is selected to be an edge of  $G'$  with probability  $p_e(a)$ . We can consider edges labeled with probabilities as mutually independent random variables indicating whether or not the corresponding edge belongs to a discrete graph.

Assuming independence among edges, the probability distribution over discrete graphs  $G' = (V, E') \sqsubseteq G = (V, E)$  is given by

$$P(G'|G) = \prod_{a \in E'} p_e(a) \prod_{a \in E \setminus E'} (1 - p_e(a)). \quad (1)$$

**Definition 2 (Simple path).** *Given an uncertain graph  $G$ , a simple path of a length  $k$  from  $u$  to  $v$  in  $G$  is an acyclic path denoted as a sequence of edges*

<sup>1</sup> Sometimes called *certain graph*.

$p_{u,v} = \langle e_1, e_2, \dots, e_k \rangle$ , such that  $e_1 = (u, v_1)$ ,  $e_k = (v_{k_1}, v)$ , and  $e_i = (v_{i-1}, v_i)$  for  $1 < i < k$ .

Given an uncertain graph  $G$ , and  $p_{u,v}$  a path in  $G$  from node  $u$  to node  $v$ ,  $\ell(p_{u,v}) = \ell(e_1)\ell(e_2)\cdots\ell(e_k)$  denotes the concatenation of labels of all the edges in  $p_{u,v}$ .

We adopt a *regular expression*  $R$  to denote what is the exact sequence of labels that the path must contain. In this way we are not interested in all the paths in the uncertain graph of length  $k$  but only in those who have exactly the labels expressed by the regular expression. Now we can define a language-constrained simple path.

**Definition 3 (Language-constrained simple path).** *Given an uncertain graph  $G$  and a regular expression  $R$ , a language constrained simple path is a simple path  $p$  such that  $\ell(p) \in L(R)$ .*

### 3.1 Querying Uncertain Graphs

The concept of existence probability of an edge in an uncertain graph can be extended to paths. We want to calculate the probability that there exists a simple path between two nodes  $u$  and  $v$ , that is, querying for the probability that a randomly sampled discrete graph contains a simple path between  $u$  and  $v$ . More formally, the *existence probability*  $P_e(q|G)$  of a simple path  $q$  in a probabilistic graph  $G$  corresponds to the marginal  $P((q, G')|G)$  with respect to  $q$ :

$$P_e(q|G) = \sum_{G' \sqsubseteq G} P(q|G') \cdot P(G'|G) \quad (2)$$

where  $P(q|G') = 1$  if there exists the simple path  $q$  in  $G'$ , and  $P(q|G') = 0$  otherwise. Hence, the existence probability of the simple path  $q$  is the probability that the simple path  $q$  exists in a randomly sampled discrete graph.

**Definition 4 (Language-constrained simple path probability).** *Given an uncertain graph  $G$  and a regular expression  $R$ , the language-constrained simple path probability of  $L(R)$  is*

$$P_e(q|L(R), G) = \sum_{G' \sqsubseteq G} P(q|G', L(R)) \cdot P(G'|G) \quad (3)$$

where  $P(q|G', L(R)) = 1$  if there exists a simple path  $q$  in  $G'$  such that  $\ell(q) \in L(R)$ , and  $P(q|G', L(R)) = 0$  otherwise.

The existence probability computation adopting (2) or (3) is intensive and intractable for large graphs since the number of discrete graphs to be checked is exponential in the number of probabilistic edges. In order to overcome this problem the solution is to approximate it using a Monte Carlo sampling

approach [11] in which we do not generate all the possible certain graphs but only a random subset providing the following basic sampling estimator for  $P_e(q|G)$ :

$$P_e(q|G) \approx \widehat{P_e(q|G)} = \frac{\sum_{i=1}^n P(q|G'_i)}{n} \quad (4)$$

We proposed, as reported in [16], an iterative depth first search procedure to check the path existence. When a node is just visited, we will sample all its adjacent edges and pushing them into the stack used by the iterative procedure. We will stop the procedure either when the target node is reached or when the stack is empty which means that there isn't a path between the two nodes. In this way we can avoid to sample all edges to check whether the graph contains the path.

## 4 Uncertain Graphs for Digital Library

The task of Information Filtering in DL aims to suggest a new item for a user. In this way an user can find interesting content even if the size of the DL are prohibitive or there are no effective methods to search a particular item. A classical approach is to exploit the information deriving from the adoption of a neighbourhood model. As we have shown in the related works section the two widely used methods are the user-oriented and the item-based approaches. The former estimates unknown ratings exploiting past ratings of similar users, while the latter estimates a rating using known ratings made by the same user on similar items. Let  $U$  be a set of  $n$  users and  $I$  a set of  $m$  items. A rating  $r_{ui}$  indicates the preference by user  $u$  of item  $i$ , where high values mean stronger preference. Let  $S_u$  be the set of items rated from user  $u$ . A user-based approach predicts an unobserved rating  $\widehat{r}_{ui}$  as follows:

$$\widehat{r}_{ui} = \overline{r}_u + \frac{\sum_{v \in U | i \in S_u} \sigma_u(u, v) \cdot (r_{vi} - \overline{r}_v)}{\sum_{v \in U | i \in S_u} |\sigma_u(u, v)|} \quad (5)$$

where  $\overline{r}_u$  represents the mean rating of user  $u$ , and  $\sigma_u(u, v)$  stands for the similarity between users  $u$  and  $v$ , computed, for instance, using the Pearson correlation:

$$\sigma_u(u, v) = \frac{\sum_{a \in S_u \cap S_v} (r_{ua} - \overline{r}_u) \cdot (r_{va} - \overline{r}_v)}{\sqrt{\sum_{a \in S_u \cap S_v} (r_{ua} - \overline{r}_u)^2 \sum_{a \in S_u \cap S_v} (r_{va} - \overline{r}_v)^2}} \quad (6)$$

On the other side, item-based approaches predict the rating of a given item using the following formula:

$$\widehat{r}_{ui} = \frac{\sum_{j \in S_u | j \neq i} \sigma_i(i, j) \cdot r_{uj}}{\sum_{j \in S_u | j \neq i} |\sigma_i(i, j)|} \quad (7)$$

where  $\sigma_i(i, j)$  is the similarity between the item  $i$  and  $j$ .

The idea behind the neighbourhood model is to consider each object as a point of a network structure and to adopt a similarity function in order to connect the objects similar to each other. The limit of this approach is to consider only the direct connections among the entities involved in the domain, and ignoring all the information available from indirect connections and from the contextual information [18,15]. As already presented in [16], the proposed approach is used to represent a dataset consisting of user ratings,  $\mathcal{K} = \{(u, i, r_{ui}) | r_{ui} \text{ is known}\}$  with an uncertain graph and then performing inference on this graph to solve classical collaborative filtering tasks. In particular, in this paper we extended the uncertain graphs definition to that of multigraphs in order to be able to manage multiple connections among nodes.

#### 4.1 Uncertain Graph Construction

In order to construct an uncertain graph from raw data, we start by analyzing the set of ratings  $\mathcal{K} = \{(u, i, r_{ui}) | r_{ui} \text{ is known}\}$ . For each user in  $\mathcal{K}$  we add a node with label *user* and for each item in  $\mathcal{K}$  a node with label *item*. As in the approach based on the neighbourhood model we add the connections among nodes. We add two kind of connections: *simU* and *simI*. For the *simU* connections, for each user  $u$  we added an edge between  $u$  and the  $k$  most similar users to  $u$ . The probability of the edge *simU* connecting two users  $u$  and  $v$  is computed as:

$$P(\mathbf{simU}(u, v)) = \sigma_u(u, v) \cdot w_u(u, v) \quad (8)$$

where  $\sigma_u(u, v)$  is the Pearson correlation between the vectors of ratings corresponding to the set of items rated by both user  $u$  and user  $v$ , and  $w_u(u, v) = \frac{|S_u \cap S_v|}{|S_u \cup S_v|}$ , where  $S_u$  is the set of items rated from user  $u$ . For the *simI* connections, for each item  $i$  we added an edge between  $i$  and the most  $k$  similar items to  $i$ . The probability of the edge *simI* connecting the item  $i$  to the item  $j$  has been computed as:

$$P(\mathbf{simI}(i, j)) = \sigma_i(i, j) \cdot w_i(i, j), \quad (9)$$

where  $s_{ij}$  is the Pearson correlation between the vectors corresponding to the histogram of the set of ratings for the item  $i$  and the item  $j$ , and  $w_i(i, j) = \frac{|\bar{S}_i \cap \bar{S}_j|}{|\bar{S}_i \cup \bar{S}_j|}$ , where  $\bar{S}_i$  and  $\bar{S}_j$  are the set of users rating the item  $i$  and  $j$ .

In this paper we adopt a multigraph, hence we can describe multiple connections between two nodes. Supposing that users and items are described using a set of features, for each item  $i$ , we can add an edge with label *simIf* with respect to the feature  $f$ , between  $i$  and the most  $k$  similar items to  $i$ . In particular, the probability of the edge *simf* connecting the item  $i$  to the item  $j$  could be computed as:

$$P(\mathbf{simf}(i, j)) = \frac{|i_f \cap j_f|}{|i_f| + |j_f| + 1}, \quad (10)$$

where  $i_f$  is the value of the feature  $f$  for the item  $i$ . For instance, if a film is described using its genres and actors, the previous formula may be used to compute a similarity between films based on actors and genres. With a similar argument we can add an edge between the user  $u$  and the most  $k$  similar user to  $u$  with respect to a given feature.

The edges labelled  $\mathbf{r}_k$  have probability equal to 1.0 denoting a specific vote of a user relative to an object. Now that we have an uncertain graph we can predict an unknown rating  $\widehat{r}_{ui}$  solving the following maximization problem:

$$\widehat{r}_{ui} = \arg \max_j P(\mathbf{r}_j(u, i)|G), \quad (11)$$

where  $\mathbf{r}_j(u, i)$  is the unknown link with label  $\mathbf{r}_j$  between the user  $u$  and the item  $i$ . Adopting this approach we can simulate user-based collaborative filtering by querying the probability of the paths, starting from a user node and ending to an item node, belonging to the regular expression  $L_i = \{\mathbf{simU}^1\mathbf{r}_i^1\}$ . In particular, predicting the probability of the rating  $j$  as  $P(\mathbf{r}_j(u, i)$  in (11) corresponds to compute the probability  $P(q|G)$  for a query path in  $L_i$ , i.e., computing  $P(L_i|G)$  as in (3):

$$\widehat{r}_{ui} = \arg \max_j P(\mathbf{r}_j(u, i)|G) \approx \arg \max_j P(L_j|G). \quad (12)$$

We can simulate item-based collaborative filtering in the same way by computing the probability of the paths belonging to the regular expression  $L_i = \{\mathbf{r}_i^1\mathbf{simI}^1\}$ . Adopting a regular expression based approach we can construct any type of query: simple, complex, hybrid (combining a user-based and an item-based approach) and exploiting contextual information.

## 5 Experiments

In order to validate the proposed approach the HetRec2011<sup>2</sup> dataset has been used. This dataset is an extension of the MovieLens dataset and contains user ratings expressing preferences for different movies. The dataset contains 2113 users, 10197 films and 855598 ratings. The ratings are one of 10 distinct values ranging from 0.5 to 5.0 with increments of 0.5. The meta-data available include user-movie tag information, movie genres, movie directors, country assignments, and aggregate statistics of audience and critics ratings. The dataset has been divided in training and testing data. The testing part includes the last four ratings for each user, while the training part includes all the previous ones. Then, the validation procedure has been conducted following the steps: a) creating the uncertain graph from the training data as reported in Section 4; b) defining a regular expression corresponding to a specific information filtering task; and c) testing the ratings reported in the testing dataset  $\mathcal{T}$  by computing, for each pair  $(u, i) \in \mathcal{T}$  the predicted rating as in Equation (12) and comparing the prediction with the true rating as reported in  $\mathcal{T}$ . In this particular dataset we have a uncertain graph with nodes labeled as `user` or `film`. There are edges

<sup>2</sup> <http://ir.ii.uam.es/hetrec2011/datasets.html>



between two **film** nodes labeled as **simF** or **simFA**, and edges with label **simU** or **simUG** between two **user** nodes. These edges are added using the procedure presented in the previous section. In particular, **simF** denotes the probability that two films could be similar and it has been computed using (9), while **simU** indicates the probability that two users are similar computed with (8). **simFA** edges connecting two films whose probability has been computed using (10), in particular **simFA** has been computed using the actors of the films. **simUG** connects two users with a probability corresponding to the similarity computed using the histogram of the rated films' genres. For each rating  $(u, i, r_{ui} = k)$  belonging to the training set there is an edge between the user  $u$  and the film  $i$  whose label is  $r_k$ . The goal is to predict the correct rating for each instance belonging to the testing set  $\mathcal{T}$ . The predicted rating has been computed using a Monte Carlo approach by sampling certain graphs and adopting the function in (12).

The accuracy of the proposed framework has been evaluated according to the *mean absolute error* (MAE) and to the *root mean squared error* (RMSE), that are the two most commonly applied evaluation metrics for rating predictions. Given  $N$  computed rating predictions the functions are computed as:

$$MAE = \frac{1}{N} \sum_{i=1}^N |\widehat{r}_{ui} - r_{ui}| \quad (13)$$

and

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^N (\widehat{r}_{ui} - r_{ui})^2} \quad (14)$$

In order to evaluate the framework we proposed to query the paths belonging to the regular expressions reported in Table 1. The first language constrained simple paths  $L_1$  corresponds to solve a user-based information filtering problem, while the third language  $L_3$  gives us the possibility to simulate an item-based information filtering approach. As we can see from Table 2 results improve when we go from a user-based approach to a item-based in terms of MAE. We can see also that adopting languages  $L_2$  and  $L_4$ , that consider contextual edges amongs users or items, we have improving results. In the second experiment, we proposed to extend the basic languages  $L_3$  and  $L_4$  in order to consider a neighbourhood with many nested levels. In particular, instead of considering the direct neighbours only, we inspect the uncertain graph following a path with a maximum length of two edges ( $L_5, L_6$ ) and three edges ( $L_7$ ). As we can see in Table 3 languages  $L_5, L_6$  and  $L_7$ , where we extend the neighborhood of the explored graph, when compared with languages  $L_3$  and  $L_4$  achieved better results. Furthermore, languages  $L_8, L_9$  and  $L_{10}$  corresponds to a hybrid system combining both user-based and item-based approach, whose corresponding results are shown in Table 4. In each table, reporting the MAE results, the first column reports the neighbourhood of the most similar nodes introduced in the graph for each similarity function, and the second column reports the number of sampling adopting for each languages.

**Table 1.** Language constrained simple paths used for the HetRec2011 dataset
$$\begin{aligned}
L_1 &= \{\text{simU}^1 \mathbf{r}_k^1\} \\
L_2 &= \{\text{simU}^1 \mathbf{r}_k^1\} \cup \{\text{simUG}^1 \mathbf{r}_k^1\} \\
L_3 &= \{\mathbf{r}_k^1 \text{simF}^1\} \\
L_4 &= \{\mathbf{r}_k^1 \text{simF}^1\} \cup \{\mathbf{r}_k^1 \text{simFA}^1\} \\
L_5 &= \{\mathbf{r}_k^1 \text{simF}^n : 1 \leq n \leq 2\} \\
L_6 &= \{\mathbf{r}_k^1 \text{simF}^n : 1 \leq n \leq 2\} \cup \{\mathbf{r}_k^1 \text{simFA}^n : 1 \leq n \leq 2\} \\
L_7 &= \{\mathbf{r}_k^1 \text{simF}^n : 1 \leq n \leq 3\} \cup \{\mathbf{r}_k^1 \text{simFA}^n : 1 \leq n \leq 3\} \\
L_8 &= \{\text{simU}^1 \mathbf{r}_k^1\} \cup \{\mathbf{r}_k^1 \text{simF}^1\} \\
L_9 &= \{\text{simU}^n \mathbf{r}_k^1 : 1 \leq n \leq 2\} \cup \{\mathbf{r}_k^1 \text{simF}^n : 1 \leq n \leq 2\} \\
L_{10} &= \{\text{simU}^n \mathbf{r}_k^1 : 1 \leq n \leq 3\} \cup \{\mathbf{r}_k^1 \text{simF}^n : 1 \leq n \leq 3\}
\end{aligned}$$
**Table 2.** MAE with the languages  $L_1$ ,  $L_2$  and  $L_3$ 

Neighborhood	Sampling	$L_1$	$L_3$	$L_2$	$L_4$
<b>5</b>	100	1.0070	0.9878	0.7493	0.7316
<b>5</b>	500	1.0040	0.9840	0.7314	0.7300
<b>10</b>	100	0.9740	0.9661	0.6850	0.6788
<b>10</b>	500	0.9687	0.9631	0.6745	0.6720
<b>15</b>	100	0.9446	0.9404	0.6545	0.6521
<b>15</b>	500	0.9395	0.9380	0.6526	0.6488
<b>20</b>	100	0.9383	0.9308	0.6415	0.6409
<b>20</b>	500	0.9297	0.9263	0.6390	0.6339

**Table 3.** MAE with the languages  $L_2, L_4, L_5, L_6$  and  $L_7$ 

Neighborhood	Sampling	$L_2$	$L_4$	$L_5$	$L_6$	$L_7$
<b>5</b>	100	0.7493	0.7316	0.6940	0.6911	0.6761
<b>5</b>	500	0.7314	0.7300	0.6812	0.6809	0.6633
<b>10</b>	100	0.6850	0.6788	0.6503	0.6404	0.6311
<b>10</b>	500	0.6745	0.6720	0.6309	0.6282	0.6225
<b>15</b>	100	0.6545	0.6521	0.6305	0.6227	0.6207
<b>15</b>	500	0.6526	0.6488	0.6176	0.6168	0.6140
<b>20</b>	100	0.6415	0.6409	0.6217	0.6196	0.6173
<b>20</b>	500	0.6390	0.6339	0.6162	0.6150	0.6087

**Table 4.** MAE with the languages  $L_8, L_9$  and  $L_{10}$ 

Neighborhood	Sampling	$L_8$	$L_9$	$L_{10}$
<b>5</b>	100	0.7187	0.6781	0.6629
<b>5</b>	500	0.7100	0.6706	0.6564
<b>10</b>	100	0.6662	0.6386	0.6211
<b>10</b>	500	0.6609	0.6255	0.6111
<b>15</b>	100	0.6361	0.6201	0.6196
<b>15</b>	500	0.6322	0.6102	0.6072
<b>20</b>	100	0.6255	0.6179	0.6160
<b>20</b>	500	0.6237	0.6050	0.5912

Table 5 shows the results on HetRec2011 dataset, using a 10-fold cross-validation, comparing the proposed framework with respect to neighborhood-based recommendation methods reported in [2]. The approach proposed in [2] exploit also the tags assigned by the users in order to extract latent semantics by using Latent Semantic Analysis. The first recommender was based on collaborative filtering using the cosine similarity to build user neighbourhoods, the second uses content analysis on latent topic analysis, while the third was based on a simple average rating. As we can see in Table 5, even without using tag information, the obtained results adopting our system are better than, or comparable to, those obtained with the approaches exploited in [2].

**Table 5.** RMSE error on HetRec2011 adopting 10-fold cross-validation

Method	RMSE
Average Recommender Rating [2]	1.0880
Content Analysis [2]	0.9436
Collaborative Filtering [2]	0.8876
$L_7$	0.9071
$L_9$	0.9005
$L_{10}$	0.8891

## 6 Conclusions

In this paper a framework based on uncertain (multi)graphs able to deal with information filtering problems in DL has been presented. The evaluation of the proposed approach has been reported by applying it to a real world dataset and proving its validity in solving simple and complex information filtering tasks when compared with respect to other competing systems. In particular, we studied the behavior of the system by varying the neighborhood considered for each node, by varying the inference accuracy, and by considering contextual information. We have noticed that the contextual information provides a very strong improvement, especially for those regular expressions that make use of short paths and that consider the similarity of users and objects as something detached from the context.

## References

1. Adomavicius, G., Tuzhilin, A.: Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *IEEE Transactions on Knowledge and Data Engineering* 17(6), 734–749 (2005)
2. Bothos, E., Christidis, K., Apostolou, D., Mentzas, G.: Information market based recommender systems fusion. In: *Proceedings of the 2nd International Workshop on Information Heterogeneity and Fusion in Recommender Systems*, pp. 1–8. ACM (2011)

3. Burke, R.: Hybrid Web Recommender Systems. In: Brusilovsky, P., Kobsa, A., Nejdl, W. (eds.) *Adaptive Web 2007*. LNCS, vol. 4321, pp. 377–408. Springer, Heidelberg (2007)
4. Candela, L., Castelli, D., Ferro, N., Ioannidis, Y., Koutrika, G., Meghini, C., Pagano, P., Ross, S., Soergel, D., Agosti, M., Dobрева, M., Katifori, V., Schuldtt, H.: *The DELOS Digital Library Reference Model*. Foundations for Digital Libraries (2007)
5. Colbourn, C.J.: *The Combinatorics of Network Reliability*. Oxford University Press (1987)
6. Gao, F., Xing, C., Du, X., Wang, S.: Personalized service system based on hybrid filtering for digital library. *Tsinghua Science & Technology* 12(1), 1–8 (2007)
7. Getoor, L., Diehl, C.P.: Link mining: a survey. *SIGKDD Explorations* 7(2), 3–12 (2005)
8. Goldberg, D.S., Roth, F.P.: Assessing experimentally derived interactions in a small world. *Proceedings of the National Academy of Sciences* 100(8), 4372–4376 (2003)
9. Hintsanen, P., Toivonen, H.: Finding reliable subgraphs from large probabilistic graphs. *Data Min. Knowl. Discov.* 17(1), 3–23 (2008)
10. Itmazi, J.A., Megías, M.G.: Using recommendation systems in course management systems to recommend learning objects. *Int. Arab J. Inf. Technol.* 5(3), 234–240 (2008)
11. Jin, R., Liu, L., Ding, B., Wang, H.: Distance-constraint reachability computation in uncertain graphs. *Proc. VLDB Endow.* 4, 551–562 (2011)
12. Nguyen, S.H., Chowdhury, G.: Digital library research (1990-2010): A knowledge map of core topics and subtopics. In: *ICADL*, pp. 367–371 (2011)
13. Popescul, A., Ungar, L.H.: Statistical relational learning for link prediction. In: *IJCAI 2003 Workshop on Learning Statistical Models from Relational Data* (2003)
14. Potamias, M., Bonchi, F., Gionis, A., Kollios, G.: k-nearest neighbors in uncertain graphs. *Proc. VLDB Endow.* 3, 997–1008 (2010)
15. Taranto, C., Di Mauro, N., Esposito, F.: Probabilistic Inference over Image Networks. In: Agosti, M., Esposito, F., Meghini, C., Orio, N. (eds.) *IRCDL 2011*. CCIS, vol. 249, pp. 1–13. Springer, Heidelberg (2011)
16. Taranto, C., Di Mauro, N., Esposito, F.: Uncertain graphs meet collaborative filtering. In: *3rd Italian Information Retrieval Workshop* (2012)
17. Taskar, B., Wong, M.F., Abbeel, P., Koller, D.: Link Prediction in Relational Data. In: *Neural Information Processing Systems* (2003)
18. Witsenburg, T., Blockeel, H.: Improving the Accuracy of Similarity Measures by Using Link Information. In: Kryszkiewicz, M., Rybinski, H., Skowron, A., Raś, Z.W. (eds.) *ISMIS 2011*. LNCS, vol. 6804, pp. 501–512. Springer, Heidelberg (2011)
19. Zhen-ming, Y., Tianhao, Y., Jia, Z.: A social tagging based collaborative filtering recommendation algorithm for digital library. In: *ICADL*, pp. 192–201 (2011)
20. Zou, Z., Gao, H., Li, J.: Discovering frequent subgraphs over uncertain graph databases under probabilistic semantics. In: *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 633–642. ACM (2010)
21. Zou, Z., Li, J., Gao, H., Zhang, S.: Finding top-k maximal cliques in an uncertain graph. In: *International Conference on Data Engineering*, pp. 649–652 (2010)
22. Zou, Z., Li, J., Gao, H., Zhang, S.: Mining frequent subgraph patterns from uncertain graph data. *IEEE Transactions on Knowledge and Data Engineering* 22, 1203–1218 (2010)

# Improving Online Access to Archival Data

Vittore Casarosa<sup>1</sup>, Carlo Meghini<sup>1</sup>, and Stanislava Gardasevic<sup>2</sup>

<sup>1</sup> ISTI-CNR, Pisa, Italy

<sup>2</sup> DILL International Master, University of Parma, Italy

**Abstract.** Archives are memory institutions whose original mission was to preserve and provide access to a set of carefully selected, arranged and described documents to a small number of scholars interested in their contents. For those specialists, the usual way to find information in an archive is by way of “finding aids”, i.e. descriptions of the archive contents that reflect the hierarchical structure by which data are physically arranged in an archive. With the increased availability of archival holdings accessible on the Web, archives are now widening the range of users, and the use of online finding aids has proved to be too complicated for the non-specialists. This is mostly due to the hierarchical nature of the description, usually represented on line with a standard called EAD (Encoded Archival Description). This paper is the synopsis of a Master Thesis, where a methodology has been developed to represent the information contained in finding aids with a different standard, namely EDM (Europeana Data Model), which is used by the Europeana digital library and is becoming the de-facto standard for metadata interoperability. EDM allows a much more intuitive representation of the archive content and the possibility to access data from many different access points.

**Keywords:** Archive, EAD, finding aid, EDM, Europeana Data Model.

## 1 The Structure of Archives

### 1.1 The Archival Fond

Archives differ from other memory institutions in the nature of materials they have. Contrary to libraries, where usually the material collected are just “copies” of books and journals, the material in archives and manuscript libraries are the unique records of corporate bodies and the papers of individuals and families. Therefore archival descriptions have to reflect this peculiarities, retaining all the informative power of a record, and keeping trace of the provenance and original order in which resources have been collected and filed by archival institutions.

This approach emphasize the central concept of archival science, which is “fond”, i.e. “all of the documents naturally generated and/or accumulated and/or used by a particular person, family or corporate body in the conduct of personal or corporate activity”. This definition leads to the fundamental archival principle (respect des fonds), which is dictating that resources of different origins are to be kept separate, in order to preserve the context in which they were found and the context in which they

were created. Furthermore, documents or records kept in archive are usually related to other documents, and are grouped into identifiable subgroups. This kind of record keeping and describing fosters the use of a hierarchical model. The hierarchical structure of the archive expresses the relationships and dependency links between the records of the archive. Therefore, a fond is usually organized in sub-fonds, which in turn can be organized in series and sub-series, formed by archival units. Following this structure, archival descriptions also proceed from general to specific, and for every unit of description they show its relationships and links with other units and with the general fonds. Archival descriptions can be presented as a tree, as shown in Figure. 1.

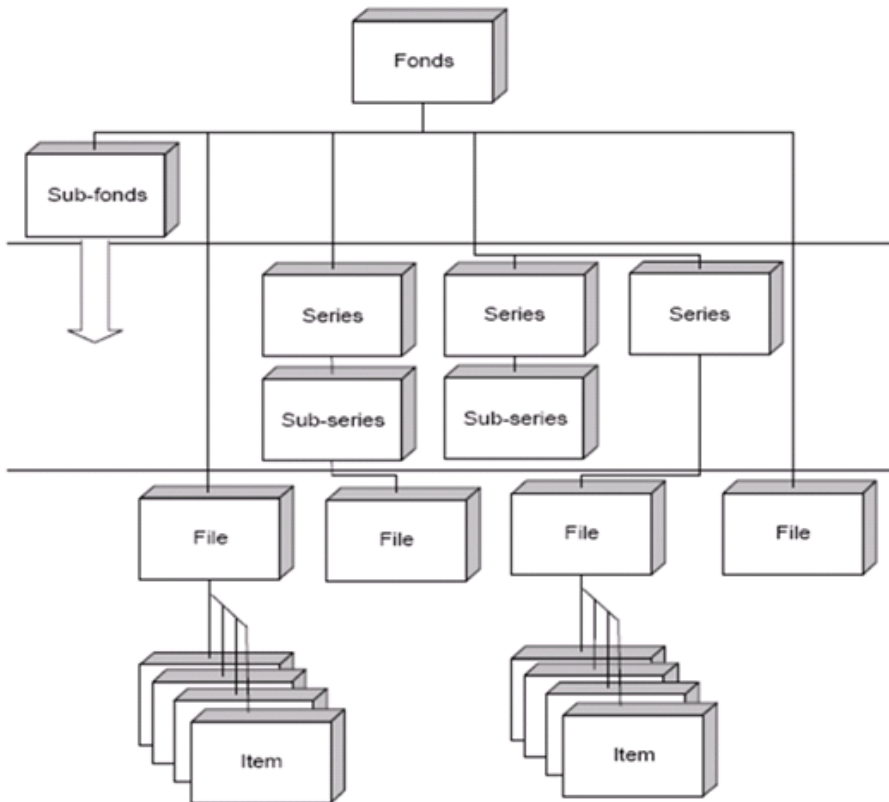


Fig. 1.

## 1.2 The Finding Aids

The gate to the archival holdings are finding aids, based on archival description practice. Finding aid is a term ordinarily used only in archives, and in general it can

include also card indexes for manuscript collections, administrative histories, and inventories for archives. Finding aids are used to access archival materials, and they contain far more information about a collection than can be found in a summary catalog record. Finding aids are generally created in the course of processing a collection and usually reflect the hierarchical arrangement of the materials. Often, many finding aids start by describing a large group of materials, usually the entire collection or record group, and then move to the description of the series of the first level components, followed by the description of smaller and smaller components, such as subseries, files and possibly even items. The description of lower levels inherits the description of the preceding levels. At the same time, finding aid acts as a collection management tool for archivist and access point for the researchers

### 1.3 The EAD Standard

EAD, Encoded Archival Description, a standard for representing finding aids, was started in the early nineties. The design of EAD was based on the following criteria: “1) ability to present extensive and interrelated descriptive information found in archival finding aids, 2) ability to preserve the hierarchical relationships existing between levels of description, 3) ability to represent descriptive information that is inherited by one hierarchical level from another, 4) ability to move within a hierarchical informational structure, and 5) support for element-specific indexing and retrieval”. Based on these requirements, XML was chosen as the formal syntax to represent the finding aids, so that an EAD encoded finding aid becomes an XML document written according to the specifications of the EAD XML Schema. Today, after several revisions, EAD is a really global standard being used by a wide variety of institutions throughout the world.

At the same time, EAD has also become the target of several critiques from archival theorists, many of them addressing its usability. The main problems that have been reported when using online finding aids encoded in EAD can be summarized as follows.

- the lack of alternative access points for users, because of the arrangement of materials according to provenance or original order of records;
- the complicated terminology; archivist should map technical terminology used as subject access points and for labeling data elements to a less technical vocabulary in order to facilitate resource discovery by non-expert users;
- finding aids consist of extensive contextual description of the circumstances surrounding the creation of its materials, and de-contextualized access to archival materials is very difficult;
- the length of the files and navigational complexity makes the process of discovery very hard;
- the administrative information that is woven throughout the finding aids is confusing;

- the collective and hierarchical description of the material and the lack of item-level description prevents an easy access at item level and a quick finding of a known item;
- the traditional finding-aid is designed to be used in an environment where the archivist acts as a mediator between the user and the finding aids, which is almost impossible over the Internet.

## 2 EDM, the Europeana Data Model

The Europeana Data Model (EDM) is a model for structuring the data that the Europeana Digital Library will be ingesting, managing and publishing. Europeana is a major effort of the European Union to create a digital library containing the cultural heritage of Europe. Today it already contains about 18 millions items, provided by a number of memory institutions all over the world. EDM was defined not only to support the richness of the content providers' metadata but also to enable data enrichment from a range of third party sources and to facilitate the publishing of (some of) Europeana content in the Linked Open Data cloud. The main requirements considered for the design of EDM were:

- distinction between “provided object” (painting, book, movie, archaeology site, archival file, etc.) and the digital representation(s) of the object
- distinction between the object and the metadata record describing the object
- multiple records for the same object should be allowed, containing potentially contradictory statements about an object
- support for objects that are composed of other objects
- compatibility with different abstraction levels of description
- provide a standard metadata format that can be specialized
- provide a standard vocabulary format that can be specialized
- allow data integration in an open environment, where it is impossible to anticipate all the data that will be contributed
- allow for rich functionality, possibly via extensions
- re-use existing (standard) models as much as possible

These design criteria have been the basis for the choice of the Semantic Web principles for EDM, providing a model which can be seen as an anchor to which various finer-grained models can be attached, making them (at least partly) interoperable at the semantic level, while retaining original expressivity and richness of original data.

The low level syntax for representing resources and their properties is RDF (Resource Description Framework), usually represented as graphs for “human consumption” or as XML documents for “computer consumption”; the high level syntax is OAI-ORE (Object Re-use and Exchange), which easily supports the ideas of the



Linked Data approach, emphasizing the re-use and linkage of richly described resources over the web. Fundamental for EDM is the OAI-ORE notion of “aggregation”, which allows to link together an object and its digital representation(s), and the notion of “proxy” which allows to represent different views on the same resource. In Figure 2 we illustrate these ideas using as an example the painting of Mona Lisa.



Fig. 2.

The top element is an OAI-ORE aggregation, identified by the URI `ex1:aggregation000PE025604`, which links together (`ore:aggregates`) the resource Monna Lisa, identified by the URI `ex1:object000PE025604`, provided (`dc:creator`) by the Direction des Musées de France and and two digital representations (`ens:WebResource`) of this resource. Additional information about the resource (`dc:creator`, `dcterms:title`, i.e. metadata records) are provided through the proxy `ex1:proxy000PE025604`. This allows to attach to the same resource another proxy (possibly coming from another provider) containing additional information for that same resource, and maintaining a clear distinction about the provenance of the two different sets of information.

As can be seen from the example above, in addition to defining terms in its own name space (abbreviated in `ens:`), EDM (re) uses as much as possible existing name spaces (i.e. their semantic), such as those defined for RDF, RDFS, SKOS, OAI-ORE, Dublin Core.

### 3 Mapping EAD to EDM

The EAD data has a hierarchical structure with descriptions associated with the nodes of the hierarchy. For this reason it is convenient to divide the general problem of mapping EAD into EDM into two parts: the structural mapping, i.e. the

transformation of an EAD hierarchy into an equivalent EDM aggregation; and the metadata mapping, that is the transformation of the descriptions found in the EAD nodes into an equivalent EDM metadata record. It is important to remember that in EAD the description associated with a node inherits all the descriptions of its ancestors.

### 3.1 Transforming an EAD Hierarchy into an EDM Aggregation

The steps for the first part of the transformation are as follows:

1. transform each EAD tree node C into an EDM Aggregation A
2. associate an OAI-ORE Proxy P to the Aggregation A, by means of the OAI-ORE property ore:proxyIn;
3. use the Proxy P as a representative of the real-world entity that node C is about, i.e. the content described by node C;
4. use the Dublin Core property dc:hasPart to relate the proxy P with the proxies defined for the children of node C in the EAD tree. In this way, the EAD tree is represented by the tree induced by the dc:hasPart property;
5. retain the order of the sibling nodes of C by means of the property ens:isNextInSequence.

The first steps of the transformation is shown in Figure 3, where the proxies have been omitted.

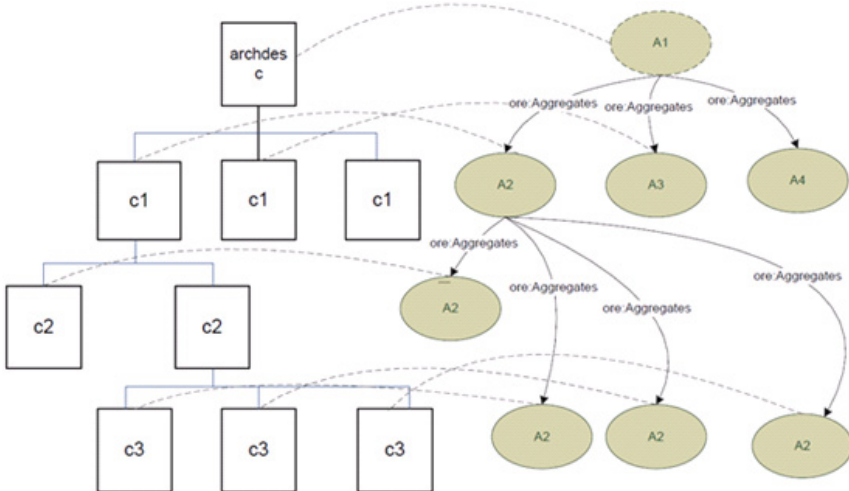


Fig. 3.

### 3.2 Mapping EAD Values into EDM Values

In the second part of the mapping, EAD elements and their possible attributes are mapped to corresponding EDM properties. To find in EDM a property equivalent (or as close as possible) to a source element, the EDM element specification should be consulted in order to see the definitions, constraints and examples of usage for all EDM classes and properties. When mapping to EDM properties, one should choose those properties carrying as much as possible semantic similarity to the elements or attributes of the original EAD schema, in order to retain as much original information as possible. EDM offers a range of properties, which are mostly defined in Dublin Core and Europeana namespaces, and to which more specialized ones can be attached and declared as subproperties.

The core idea behind converting EAD data into EDM is that every complex element, i.e. an element carrying all the information related to its ancestors (the EAD hierarchy) maps to a resource, i.e. a node of the EDM aggregation (more precisely, it maps to the proxy of the EDM node representing the corresponding node in the EAD hierarchy), and every atomic attribute maps to an attribute of this node. It should be remembered that, based on the EDM model, the metadata values are attached to the proxy of a resource, and not to the resource itself.

### 3.3 Validation of the Process

The process described above was validated by applying it to archival data coming from the Multimedia Archive of Accademia Nazionale di Santa Cecilia (ANSC). ANSC is a musical academy located in Rome, Italy and one of the oldest musical institutions in the world. The entire patrimony of this institution is about 120,000 volumes and publications, mainly scores, monographs and periodicals about music. Two fonds of this archive (Ethnomusicology Fond and Audio Video Fond) were mapped to EDM for the purpose of validating the method and analyzing the process.

The descriptions of the two fonds chosen was made available as two separate EAD XML files, and each fond was processed separately. The separation of the different levels found in the description of the fonds was performed by using ad hoc software developed at ISTI-CNR. For each extracted level a separate XML file was created, and each level was analyzed to make sure that the mapping of the nodes of a given level would cover all the possible elements at that level.

The result of this work was summarized in two metadata mapping tables, one for each fond. An excerpt of one table is shown in Figure 4. In column (a) there is the path in which the EAD elements were encountered in the original file; in column (b) there is the meaning (the semantics) of these elements; in column (c) there is the default values of these element and in column (d) there is the most appropriate EDM counterpart. Finally, column (e) contains the the RDF objects created for the composite elements.

(a)	(b)	(c)	(d)	(e)
<c>	fond , highest node		create instance of ens:ArchivalFond, domain of: ens:IsPartOf (to recordgrp Proxies)	create subclass of ens:NonInforma tionResource called ens:ArchivalFo nd
#level	Fond	fond	dc:type	
#id	Identifier		dc:identifier	
#audience	internal			/
<did>				/
<unit- title>		Archivio di Etnomusi- cologia	dc:title	dc:title
<uitid>	call num- ber/referen ce code, value not mapped		/ ens:currentLocation	create instance 1 of ens:Place
#country- code	IT			ens:country (to 1:Place)
#reposito- rycode	ANSC		dc:source (to 1:Place) instance 1 of class:Agent, sko- salttable:ANSC, this URI will hold all the data on ANSC, ad- dress..+ skos alttable ANSC (to 1:Agent)	

Fig. 4.

## 4 Conclusions

The main purpose of transforming the EAD representation into EDM is an attempt to make online access to finding aids of archives more “user-friendly” for the casual user. From the insight gained through the validation of the transformation process, despite the limited size of the archival data used, it seems that (at least to some extent) the goal has been achieved. The main improvements to on-line access for the general public of archives can be summarized as follows:

- The specialized archival terminology, through the mapping defined in the mapping tables, is translated to the more general terms used in EDM, making access more intuitive and easy; in addition, it eliminates the many inconsistencies of the terms

used in different archives (and on their web sites), which makes archival research even more confusing.

- In the hierarchical structure of finding aids discovery is usually available through a top-down approach, while using EDM as a query language any node can now be reached directly, and from there a user can go in any possible direction.
- In EAD the information is often buried so deep in the hierarchical structure of the file, that the Web crawlers have problems in indexing it; in the mapping to EDM, the information from the inner levels is extracted and is equally accessible to search engines as the one from the upper ones.
- If the mapping to EDM from different archives is done in a consistent way, it would allow to search for information over more than one archive, providing the same functionality as the union catalog for libraries.

Along the lines of the last point, we might add that the use of existing authority files for person names and for geographical names would provide a great added value to the archival data. As a general rule, genealogists and historians account for more than 50% of archive users, and they usually search for information starting with person or place name. Authority files would help overcome problems caused by different spelling for the name of a person or a location, or to account for the change of names over time.

In a broader perspective, we should consider also that once that the archival data is available in EDM representation it would be possible to overcome the “principle of provenance”, i.e. the fundamental archival principle by which records of different origins (provenance) should be kept separate in order to preserve their context. The consequence of this principle is that archival researchers often need to access several fonds in order to collect material of interest that is kept (and described) in separate fonds, but that is logically connected in some way. By applying the ideas of Linked Open Data, and creating links between archival collections and other(re)sources on the Web it would be possible for a researcher to easily discover contextually related material of different provenance, possibly getting new (and may be unexpected) perspectives on the subject of interest.

## References

1. Carpenter, B., Park, J.: Encoded Archival Description (EAD) Metadata Scheme: An Analysis of Use of the EAD-Headers. *Journal of Library Metadata* 9(1), 134 (2009)
2. Chan, L.M., Zeng, M.L.: Metadata Interoperability and Standardization—A Study of Methodology. Part I. *D-Lib Magazine* 12(6) (2006)
3. Coats, L.R.: Users of EAD - Finding Aids: Who Are They and Are They Satisfied? *Journal of Archival Organization* 2(3), 25 (2004)
4. Definition of the Europeana Data Model elements Version 5.2.1, Europeana v1.0 (2011)
5. Europeana Data Model Primer. Europeana v1.0 (2010)
6. International Council on Archives, Statement of Principles Regarding Archival Description. *Archivaria* 34 (1992)

7. Meghini, C., Isaac, A., Gradmann, S., Schreiber, G., et al.: The Europeana Data Model. In: ECDL Workshop on Very Large Digital Libraries, Glasgow, September 10 (2010)
8. Pitti, D.V.: Encoded Archival Description: An Introduction and Overview. *ESARBICA Journal* 20, 71–80 (2001)
9. Pitti, D.V., Duff, W.M.: Encoded Archival Description on the Internet. Haworth Information Press, Binghamton (2001); Also published as *Journal of Internet Cataloging* 4(3/4)
10. Theodoridou, M., Doerr, M.: Mapping the Encoded Archival Description DTD Element Set to The CIDOC-CRM. Technical Report 289, ICS-FORTH (2001)

# Quick and Easy Implementation of Approximate Similarity Search with Lucene\*

Giuseppe Amato, Paolo Bolettieri, Claudio Gennaro, and Fausto Rabitti

ISTI - CNR, Pisa, Italy

{giuseppe.amato,paolo.bolettieri,  
claudio.gennaro,fausto.rabitti}@isti.cnr.it

**Abstract.** Similarity search technique has been proved to be an effective way for retrieving multimedia content. However, as the amount of available multimedia data increases, the cost of developing from scratch a robust and scalable system with content-based image retrieval facilities is quite prohibitive.

In this paper, we propose to exploit an approach that allows us to convert low level features into a textual form. In this way, we are able to easily set up a retrieval system on top of the Lucene search engine library that combines full-text search with approximate similarity search capabilities.

## 1 Introduction

Very often multimedia content is not associated with any text or metadata, therefore traditional search techniques cannot be used and content-based retrieval or similarity-based retrieval is the only way to access this information. Moreover, even when textual information is available, the combination of similarity search with the full-text search is very useful.

However, if the digital data we want to search for similarity are just a few thousand, a sequential search could be enough. But when the amount of data becomes large (hundreds of thousands), a single similarity search can last minutes.

On the other hand, the continuous price reduction of digital production tools, such as cameras, camcorders, and smartphones, is driving the demand for content-based retrieval tools.

Several attempts are currently being made to provide these capabilities, for instance Google images allows the user to upload a photo to find out similar images in the web. However, the cost of developing and deploying from scratch a robust and reliable system with content-based image retrieval facilities could not be within the range of possibilities for everyone.

But how easy is it to add these features to an existing Digital Library Management System? In this paper, we would like to approach the problem of similarity

---

\* This work was partially supported by the ASSETS project funded by the European Commission.

search by enhancing the full-text retrieval library Lucene<sup>1</sup> with content-based image retrieval facilities. Apache Lucene is a high-performance, full-featured text search engine library written entirely in Java that is suitable for nearly any application requiring full-text search abilities.

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

The basic idea exploited in our approach has been independently introduced by Amato et al.<sup>2</sup> and Chavez et al.<sup>4</sup> and consists on observing that two objects  $x_1$  and  $x_2$  are very similar (which in metric spaces means that they are close one to each other), if their view of the surrounding world (their perspective) is similar as well. This implies that, if we take a set of objects from the database and we order them according to their similarity to  $x_1$  and  $x_2$ , the obtained orderings are also similar. Therefore, we can approximatively judge the similarity between any two arbitrary objects  $x_1$  and  $x_2$ , by comparing the ordering, according to their similarity to  $x_1$  and  $x_2$ , of a group of reference objects, instead of using the actual distance function between the two objects.

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

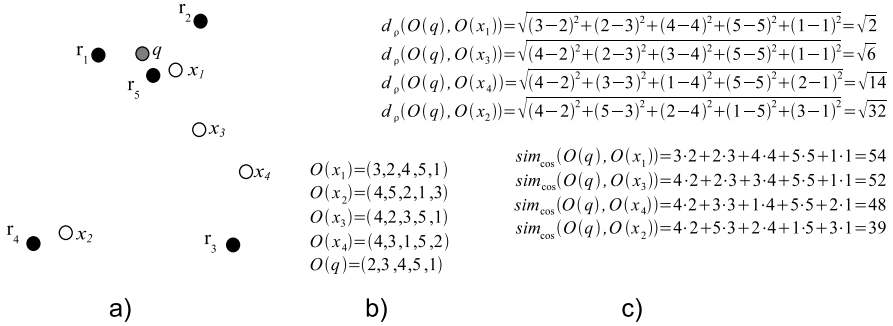
Capitalizing on the work of Amato et al.<sup>2</sup>, we also use the inverted files in our research. Another similar approach, called PP-Index<sup>5,6</sup>, uses a compact prefix tree for estimating the real distance order of the indexed objects with respect to a query. All these above mentioned approaches make use of index methods completely designed and developed from scratch. Although the results of these systems are quite impressive<sup>3</sup>, they probably will not easily move from research prototypes to commercial applications due to the strong effort required to maintain and support such information systems. Consider, for example, Lucene: at the time of this writing, Lucene’s core team includes about half a dozen active developers. In addition to the official project developers, Lucene has a fairly large and active technical user community that frequently contributes patches, bug fixes, and new features.

---

<sup>1</sup> <http://lucene.apache.org>

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





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

Moreover, only the approach in [5] provides a full-text search on descriptive textual metadata, which is, however, not combined with the content-based similarity search. Our approach instead since it is built on top of Lucene provides complex query processing by combining similarity search with the full-text search.

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

## 2 Perspective Based Space Transformation

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

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

We denote the rank in  $O(x)$  of a reference object  $r_i \in R$  as  $O_i(x)$ . For example, if  $O_4(x) = 3$ ,  $r_4$  is the 3rd nearest object to  $x$  among those in  $R$ .

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

As we anticipated before, we assume that if two objects are very close one to each other, they have a similar view of the space. This means that also the orderings of the reference objects according to their distance from the two objects should be similar. There are several standard methods for comparing two ordered lists, such as *Kendall's tau*, the *Spearman Footrule Distance*, and the *Spearman Rho Distance* [7]. In this paper, we concentrate our attention on the latter distance, which is also used in [4]. The reason of this choice (explained later on) is tied to the way standard search engines process the similarity between documents and query. Given two ordered lists  $O(x)$  and  $O(q)$  ( $x, q \in \mathcal{D}$ ), containing the ranks of all objects of  $R$ , the Spearman Rho Distance  $d_\rho$  between  $O(x)$  and  $O(q)$  is computed as in the following:

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

where  $m$  is the dimension of the vector  $R$ . This distance measures the degree in which rankings correspond with each other and it can be used in place of the metric distance  $d$  (see Figure 1c)).

In order to reduce the search cost and also, as we will see, the size of the index, it is convenient to take just the closest reference objects to represent any object that has to be indexed. Let  $k_x \leq m$  be the number of reference objects used for representing the objects. Note that, in this case, different objects will be typically represented by different reference objects, given that different objects will have different neighbor reference objects. This idea can be extended also to the query, for which we can exploit a number  $k_q \leq k_x$  of nearest reference objects. If we define two approximate version of the vectors  $\tilde{O}^k$ , such that  $\tilde{O}_i^k(x) = k + 1$  for all  $i$  such that  $O_i(x) > k$  (with either  $k = k_x$  or  $k = k_q$ ), we can still use the distance in Eq. (1), i.e:

$$d_\rho(\tilde{O}^{k_x}(x), \tilde{O}^{k_q}(q)) = \sqrt{\sum_{i=1}^m (\tilde{O}_i^{k_x}(x) - \tilde{O}_i^{k_q}(q))^2}. \quad (2)$$

In this case, we assume that  $x$  belongs to the dataset and  $q$  is the query. This is a generalization of the Spearman Rho Distance with location parameter for the special case  $l = k_x = k_q$  [7], which evaluates the distance (or dissimilarity) of two top-k ranked lists.

Up to now, we have discussed how to compare two partial rankings of reference objects corresponding to objects and query. However, we did not say how to implement the proposed idea into a standard full-text search engine.

Most text search engine, including Lucene, use the Vector Space model to represent text. In this representation, a text document is represented as a vector of terms each associated with the number of occurrences of the term in the document. Therefore, we have to define a textual representation each metric object of the database so that the inverted index produced by Lucene looks like the one presented above and that its built-in similarity function behaves like the

Spearman Similarity rank correlation used to compare ordered lists. This can be achieved in several ways, in the following we outline our solution.

First, we associate each element  $r_i \in R$  with a unique alphanumeric keyword  $\tau_i$ . Then we use the function  $t^k(x)$ , defined in the following, to obtain a space-separated concatenation of zero or more repetitions of  $\tau_i$  words:

$$t^k(x) = \bigcup_{i=1}^i \bigcup_{j=1}^{k+1-\tilde{O}_i^k(x)} \tau_j$$

where, by abuse of notation, we denote the space-separated concatenation of words with the union operator  $\bigcup$ . The function  $t^k(x)$  returns a text representation of  $x$  such that, if  $r_i$  appears in position  $p$  in the list of the  $k$  reference objects nearest to  $x$ , then the term  $\tau_i$  is repeated  $(k+1) - p$  times in the text. The function  $t^k(x)$  is used to generate the textual representation of the object  $x$  to be used for both indexing and querying purposes. Specifically, we use  $k = k_x$  for indexing and  $k = k_q$  for querying.

In our case, this means that, if for instance term  $\tau_i$  corresponding to the reference descriptor  $r_i$  ( $1 \leq i \leq m$ ) appears  $n$  times, the  $i$ -th element of the vector will contain the number  $n$ , and whenever  $\tau_i$  does not appear it will contain 0. To summarize, we finally get the vectors of size  $m$ ,  $\tilde{O}^{k_x}(x)$  and  $\tilde{O}^{k_q}(q)$ , which correspond to  $t^{k_x}(x)$  and  $t^{k_q}(q)$ , respectively. The cosine similarity is typically adopted to determine the similarity of the query vector and a vector in the database of the text search engine, and it is defined as:

$$sim_{cos}(\tilde{O}^{k_x}(x), \tilde{Q}^{k_q}(q)) = \frac{\tilde{O}^{k_x}(x) * \tilde{Q}^{k_q}(q)}{\|\tilde{O}^{k_x}(x)\| \|\tilde{O}^{k_q}(q)\|},$$

where  $*$  is the scalar product.  $sim_{cos}$  can be used as a function that evaluates the similarity of the two ranked lists in the same way as  $d_\rho(x, q)$  defined in Eq. (2) does (although it is defined as a distance), and it is possible to prove that the first one is an order reversing monotonic transformation of the second one, and then that they are equivalent for practical aspects<sup>3</sup>. This means that if we use  $d_\rho(\tilde{O}^{k_x}(x), \tilde{O}^{k_q}(q))$  and we take the first  $k$  nearest metric objects from dataset (i.e., from the shortest distance to the highest) we obtain exactly the same descriptors in the same order if we use  $sim_{cos}(\tilde{O}^{k_x}(x), \tilde{Q}^{k_q}(q))$  and take the first  $k$  similar objects (i.e., the greater values to the smaller ones). This is illustrated in Figure 11. The proof of this proposition is omitted due to space limitations of this paper but may be demonstrated using simple mathematical steps. To have an idea on how these textual representations look like, consider the example reported in Figure 11, and let us assume  $\tau_1 = \text{RO1}$ ,  $\tau_2 = \text{RO2}$ , etc. The function  $t$  will generate the following output

<sup>3</sup> To be precise, it is possible to prove that  $sim_{cos}(x, q)$  is an order reversing monotonic transformation of  $d_\rho^2(x, q)$ . However, since  $d_\rho(x, q)$  is monotonous this does not affect the ordering.

$t^5(x_1) = \text{"RO5 RO5 RO5 RO5 RO5 RO2 RO2 RO2 RO2 RO1 RO1 RO1 RO3 RO3 RO4"}$   
 $t^5(x_2) = \text{"RO4 RO4 RO4 RO4 RO4 RO3 RO3 RO3 RO3 RO5 RO5 RO5 RO1 RO1 RO2"}$   
 $t^5(x_3) = \text{"RO5 RO5 RO5 RO5 RO5 RO2 RO2 RO2 RO3 RO3 RO3 RO1 RO1 RO4"}$   
 $t^5(x_4) = \text{"RO3 RO3 RO3 RO3 RO3 RO5 RO5 RO5 RO5 RO2 RO2 RO2 RO1 RO1 RO4"}$

and for the query  $q$ :

$t^5(q) = \text{"RO5 RO5 RO5 RO5 RO5 RO1 RO1 RO1 RO1 RO2 RO2 RO2 RO3 RO3 RO4"}$

If we exploit the idea of taking just the closest reference objects to represent any object that has to be indexed, and assuming, for instance,  $k_x = 3$  (the number of reference objects used for indexing), and  $k_q = 2$  (the number of reference objects used for generating the query), the textual representations become:

$t^3(x_1) = \text{"RO5 RO5 RO5 RO2 RO2 RO1"}$   
 $t^3(x_2) = \text{"RO4 RO4 RO4 RO3 RO3 RO5"}$   
 $t^3(x_3) = \text{"RO5 RO5 RO5 RO2 RO2 RO3"}$   
 $t^3(x_4) = \text{"RO3 RO3 RO3 RO5 RO5 RO2"}$

and for the query  $q$ :

$t^2(q) = \text{"RO5 RO5 RO1"}$

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

## 2.1 Reordering Search Result

The algorithms described so far use an object representation in a transformed space and an object similarity measure based on a variation of the  $d_p$  measure to order the objects in the dataset in decreasing similarity with respect to the query. The result is an approximation of the exact result set that would have been obtained if the ordering of the objects was performed using the original distance  $d$  in the original data space.

Suppose we are searching for the  $k$  most similar (nearest neighbors) objects to the query. We can improve the quality of the approximation by re-ranking, using the original distance function  $d$ , the first  $c$  ( $c \geq k$ ) objects from the approximate result set at the cost of  $c$  more disk accesses and  $c$  distance computations. We will show that this technique significantly improves the accuracy, though only requiring a very low search cost. In fact, when  $c$  is much smaller than the size of the dataset, this extra cost can be considered negligible with respect to the cost of accessing the inverted file. For instance, when  $k$  is 10 and  $c = 1000$ , with a dataset size of 1,000,000 it means that we have to reorder a number of objects equivalent to just 0.1% of the entire dataset. Usually, as we will see in the experiments, this is not true for other access methods, for instance tree-based access methods, where the efficiency of the search algorithms strongly depends on the amount of objects retrieved.

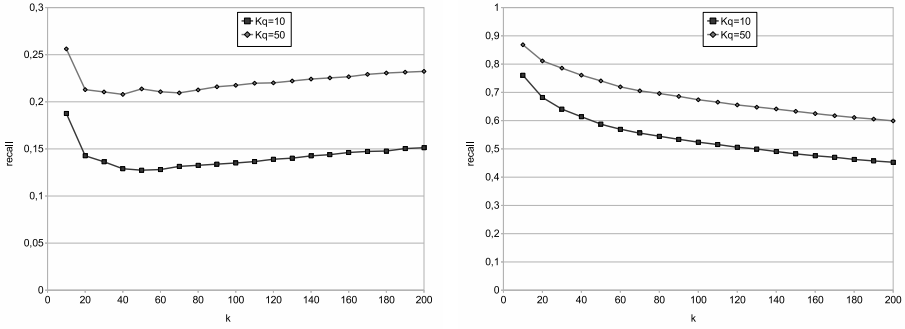
### 3 A Real Application and Performance Evaluation

In this section, we report the results of an experimental evaluation of the proposed method. For both testing and demonstration, we developed a web user interface to perform image content based retrieval on the CoPhIR dataset [3], which consists of 106 millions images, taken from Flickr ([www.flickr.com](http://www.flickr.com)), described by MPEG-7 visual descriptors. Content based retrieval can be performed by using similarity functions of the visual descriptors associated with the images.

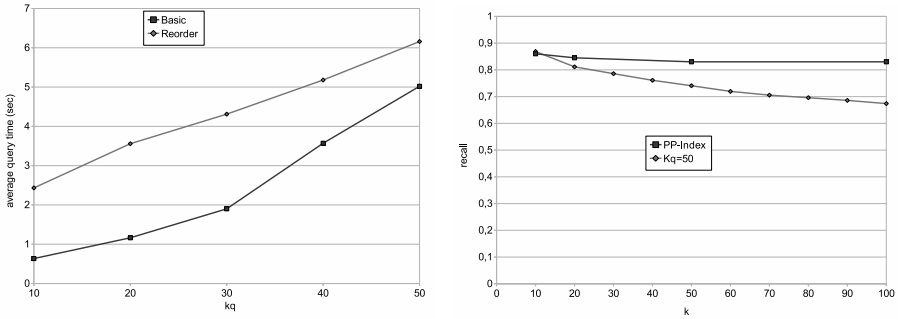
We have indexed the whole CoPhIR dataset and for each image, we created five Lucene fields which can be queried separately or in combination. The first field contains the unique identifier of the Flickr image. The second field maintains the textual information taken from title, and tags of the original Flickr image. The other three fields contain the content generated by the  $t$  function explained above for searching on three different pre-combined visual features. In particular, in order to support content based search, the CoPhIR project extracted several MPEG-7 visual descriptors from each image, three descriptors for describing the colors (SCD, CSD, and CLD) and two for describing textures (EHD and HTD). We have indexed three different aggregations of those descriptors, the first one combining the three color descriptors, the second one combining the two texture descriptors, and the third one combining all five descriptors. In this way we leave the possibility to the user to search for colors and textures independently or to search all the descriptors together. The weights used for aggregating the descriptors are the ones suggested in [2].

At the address <http://lucignolo.isti.cnr.it/> a demo web application of the developed search engine can be found. From that page it is possible to perform a full-text search, a similarity search starting from one of the random selected images. Besides the three types of visual similarities, thanks to the search functionality of Lucene, it also provides complex query processing by combining any of the three types of similarity search with the full-text search on descriptive metadata.

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



**Fig. 2.** Recall varying the number  $k$  for different values of  $k_q$  parameter. Basic (left), and Reorder (right).



**Fig. 3.** Query time for different values of  $k_q$  parameter (left) and comparison between our approach and PP-Index on the same data set (right)

Figure 3 (left graph) also shows the average query processing times as function of  $k_q$ , with and without reordering. As expected, the search cost is worse when use the reordering strategy but still acceptable, also considering the big improvement in terms of recall.

### 3.1 Comparison with PP-Index

A similar approach [6] (based on the on representing any indexed object with its view of the surrounding world), called Permutation Prefix Index (PP-Index), uses an index data structure that supports efficient approximate similarity search.

Figure 3 (right graph) shows the comparison of the recall between our approach and PP-Index on the CoPhIR dataset. Actually, PP-Index exhibits better performance. However, as explained in the introduction, the aim of our approach is to provide a tool for rapid development and integration of a multimedia object retrieval system with other digital libraries based on text. As a result, along

with the obvious advantage of having a system which relies upon an open source library that is constantly expanding, our method provides content based search combined with textual metadata.

## 4 Conclusions and Future Work

In this paper we presented an approach to approximate similarity search in metric spaces based on a space transformation that relies on the idea of perspective from a data point. We proved through a concrete implementation that the proposed approach has clear advantages over other methods existing in literature in terms of easiness in implementation. A major characteristic of the proposed technique is that it can be implemented by using inverted files, thus capitalizing on existing software investments.

This approach can take advantage of parallelism of Lucene and easily scales up to any desired dataset size. This can be obtained by distributing the inverted index in multiple Lucene segment, and exploding parallel search facilities of Lucene. For instance, our index consists of ten separated Lucene indexes each one including about 1/10 of the whole dataset. If the indexes reside on different physical disks, we may obtain performance improvements; however, in our tests conducted with a single physical disk, the performance with multi-thread search was slightly better than with a single-thread search.

## References

1. Amato, G., Savino, P.: Approximate similarity search in metric spaces using inverted files. In: Proceedings of the 3rd International Conference on Scalable Information Systems (InfoScale 2008), pp. 1–10. ICST (2008)
2. Batko, M., Kohoutkova, P., Novak, D.: Cophir image collection under the microscope. In: International Workshop on Similarity Search and Applications, pp. 47–54 (2009)
3. Bolettieri, P., Esuli, A., Falchi, F., Lucchese, C., Perego, R., Rabitti, F.: Enabling content-based image retrieval in very large digital libraries. In: Second Workshop on Very Large Digital Libraries (VLDL 2009), pp. 43–50. DELOS (2009)
4. Chavez, E., Figueroa, K., Navarro, G.: Effective proximity retrieval by ordering permutations. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 30, 1647–1658 (2007)
5. Esuli, A.: Pp-index: Using permutation prefixes for efficient and scalable approximate similarity search. In: Proceedings of the 7th Workshop on Large-Scale Distributed Systems for Information Retrieval (LSDS-IR 2009), pp. 17–24 (2009)
6. Esuli, A.: Use of permutation prefixes for efficient and scalable approximate similarity search. *Information Processing & Management* (2011)
7. Fagin, R., Kumar, R., Sivakumar, D.: Comparing top-k lists. *SIAM J. of Discrete Math.* 17(1), 134–160 (2003)

# Establishing a Digital Library in Wide-Ranging University's Context

## The Sapienza Digital Library Experience

Angela Di Iorio<sup>1</sup>, Marco Schaerf<sup>1</sup>, and Matteo Bertazzo<sup>2</sup>

<sup>1</sup> Sapienza Università di Roma, Rome, Italy  
{angela.diiorio,marco.schaerf}@uniroma1.it

<sup>2</sup> CINECA, Bologna, Italy  
m.bertazzo@cineca.it

**Abstract.** The Sapienza Digital Library (SDL) is a research project undertaken by Sapienza Università di Roma, the largest Europe's campus, and the Italian supercomputer center Cineca.

The SDL project aims to build an infrastructure supporting preservation, management and dissemination of the past, present and future digital resources, that contain the overall intellectual production of the Sapienza University. The solution adopted tries to find a tradeoff between the standardization of the digital processes and products (that allows a cost-effective centralized and shared management and curation), and the preservation of the peculiarities of scientific materials, belonging to disparate knowledge disciplines (that need to be digitally available for future initiatives, more specifically tailored to the designated communities).

**Keywords:** Digital library, Long term digital preservation, Digital curation, OAIS, METS, MODS, PREMIS, Controlled vocabularies.

## 1 Introduction

The Sapienza Digital Library<sup>1</sup> (SDL) is a research project undertaken by Sapienza Università di Roma (Sapienza), the largest Europe's campus, and the Italian supercomputer center Cineca, which is a non profit consortium made up of 47 Italian universities.

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.

Setting the future scenario of the SDL application, it has been evaluated and prefigured the large amount of research and knowledge materials, coming from such large and ancient University, as well as the variety of interests coming from such large and multidisciplinary community of stakeholders, and, last but not least, the potential uses of that material, for general and specialized communities of users.

---

<sup>1</sup> <http://sapienzadigitallibrary.uniroma1.it> (expected on January 2013).



The project was indeed conceived to manage the integration of a large volume of multiformat materials, and to enable their access through different devices, in order to fulfill the needs and the expectations of diverse communities, local, global, and future.

The actual state of experiences in digital libraries, in digital resources management, in digitization and in evolution of dissemination tools have suggested to examine new cost-effective solutions in the weaving factory of submission, archiving, and dissemination of digital resources. The solution adopted tries to find a tradeoff between the standardization of the digital processes and products, that allows a cost-effective centralized management, and the preservation of the peculiarities of scientific materials that need to be digitally available for future and specific initiatives.

## 2 Mission

The primary objective of the project is to provide Sapienza University with a modern digital library, comprehensive and open, which contains all digital materials produced by, held by, with ownership of, or granted to Sapienza.

The materials will be organized, catalogued, enriched and made accessible to the whole academic community and over.

## 3 Project's Objectives

The initial objectives that was detected are those essentially applicable to any kind of university or educational institution, and extending the vision, also to any institution that needs to manage digital material.

The objectives were firstly defined from the users point of view:

- Offering to the Sapienza's designated communities the opportunity to exploit digital materials owned and/or produced by Sapienza;
- Managing a broad variety of digital materials, born-digital and digitized;
- Archiving and preserving collections of images, audio/videos, 3D materials, scientific articles and datasets, special and valuable collections (private archives of scientists, work archives, etc.), museums/archives/libraries materials, scientific learning and teaching materials;
- Organizing, grouping, and indexing materials, supporting their browsing and searching on different dimensional views, and their reuse in different contexts;
- Optimizing, improving, and enhancing the value of digital materials throughout the web semantic technologies and social tools;
- Building a framework in the forefront, for the submission, dissemination, and preservation of Sapienza's digital assets, interconnected with the most important Italian, European, and International digital resources aggregators;
- Allowing the interoperable conversation with other kinds of information management systems (libraries/archives/musems/universities, open access repositories...).

More technical objectives connected to the more stringent organizational and technical requirements in harmonization with the global digital libraries, and the digital curation scenario, were defined as the following:

- Managing digital materials coming from former digitization projects, making retrospective conversions of existing materials;
- Gathering as much as possible information, lowering the threshold of information lost;
- Achieving a satisfying level of information not only for user needs, but also for enabling advanced services for preservation and dissemination;
- Enriching the information making it reusable and connectable with other application contexts,
- Adhering to the Open Archival Information System (OAIS)[1] functional model and developing compliant services supporting the Long Term Digital Preservation (LTDP);
- Adopting the most spread digital libraries and digital preservation metadata standards, in order to maintain and to guarantee the interoperability of the SDL system with other systems, supporting the worldwide dissemination of digital resources;
- Adopting platforms and tools based on open source solutions.

## 4 Application Context

Sapienza university was founded in 1303 by Pope Boniface VIII and nowadays has 145,000 students, over 4,500 professors, almost 5,000 administrative and technical employees. The Sapienza organizational units for learning and scientific investigations, cover almost all disciplines of knowledge, and are divided into 11 colleges and 68 departments. The Sapienza memory organizations are represented by 59 libraries, 20 museums and 2 main archives, current and historical.

Collecting and managing the intellectual materials that was, is, and will be produced by that large organizational scenario, needs of a common, and a cost-effective solution, which leveraging on standardized digital resources, will allow their management and exploitation in the long term.

By the digital management point of view, the application context is extremely fragmented, because of the multiplicity of information sources that had produced digital resources in a local view and with personalized methodologies. Actually, scientists are simply more focused on their studies, researches and interests, than on the digital management of their information resources. For this reason, finding a common way to organize, manage, and exploit the content of digital materials, is essential to provide useful tools that support and ease the intellectual work, and lower the weight of the daunting task of managing digital resources.

## 5 Digital Materials Used

At the beginning of the project was necessary to prepare an initial census of the existing digital materials that were representative of specific type of objects like for example videos, images, digitized books, text documents, big images like historical maps. All the materials were gathered and stored into a dark repository of the Sapienza computing center. The materials were used as samples for the workflow of digital resources building that has led to the creation of the Submission Information Package (SIP) as required by the OAIS model[1].

Different types of SIPs were modelled on resources' types (i.e. image, video, map) and system's services were coherently modelled for ingesting, managing, and accessing the content. For example, even though maps and a photographs are both images, the fruition service provided by the system is different in regard to the image's dimension.

In general, the materials that, the SDL will be able to manage, are:

- Books, ancient (before 1831) and modern, prints, maps and other digitized materials;
- Scientific digital products (Ph.D. thesis, materials with rights, datasets...);
- Images, Audio/Video materials digitized and born-digital;
- Learning objects
- User Generated Contents
- Special materials: i.e. archaeological documentations, personal archives...
- 3D objects

## 6 SDL Reference Models and the Preservation Strategy

The reference models that lay down the digital library design and conception are the Open Archival Information System (OAIS)[1] and the DELOS digital library Reference Model (DELOS)[2].

Specifically, the DELOS Three-tier framework composed by Digital Library(DL), Digital Library System(DLS), and the Digital Library Management System(DLMS) were envisaged in the Sapienza context.

In conformance with DELOS model, the Sapienza DL is a set of “real” persons and organizational units that *“collects, manages and preserves for the long term rich **digital content**, and offers to its **user** communities specialised **functionality** on that content, of measurable **quality** and according to codified **policies**”*.

The Sapienza DLS is a distributed architecture tailored on and used by the different communities and provides specific software tools.

The Sapienza DLMS is the software infrastructure which was conceived following the philosophy of the “extensibility”, in order to implement tailored services in harmonization of the integration of new content types (i.e. specific visualization tool for big images) as well as the introduction of new requirements for the system (i.e. application of new information classification system).

A supposed DLMS is usually founded on the OAIS conceptual model, and usually its archiving repository provides the basic functions like ingestion, archival storage, data management, administration, preservation planning, access. Actually, very few DLMS are equipped with a complete and overall preservation planning, likewise coherently, the relevant administration. Indeed, the LTDP needs more integration between the technological support and the digital preservation organizational need, which is especially expressed throughout the organizational commitment, as evidence of the sufficient level of awareness about the LTDP.

## 7 Activities Project's Overview

The project has started in January 2011 and it was divided into two work phases.

The objective of the first phase, was to release a DLMS prototype implementing all the macrofunctionalities defined by the OAIS: ingesting, archiving and access. To release the DLMS prototype, it was necessary to design the pre-ingestion activities for the SIP building, and contemporaneously, it was defined and progressively improved the metadata framework, useful to support the information management. Consequently, the outcome of the first phase was prototyping, the SDL metadata framework (7.3), the Sapienza pre-ingestion workflow (7.4), the Sapienza SIP building (7.5), the SDL DLMS (7.6).

The first phase of the project has been closed in December 2011.

The second phase started in January 2012 and the objectives are: 1) developing DLSs for making communities to interoperate with SDL, 2) enriching the prototyped elements, released during the first phase, by adding metadata enabling the digital preservation strategies implementation, 3) optimizing the overall DLMS functionalities.

The following describes the activities performed for the first phase of the project.

### 7.1 System's Requirements Analysis and Matching of Digital Materials Characteristics

The census of available digital resources has resulted in a first categorization of materials types, and in a list of characteristics, that need to be managed by the system's services in supporting the main OAIS functionalities like ingest, archiving and access. The characteristics of materials were modelled taking into account the user needs, and consequently the differentiated access types, the variety of searching/browsing, the preservation needs and the draft of the rights management with the *corpus* of permissions, statements and other generic constraints.

From that initial analysis was designed a workflow of the materials processing, in order to prepare them for the submission to the SDL digital repository. The processing objective is the creation of a SIP, which in real implementations is a compound object made of content objects and metadata objects.

## 7.2 Selection of the Digital Repository Application System

The choice of the digital repository management system is a strategic and constraining decision for the digital curation of the materials. Consequently, an initial evaluation of the most spread *open source* software projects was done, and the Fedora Commons (FC)[3] has been chosen because, in spite of its complexity is more oriented to the web services integration, the semantic web technologies, and the LTDP. Furthermore, FC gives chances to use *content models*, that can be customized in regard to the originating models of the SDL materials.

As usual in implementing FC, an analysis of the *pros* and *cons* about the atomistic and compound paradigm was done, in relation to the projects requirements. The choice was led on the atomistic model, mainly considering the long term perspective of the project, which foresees to use and reuse the digital materials in diverse contexts. The greater flexibility, in reusing digital objects, was considered a good reward respect to the major complexity in managing the atomistic model, due to the system maintenance of information about relationships among objects.

## 7.3 Definition and Design of the SDL Metadata Framework

The metadata framework conceived for SDL had taken into account the wide-ranging general requirements that set three specific characteristics: completeness (gathering as much as possible information), flexibility (adapting to different contexts) and extensibility (integrating with new information).

Consequently, the framework has to be able to hold any kind of resources' description. The holding of information does not mean that the managing system has necessarily to manage it, but holding information and maintaining it available, would allow its reuse in future focused projects.

The metadata framework has to support the following requirements:

- conformant with OAIS;
- prearranged to hold different standard descriptions on which implementing integration services, supporting the use of wide-ranging knowledge's materials;
- prearranged to the exchange with other digital library systems or other information management systems;
- prearranged to the LTDP and equipped with the minimal and essential metadata, enabling the long term management.

The metadata is generally categorized in descriptive, administrative, structural rights management, preservation<sup>2</sup>, and technical and use<sup>3</sup>, even though same metadata can

---

<sup>2</sup> *Understanding Metadata*, National Information Standards Organization, 2004, [www.niso.org/standards/resources/UnderstandingMetadata.pdf](http://www.niso.org/standards/resources/UnderstandingMetadata.pdf)

<sup>3</sup> Tony Gill, Anne J. Gilliland, Maureen Whalen, and Mary S. Woodley Edited by Murtha Baca Introduction to metadata Online Edition, Version 3.0 [http://www.getty.edu/research/conducting\\_research/standards/intrometadata/index.html](http://www.getty.edu/research/conducting_research/standards/intrometadata/index.html)

be assigned to different categories, in regard to the use perspectives. The actual scenario, in DL implementations, highlights the broad adoption of metadata standards, coming from metadata specialists international workgroups, supported by the Library of Congress. The most adopted combination is METS[4]/MODS[5]/PREMIS[6], where PREServation Metadata Implementation Strategies (PREMIS), is used for preservation metadata, Metadata Object Description Schema (MODS) for descriptive metadata and Metadata Encoding & Transmission Standard (METS) for wrapping metadata all together. Usually the metadata standards are available to the communities by means of XML schemas<sup>4</sup> that enable the information systems to interchange, set of information encoded in XML files.

The SDL has adopted primarily LOC standards because of the wide-adoption in DL projects and, the more the standards are spread and the more the adopting systems are interoperable. In addition, because they are open standards it follows that the longevity of their knowledge-base is likely longer.

### **Descriptive Metadata Set**

The SDL metadata framework was designed to support as much as possible information, even coming from different kinds of knowledge provider. The MODS metadata is the “core description” on which are configured the DLMS’s services for searching and browsing of the SDL collections. A stable MODS profile will be released during the second phase of the project, and it will be the reference descriptive framework for describing new digital materials. All the varied information sets, collected from the different Sapienza organizational units are mapped and encoded in MODS, and enriched by controlled values taken from the MODS controlled vocabularies as well as, the SDL controlled vocabularies.

The translation of different information sources into the MODS has respected and followed the Digital Library Federation/Aquifer Implementation Guidelines for Shareable MODS records[7]. The elements required by the DLF/Aquifer requirement level, has been adopted as one of the SDL policies for the basic requirement level in resource’s description. Furthermore, for special collections it has been taking into account the Master Data Element List of Library of Congress Metadata for Digital Content[8].

The MODS has been used for describing materials, not only at the single item level, but also at the collection level. Every item or resource (here meant as a discrete unit, conceptually equivalent to the OAI Information Package (IP), in this article specifically qualified as SIP in 7.1), existing in SDL, must belong to an identifiable collection, that indeed is described by MODS elements.

The MODS was considered more suitable for the SDL metadata framework, because is richer than the easiest implementable Dublin Core, it being understood that the DL system can dumb down from MODS, to Dublin Core<sup>5</sup>, and similar in simplicity, as well as to map toward open data standards<sup>6</sup>.

---

<sup>4</sup> XML Schema, <http://www.w3.org/XML/Schema>

<sup>5</sup> The Dublin Core Metadata Initiative, <http://dublincore.org/>

<sup>6</sup> Linked Data, <http://linkeddata.org/>

Mostly, all advantages and features, listed on MODS official website, were considered important in implementing it, but one of them deserves to be cited: MODS can be used also for Search/Retrieval via URL(SRU)<sup>7</sup> as specified format, enabling federated searching and similar automated queries via URL, which means to offer further advanced services to the users.

### **Metadata Container and Structural Metadata**

For packaging all metadata together into the defined SIP, the SDL metadata framework devised, has exploited the flexibility of METS, making the system available 1) to collect other kinds of metadata set, over those specifically adopted, and 2) to dumbing down toward standards less complex like Dublin Core or European Semantic Elements<sup>8</sup>.

Whenever is necessary to collect resources' descriptions more detailed than MODS, these descriptions are stored "as is" and summarized, according to the SDL MODS profile, into the SDL MODS core description.

Thanks to the METS flexibility, during the development of the per-ingestion activities and the improvement of the metadata framework design, other metadata standards have been embedded, like for example technical standards specific for different kinds of materials (MIX<sup>9</sup>, VideoMD<sup>10</sup>).

A stable METS profile will be released during the second phase of the project.

### **Preservation Metadata**

The overall SDL SIP building workflow was pervaded by the LTDP philosophy, ensuring the basic provision of the preservation metadata, considered mandatory by the PREMIS, that is the preservation metadata framework mapped from the conceptual structure of the OAIS model. The SDL metadata framework was designed to guarantee the minimum conformance with the PREMIS standard both on semantic unit and data dictionary level, following requirements and constraints, and by collecting all the metadata defined as mandatory by the PREMIS Data Dictionary[9]. Although PREMIS is not formally and completely adopted yet, all the mandatory information were encapsulated into the metadata framework, and all the other useful information were stored by the SDL black repository system and will be encapsulated during the project's second phase.

This means that the SDL resources are already equipped for the management of the LTDP strategies, even though the DL prototype is not managing them yet. The preservation planning and administration will be implemented in the second phase of the project.

---

<sup>7</sup> Search/Retrieval via URL, <http://www.loc.gov/standards/sru/>

<sup>8</sup> European Semantic Elements, <http://www.europeanlocal.eu/eng/Document-Library/Reports/ESE-Semantic-Elements-ver-3.1>

<sup>9</sup> NISO Technical Metadata for Digital Still Images Standard, <http://www.loc.gov/standards/mix/>

<sup>10</sup> AudioMD and VideoMD technical metadata for Audio and Video, <http://www.loc.gov/standards/amdvmd/index.html>

## 7.4 Pre-ingestion Workflow

Considering the extension of the University, in numerical as well as in geographical terms, the workflow was designed following a distributed view of the work, even though in order to design the workflow, at the beginning a centralized system of the materials' treatment is necessary. The workflow will be applied and distributed in the different institutional units belonging to Sapienza at the first stable system release.

The materials gathered into the SDL dark repository were of different varieties, in terms of contents and metadata structure. The first step was to maintain the provenance source of materials identifying firstly the real Sapienza Organizational Unit, that asked to submit materials to the system, and secondly, identifying the collections and items contained.

At the beginning, the workflow was designed and tested on a sample collection of almost 2000 images, and progressive tests of SIP building processes(7.5) had fixed and integrated both metadata framework(7.3), pre-ingestion workflow activities and the SIP building processes.

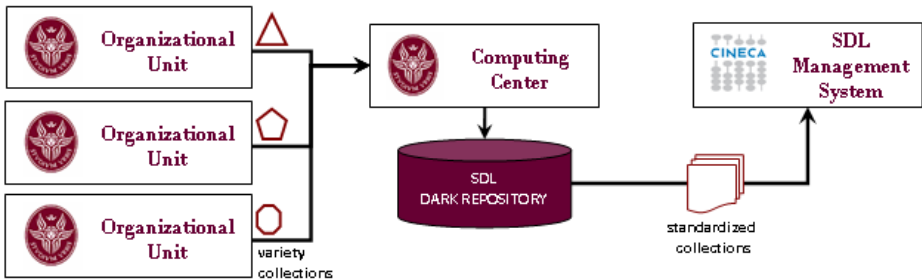


Fig. 1. Pre-ingestion workflow overview, from variety to standardized collections

The pre-ingestion workflow consists of all those activities necessary to prepare digital objects and metadata for the automatic building of the SIP, in conformance with the SDL metadata framework defined. In practice it organizes, structures and enriches the flow of digital materials from Sapienza organization units toward the Sapienza SIP building DLS.

The design and development of the workflow has essentially consisted of the following activities:

- Detection of available digital materials;
- Analysis, normalization and enrichment of both metadata and digital objects;
- Modelling of resources and resources' collections information for the submission objective;
- Modelling of provenance information being collecting in pre-ingestion activities;
- Designing and development of a local database as “metadata nursery” for the production of the final SIP's version.



## 7.5 Selection of Representative Samples and SIP Building

From the analysis of the existing materials were identified the representative samples of diverse types of materials (collections, videos, images, maps, books), to which applying specific content model. The building process of the SIP was applied to the samples selected and was performed throughout activities like the normalization of files' naming, the organization of collections, resources and digital objects, the creation and encoding of metadata files. The building process was tested and integrated many times during the experiment of the *ingestion*, until the maturity of the metadata framework.

At the conclusion of the first phase of project the SIP building was tested in 6 retrospective conversions of existing collections, compounding images and video, and described by spreadsheets and database information.

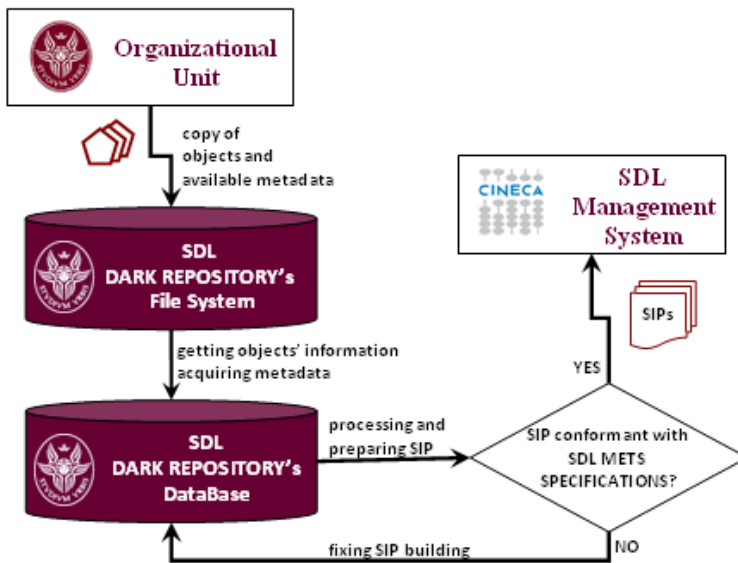


Fig. 2. SIP building development

For each sample were defined constraints on the information structure, which characterizes the originating model and unleashes the relevant FC content model.

The SIP building process consists of transforming the sources information (databases/datasets/spreadsheets/systems folders) in metadata, enriching it with provenance, context, reference, fixity (OAIS) metadata, and with basic knowledge domain semantics, encoding it in XML files valid for the XML schemas specifications, and combining them following the METS profile specification of the SDL metadata framework.

The SIP built has the following characteristics:

- based on METS as metadata container;
- encompasses different descriptive and administrative standards;

- conformant to specific metadata standard guidelines (i.e. DLF aquifer);
- conformant to Sapienza customized FC content models;
- exists independently and redundantly apart from the DLMS;
- open to the inclusion of other metadata standards that are more descriptive than MODS core description.

## 7.6 SDL DLMS Prototype

The SDL DLMS gets in ingestion the SIP built in conformance with the SDL metadata framework, it archives the objects and information in FC, and makes the SIP's resources available to the SDL exploration's system which allows to navigate, browse and search resources.

The prototype is internally available in order to give the opportunity to the project's community of participating to the optimization of the UI and Users' services, as well as testing functionalities, and improving the architectural information structure of the web interface. At this moment the prototype was peopled with 6 collections containing almost 15,000 items, differentiated in four resources types, images, books, videos, and maps. Furthermore, was uploaded a collection created *ad hoc* for the prototype's homepage, that was created by reusing resources, already ingested in the system. This has verified that the system holds all necessary information useful to create new digital objects (items or collections), that are aggregations of existing resources.

## 8 Future Developments

The second phase of the project will provide the DL with a set of specific DLS which will essentially support 1) the submission of new digital items and collections, 2) the participation of Sapienza community for collecting new materials, 3) the optimization of dissemination tools by means of customized interfaces (web users, OAI-PMH, Web Services...), and the integration of services for the exhibition/export of metadata for third party resources aggregators: InternetCulturale<sup>11</sup>, Europeana<sup>12</sup>, World Digital Library<sup>13</sup>, 4) the integration of metadata supporting preservation planning and administration.

The workflow of the overall materials submission will be ruled by SDL guidelines wherein will be defined the Sapienza digital policies. The guidelines will harmonize the way of creating, and producing digital resources for the DL: specific paths for different content models will be described in order to support the community in being aware about the digital resources' management.

Stable METS and MODS profiles will be released during this phase.

At the end of 2012, the system will be released in production for the main functionalities, and will be open to other partners.

---

<sup>11</sup> <http://www.internetculturale.it/opencms/opencms/it/>

<sup>12</sup> <http://www.europeana.eu/portal/>

<sup>13</sup> <http://www.wdl.org/en/>

## References

1. Consultative Committee for Space Data Systems Reference Model for an Open Archival Information System (OAIS), Blue Book. Issue 1 (January 2002), <http://public.ccsds.org/publications/archive/650x0b1.pdf>
2. Candela, L., Castelli, D., Ferro, N., Ioannidis, Y., Koutrika, G., Meghini, C., Pagano, P., Ross, S., Soergel, D., Agosti, M., Dobрева, M., Katifori, V., Schuldt, H.: The DELOS Digital Library Reference Model - Foundations for Digital Libraries, Version 0.98 (February 2008), [http://www.delos.info/files/pdf/ReferenceModel/DELOS\\_DLReferenceModel\\_0.98.pdf](http://www.delos.info/files/pdf/ReferenceModel/DELOS_DLReferenceModel_0.98.pdf)
3. Fedora 3.5 Documentation, Content Model Architecture, <https://wiki.duraspace.org/display/FEDORA35/Content+Model+Architecture>
4. Metadata Encoding & Transmission Standard (METS), <http://www.loc.gov/standards/mets/>
5. Metadata Object Description Schema (MODS), <http://www.loc.gov/standards/mods/>
6. PREservation Metadata Implementation Strategies (PREMIS), <http://www.loc.gov/standards/premis/>
7. Digital Library Federation/Aquifer Implementation Guidelines for Shareable MODS Records, [https://wiki.dlib.indiana.edu/confluence/download/attachments/24288/DLFMODS\\_ImplementationGuidelines.pdf](https://wiki.dlib.indiana.edu/confluence/download/attachments/24288/DLFMODS_ImplementationGuidelines.pdf)
8. Metadata for Digital Content Developing institution-wide policies and standards at the Library of Congress, <http://www.loc.gov/standards/mdc/elements/>
9. PREMIS Editorial Committee, PREMIS Data Dictionary for Preservation Metadata version 2.0 (March 2008), <http://www.loc.gov/standards/premis/v2/premis-2-0.pdf>

# Digital Curators' Education: Professional Identity vs. Convergence of LAM (Libraries, Archives, Museums)

Anna Maria Tammaro<sup>1</sup>, Melody Madrid<sup>2</sup>, and Vittore Casarosa<sup>3</sup>

<sup>1</sup> University of Parma

<sup>2</sup> International Master DILL, University of Parma

<sup>3</sup> ISTI-CNR, Pisa

**Abstract.** Digital curation education is a new subject where the convergence between libraries, archives, museums and computer science seems to build an interdisciplinary bridge, with common competences needed by present and future professionals. The study methodology is based on: the literature review, on the proceedings of the Puerto Rico Conference organised by IFLA on “Education for Digital Curation” and on the findings of a Delphi study which has been done for a Thesis of the International Master DILL. Issues and problematic areas for further study and discussions are evidenced.

**Keywords :** preservation, digital curation, digital library education.

## 1 Introduction

The problem of digital preservation is a new area of study, in which converge the activities and research of different disciplinary sectors, which can be identified as computer science, digital archives, digital museums and libraries. The literature developed in various areas of research offers different definitions of the area of study, due to diverse interpretation and meaning given by professional communities to the words “preservation”, “archiving”, “information” and “data”. To overcome (to some extent) the communication problems between the different disciplinary and professional sectors, a new term “digital curation” was recently adopted, which joins two preexisting terms “curation” and “digital preservation”.

Curation is a term used both by cultural institutions, such as libraries, archives and above all museums as well as by scholars and creators of large databases such as Genome. It indicates those activities that add value and knowledge to the collections, and the added value is usually given by the curator or manager of the cultural institution. The curator is often a specialist in the field and through her competence she enriches the collection in a variety of ways. First of all the curator is an expert in the activity of selecting the collection items, so that the value of the whole collection is greater than the sum of the values of its items. The services provided (by the curator) give evidence to this added value, and also, since the curator is able to interpret the significance of the collection and to communicate it to the users, the services can assume an

educational and personalized role. The curator has also technical competences, such as the indexing of the collection to facilitate browsing and retrieval, and the enrichment of the documentation (metadata) to provide for the single objects additional information about their descriptive and historical context. In smaller institutions, the curator offers general support to users.

The American Association for Museums Curators describes the curator's role in this way:

*“Regardless of their situation, curators have distinctive responsibilities that focus upon: 1) the interpretation, study, care, and development of the collection, and 2) the materials, concepts, exhibitions, and other programs central to the identity of their museum. Because of their direct responsibilities for the collection and their role in the development of interpretive material, curators are ambassadors who represent their institution in the public sphere”.*

The term Digital preservation is used by the Library of Congress (<http://www.digitalpreservation.gov>) with this meaning:

*“Digital preservation is the active management of digital content over time to ensure ongoing access”.*

The focus of digital preservation is upon the collection, management and permanent access to digital resources, in particular those which are “born digital” and therefore have no physical counterpart. In 2010 the Library of Congress conducted a survey of the educational needs for digital preservation and identified three levels of competences: practical, managerial, executive. The practical competences are essentially technical, based upon standards and the technology applications necessary for the management and preservation of digital objects. The management competences are mainly those related to the management of digitization projects, and the executive competences are those related to a strategic vision and the continuous updating of the preservation activities. For libraries, it is interesting to note in some of the answers received during interviews that the activity of preservation is not considered a competence needed by all librarians, but rather it is perceived as a specialized competence for a small group or even to be left in the hands of specialists outside the library. The term preservation, as well as the similar one “archiving”, is in fact traditionally perceived as an activity at the end of the workflow for the management of digital resources and thus considered separated or isolated from the vital flow of the creation, organization and circulation of the resource.

The present paper intends to define the state of the art of the convergence for “digital curation” and is based upon the acts of recent IFLA Conferences in which sessions dedicated to Digital Curation were held, and upon the results of a Delphi study carried out in the research thesis of Melody Madrid, a student of the International Master DILL (Digital Library Learning). The paper does not attempt to be exhaustive, but is limited only to outlining the problems which emerge from the

meeting of different disciplinary communities, although in the convergence of a common area of interests. Surely the study of the convergence of archives, libraries, museums, together with other professional sectors such as computer science, must be elaborated on with further research.

The convergence has a notable impact on the teaching to new professionals as well as on the retraining of the staff in service. For this reason, the paper concentrates on the problem of competences which are deemed to be necessary for digital curation in an interdisciplinary approach.

## 2 Digital Curation: History of the Concept and Competences

The term “digital curation” was used for the first time during the Seminar “Digital curation: digital archives, libraries and e-science” held in London in 2001 by the Digital Preservation Coalition and British National Space Center (Beagrie 2006). The term digital curation was later adopted by JISC (<http://www.jisc.ac.uk/>), which established the Digital Curation Centre (DCC 2004).

The Digital Curation Centre (2004) so defines the concept, introducing the notion of “adding value”:

*“Digital curation broadly interpreted is about maintaining and adding value to a trusted body of digital information for current and future use”*

Beagrie N. (2006) in his paper “Digital curation for science, digital libraries and individuals”, widens the concept by introducing the notion of “entire life cycle”:

*“Actions needed to maintain digital research data and other digital materials over their entire life cycle and over time for current and future generations of users”*

The concept of digital curation was thus born with the idea of building bridges between different disciplinary approaches and arises from the initial knowledge of the scholars that a new approach is necessary for the care and preservation of digital assets during their life span. As Harvey (2011) has written, the new approach is characterized by new competences:

*“Among these are the ability to function comfortably in both digital and physical mediums, to move seamlessly and efficiently between both mediums, to recognize and respect the core differences between information disciplines as well as between the information content itself, and to negotiate the ways in which digital environments can overcome information silos to create a universe of access across institutionalized boundaries”*

The first concept which unifies the different disciplinary approaches is the life cycle of the digital resources. This common approach to the life cycle brings with it two aspects: 1) the first is that preservation should no longer be perceived as a final phase

separated from the creation and access to the digital resource; 2) the second brings forward the need for collaboration among all the stakeholders who participate in various roles and in different phases of this life cycle.

A first curriculum for digital curation was promoted in 2008 by NARA, the US National Archives and Records Administration. DIGCCURR, the project that followed this stimulus, has developed a matrix of knowledge and competences based on 23 functionalities, which are pragmatically based on the work flow (Lee, 2009) (<http://ils.unc.edu/digccurr/digccurr-functions.html>). The life cycle of digital resources which was taken as a model was the one defined in the OAIS model. Among others, one of the results of the DIGCCURR project has been to show the need for internships and hands-on experience for digital curators.

## 2.1 Operational Competences of Digital Curator

The operational competences of the digital curator are essentially related to the technical functionalities described in the OAIS model. Is the digital curator a computer scientist, or rather a professional who collaborates with a computer scientist?

The presentations at the IFLA Conferences have proposed this problem again, where the discussion was opened. Casarosa (2011) has pointed out that the professionals must be aware of the technologies and standards necessary for digital preservation, together with other competences which regard trust and trustworthiness in the context of digital preservation, and appreciation of the roles and responsibilities involved in digital preservation activities.

Repanovici (2011) has proposed a curriculum on digital curation to Engineering and LIS students asking them to quantify their preferences, with the following results:

*“Both ENG and LIB students are interested in the following courses: Conservation by digitization and Archiving web pages. The students from ENG are interested in Techniques of security against electronic theft, while the LIB students are interested in Methods of press monitoring, Legislation on culture”.*

Bahr (2011) relates the results of a survey on the educational needs of the staff involved in the Leibniz Project, pointing out that both librarians and information technicians involved in the project indicate basic educational needs:

*“Profound knowledge of content related criteria and technology related criteria exists. However, applying this knowledge in the context of digital preservation is not always integrated into digital curation practice of content experts and information technology experts alike”*

The technical competences described by DIGCCURR are listed in the table below:

Theme	Outline	DIGCCURR
Document and artefact management – physical and virtual	Focus can be at level of organization/institution, information system (e.g. record-keeping system), collection, or individual items.	<p>Characterization of digital objects within information package</p> <p>Characterization of information package</p> <p>It includes assessments of recordkeeping systems and authenticity of documents within those systems.</p>
Document design on the Internet	Services and functions used for the storage and retrieval of Archival Information Packages	<p>Disaster planning, preparation and response</p> <p>Ensuring sufficient redundancy of copies</p> <p>Error checking</p> <p>Holdings maintenance</p> <p>Management of storage hierarchy</p> <p>Providing data, Receiving data, Replacing media</p>
Information retrieval (using information systems to locate documents and information)	Making digital resources available to Consumers.	<p>Coordination of access activities</p> <p>Delivery of responses; Exposure</p> <p>Generation of access collections; Generation of Dissemination Information Package (DIP)</p> <p>Information discovery; Information retrieval; Legal discovery; Viewing</p>
Data Management	Design and maintenance of the intermediate data structures that are used to manage and provide basic access to digital data e.g. file systems, Extensible Markup Language (XML) data elements, and catalog data within data grids.	<p>Administering database</p> <p>Generating reports</p> <p>Linking/resolution services</p> <p>Performing queries</p> <p>Receiving database updates</p>



Theme	Outline	DIGCCURR
Identifying, Locating & Harvesting	Identification, locating and harvesting (i.e. "gathering up") aggregates of resources, for purposes other than direct and immediate use of the resources.	Defining and setting parameters for harvests and file requests Extracting identifier information to determine network locations of resources Harvesting metadata from external sources or repositories Making requests to appropriate locations to collect resources Synchronizing content

## 2.2 Management Competences of Digital Curator

The professional competences characterizing the different profiles are those which are usually considered the core of the profession. These competences in fact are considered to be the identity of the profession, as they are based upon the basic principles and the specific mission of each profession. What is the impact of the convergence on this professional identity?

The different identities of the information professions are still present, but there is a trend towards their change and technological convergence. In "Cyberinfrastructure, Data, and Libraries, Part 1 A Cyberinfrastructure Primer for Librarians" Anna Gold (2007) discusses the need for librarians to extend their competences to the phase preceding publication, while traditional background concentrates on the phase of dissemination and access, following publication. This knowledge is added to the needed technical skills, which may include data management, data archiving, digital preservation, the semantic web, and the linked open data.

From the world of archives, Margareth Hedstrom in 1991 highlighted the problems of digital archiving, many of which are still relevant. The first problem that was highlighted was the lack of technological competences, of which (at that time) the archivists had no knowledge due to the novelty of digital resources. Other more conceptual problems are also described in the work, such as the necessary collaboration with other professions for the gathering of contextual information, up to the point of questioning some pillars of the traditional archival theory. In synthesis, digital archiving is different from record keeping and record management.

In "The Institutional Repositories: Staff and Skills Set ", Robinson (2009) describes the knowledge and skills needed by repository managers and administrators and arranged them into nine categories: management; software; metadata; storage and preservation; content; advocacy, training and support; liaison (internal); and liaison (external); and current awareness and professional development.

### 2.3 Strategic Competences of Digital Curator

The strategic competences of the digital curator are competences at an upper level, directed at describing and defining the policy of an institution, and at a national level contributing to the strategies of information policy and administration, finalized to the management and the preservation of the digital collections (Harvey, 2010, 2011), including also high-level categories of digital curation functions (Lee, 2008).

Pomerantz et al. (2009) have compared the competences of the digital librarian and of the digital curator, coming to the conclusion that there are no major differences. Both curricula begin with the same model of the information cycle, and analyzing both the impact that the context has on the actors, and the instruments and the functionalities that are necessary, no major difference is evidenced. What seems different is only a diverse focus on the preservation of the collection and the care of the digital objects.

Harvey (2011) describes the necessity of beginning with the users and their access needs in order to obtain the balance between the different disciplinary approaches, since although they are different, they all focus on access:

*“Balancing these user needs with respect for the core theories that ground the various aspects of the heritage the materials come from requires a deep understanding of different disciplines as well as of the digital options for convergence and display of the materials of that heritage. The fundamental principles of cultural heritage convergence should relate to maintaining the balance, so that the very different, but equally relevant, missions of libraries, archives and museums are not lost or subsumed in the desire to bring cultural materials and other information together”*

## 3 Delphi Study: Core Competences of Digital Curators

In “A Study of Digital Curator Competences: A survey of experts”, Madrid (2011) defined and validated competence statements for Libraries, Archives and Museums (LAM) digital curators through a Delphi research technique. The research intended to get an equal number of participants from the Library, Archives and Museum sectors, but no reply was received from expected participants in the Museum sector. However, the panel members who responded to this study were university professors or researchers concerned with digital curation and preservation in the LAM sector, which is now considered an interconnected profession.

Using a modified Delphi method, three rounds of questionnaires with controlled feedback and space for comments and/or suggestions were sent to the panel members. The questionnaire was requesting to assess, on a five point Likert scale, the agreement with a set of statements about competences needed in Digital Curation. Consensus was determined when a competence statement received a rating higher than 3, an average value greater than 3.5, and a standard deviation smaller than 1.0. Response rates for rounds I, II and III were: 70% (n=16), 87.5% (n=14), and 94% (n=15) respectively. Of the 18 digital curator competences listed in the first round

questionnaire, 13 (70%) achieved consensus as being necessary competences, required for advanced level digital curator. Other input from respondents such as comments and suggestions were also analyzed. An additional 23 digital curator competence statements were also suggested by the panel in round I and further developed in subsequent rounds. In round II, 12 (30%) competence statements achieved consensus. The final round and editing of competence statements led to 20 statements that describe what a well-prepared digital curator, trained to participate in digital curation work, should be able to do.

The definition of Digital Curator which has been agreed by the experts participating to the Delphi study is:

*“Digital curators are individuals capable of managing digital objects and collections for long-term access, preservation, sharing, integrity, authenticity and reuse. In addition, they have a range of managerial and operating skills, including domain or subject expertise and good IT skills”*

The list of the 20 statements is divided in Operational and Managerial competences to maintain the structure of this paper, but the statements were the result of an holistic approach.

### **3.1 Operational Competences**

The operational competences of the digital curator which were agreed by the experts participating in the Delphi study are as follows.

The Digital Curator:

1. Selects and appraises digital documents for long-term preservation.
2. Has an expert knowledge of the purpose of each kind of digital entities used within the designated community and its impact on preservation.
3. Knows the data structure of different digital objects and determines appropriate support needed.
4. Understands storage and preservation policies, procedures and practices that ensure the continuing trustworthiness and accessibility of digital objects.
5. Is aware of requirements for information infrastructures in order to ensure proper access, storage and data recovery.
6. Diagnoses and resolves problems to ensure continuous accessibility of digital objects, in collaboration with IT professionals.
7. Monitors the obsolescence of file formats, hardware and software and the development of new ones (e.g. using such tools as PRONOM registry)
8. Ensures the use of methods and tools that support interoperability of different applications and preservation technologies among users in different locations.
9. Verifies the provenance of the data to be preserved and ensures that it is properly documented.
10. Has the knowledge to assess the digital objects' authenticity, integrity and accuracy over time.

### 3.2 Managerial Competences

What are the main management responsibilities of the digital curator? The managerial competences which were agreed by the experts participating in the Delphi study are as follows.

The Digital Curator:

1. Plans, implements, and monitors digital curation projects.
2. Understands and communicates the economic value of digital curation to existing and potential stakeholders, including administrators, legislators, and funding organizations.
3. Formulates digital curation policies, procedures, practices, and services and understands their impact on the creators and (re)users of digital objects.
4. Establishes and maintains collaborative relationships with various stakeholders (e.g., IT specialist, information professionals inside and outside the institution, data creators, (re)users and other stakeholders like vendors, memory institutions and international partners) to facilitate the accomplishment of digital curation objectives.
5. Organizes personnel education, training and other support for adoption of new developments in digital curation.
6. Is aware of the need to keep current with international developments in digital curation and understands the professional networks that enable this.
7. Understands and is able to communicate the risk of information loss or corruption of digital entities.
8. Organizes and manages the use of metadata standards, access controls and authentication procedures.
9. Is aware of relevant quality assurance standards and makes a well considered choice whether to employ them or not.
10. Observes and adheres to all applicable legislation and regulations when making decisions about preservation, use and reuse of digital objects in collaboration with legal practitioners.

Based on the suggestions and comments received, it is worth mentioning that the members of the panel believed that digital curation workforce has multiple levels or tiers, is multi-disciplinary and includes workers from different sectors.

## 4 Conclusion

Digital curation is a new area of research and education where different professional communities end up facing similar issues and needing similar competences. The digital nature of the resources to be curated and preserved blurs the boundaries between the three traditional professions (librarian, archivist, museum curator). Once that a resource has become (or was born) digital, the challenges, the technologies and the competences needed for its curation and preservation to a large extent do not depend on the nature of the resource.

The twenty statements that have been defined and listed in the Delphi study include the operational and managerial competences of the digital curator. In conclusion, different identities of the information professionals can be evidenced, corresponding to different focus and missions of the disciplinary approaches, but the trend of convergence of the operational and some of the managerial competences can be noted.

Since a digital curator should be involved in the entire life-cycle of a resource, from its creation to its preservation for “future generations”, it appears that regardless of the origin and the intended fruition of a resource, large segments of its life cycle are more or less the same in each of the three traditional disciplines. Of course, given the different focus and mission of the three disciplines, the value adding and the access portions of the life cycle will remain different. However, the authors of this paper believe that further collaboration for the development of a common curriculum in digital curation can be built upon the many similarities over the entire life cycle.

## References

- American Association of Museums. Curators Committee, Code of ethics for curators (2006), [http://www.curcom.org/\\_pdf/code\\_ethics2009.pdf](http://www.curcom.org/_pdf/code_ethics2009.pdf)
- Beagrie, N.: Digital curation for science, digital libraries, and individuals. *International Journal of Digital Curation* 1(1) (2006)
- Bahr, T., et al.: Puzzling over digital preservation - identifying traditional and new skills needed for digital preservation. In: IFLA WLIC Conference Puerto Rico (2011), <http://conference.ifla.org/past/ifla77/217-bahren.pdf> (retrieved)
- Casarosa, V., et al.: Training for Digital Preservation in the context of the European project PLANETS. In: IFLA WLIC Conference Puerto Rico (2011), <http://conference.ifla.org/past/ifla77/217-casarosa-en.pdf> (retrieved)
- Cunningham, A.: Digital Curation/Digital Archiving: A View from the National Archives of Australia. *American Archivist* 71(2), 530–543 (2008)
- DCC. Digital Curation Centre, What is digital curation? (2003), <http://www.dcc.ac.uk/digital-curation/what-digital-curation> (retrieved)
- DIGCCURR Project, <http://ils.unc.edu/digccurr/digccurr-functions.html> (retrieved)
- Duranti, L.: The Long-Term Preservation of Accurate and Authentic Digital Data: The INTERPARES Project. *Data Science Journal* 4 (2006), [http://www.jstage.jst.go.jp/article/dsj/4/0/4\\_106/\\_article](http://www.jstage.jst.go.jp/article/dsj/4/0/4_106/_article) (retrieved)
- Gold, A.: Cyberinfrastructure, Data, and Libraries, Part 1 A Cyberinfrastructure Primer for Librarians. *DLib Magazine* 13(9/10) (2007), <http://www.dlib.org/dlib/september07/gold/09gold-pt1.html> (retrieved)

- Harvey, R.: *Digital curation: a how-to-do-it manual*. Neil Schuman Publishers, Inc., New York (2010)
- Harvey, R., Bastian, J.: *Out of the Classroom and into the Laboratory: Teaching Digital Curation Virtually and Experientially*. In: *IFLA WLIC Conference Puerto Rico (2011)*, <http://conference.ifla.org/past/ifla77/217-harvey-en.pdf> (retrieved)
- Hedstrom, M.: *Understanding electronic incunabula: a framework for research on electronic records*. *The American Archivist* 54(3), 334–354 (1991)
- IFLA Education and Training Section with Preservation and Conservation, Information Technology; co-sponsored by ICA Section for Archival Education and Training, Education for digital curation. In: *IFLA WLIC Conference Puerto Rico (2011)*, <http://conference.ifla.org/past/ifla77/education-and-training-section-with-preservation-and-conservation-information-techno.html> (retrieved)
- Library of Congress Digital Preservation Centre, <http://www.digitalpreservation.gov> (retrieved)
- Lord, P., Macdonald, A.: *Digital Data Curation Task Force. Report of the Task Force Strategy Discussion Day Tuesday, 26th November 2002 Centre Point, London WC1 (2003a)*, [http://www.jisc.ac.uk/uploaded\\_documents/CurationTaskForceFinal1.pdf](http://www.jisc.ac.uk/uploaded_documents/CurationTaskForceFinal1.pdf) (retrieved)
- Lee, C.A.: *Matrix of Digital Curation Knowledge and Competencies*, <http://www.ils.unc.eduldigccurrdigccurr-matrix.html> (retrieved)
- Lord, P., Macdonald, A.: *e-Science curation report: Data curation for e-science in the UK – an audit to establish requirements for future curation and provision. Report prepared for the JISC Support of Research Committee, JCSR (2003b)*, [http://www.jisc.ac.uk/uploaded\\_documents/e-ScienceReportFinal.pdf](http://www.jisc.ac.uk/uploaded_documents/e-ScienceReportFinal.pdf) (retrieved)
- Madrid, M.: *A Study of Digital Curator Competences: A survey of experts*. Master Thesis, DILL International Master Digital Library Learning, University of Parma (2011)
- Pomerantz, J., Oh, S., Wildemuth, B.M., Hank, C., Tibbo, H., Fox, E.A., et al.: *Comparing curricula for digital library and digital curation education*. In: *Proceedings of DigCCurr2009 International Symposium on Digital Curation*, Chapel Hill, NC, USA (2009)
- Ray, J.: *Managing the Digital World: The role of Digital Curation (2009)*, [http://www.ims.gov/pdf/JRay\\_Edinburgh.pdf](http://www.ims.gov/pdf/JRay_Edinburgh.pdf) (retrieved)
- Repanovici, A.: *Education and training for digital repositories manager*. In: *IFLA WLIC Conference Puerto Rico (2011)*, <http://conference.ifla.org/past/ifla77/217-repanovici-en.pdf> (retrieved)
- Robinson, M.: *The Institutional Repositories: Staff and Skills Set (2009)*, <http://www.sherpa.ac.uk/news/staffandskillssecondrevision.html> (retrieved)

# A Contribution for the Dissemination of Cultural Heritage Content to a Wider Public

Maristella Agosti<sup>1</sup>, Lucio Benfante<sup>1</sup>, and Nicola Orio<sup>2</sup>

<sup>1</sup> Department of Information Engineering – University of Padua  
Via Gradenigo, 6/a – 35131 Padua, Italy  
{agosti,benfante}@dei.unipd.it

<sup>2</sup> Department of Cultural Heritage – University of Padua  
Piazza Capitaniato, 7 – 35139 Padua, Italy  
orio@dei.unipd.it

**Abstract.** Digital resources are becoming an important tool for research in all the domains related to cultural heritage. Scholars have special requirements that need to be matched when developing digital library and digital archive systems that are to be used as tools to carry out scientific research. After having designed and developed a digital library application called IPSA as a system for researchers in illuminated manuscripts, we investigated how the digital library can be evaluated by non-domain users. Our goal was to highlight the overlaps and the differences in the user requirements between specialists, who use the digital archive to fulfill their research goal, and non-domain users, who interact with the digital library system because of a general interest about its content. The results have been used to re-engineer the digital library system and extend the functions of the digital library application in order to open up its use also to non specialists.

## 1 Motivations and Background

In past years, most systems able to manage specialised collections of cultural heritage documents have been envisaged and developed with one specific category of users in mind. In fact many systems have been created for managing collections to be used by researchers and scholars with specific requirements related to the research work carried out on the collections. More recently, many institutions have started to consider the possibility of opening up the use of those specialised collections and systems to other categories of users that may be interested in searching and navigating through cultural heritage resources. For instance, the DEBORA [9] and MonArch [12] projects involved different categories of users, from end-users to specialists in different domains, to develop collaborative access to cultural heritage content – Renaissance books and archeological sites, respectively.

The opening up of those collections and systems to new categories of users becomes a new challenge for the information communication technology specialists that have to address the generalization of systems previously designed for a

specific category of users. Cultural heritage applications can also be the starting motivation for the development of innovative tools to access multimedia content, for instance to annotate multimedia content [8] or to develop ad-hoc image processing techniques [11].

As an initial approach to this challenge, we considered the requirements that we gathered from a distinct category of domain specialised users to design a digital library application and tried to generalise them in order to re-design the system for its use by different categories of users. In particular, we focused on how the requirements gathered from domain specialized users (professional researchers) and used to envisage and design a digital library system can be considered to extend the functions of the system also to non-domain users.

A digital archive system of illuminated manuscripts was used as a case study for this investigation. The digital archive is called IPSA, which stands for *Imaginum Patavinae Scientiae Archivum* (archive of images of the Paduan science) [5]. IPSA was developed, from 2001 to 2005, with the main objective of being a scientific tool for the analysis of the role played by the Paduan school during the Middle Ages and the Renaissance in the spread of the new scientific method in different sciences, from medicine to astronomy and botany, through the study of illuminated manuscripts [10]. IPSA is a digital library system able to manage the description and the digital version of documents, dated especially from the late Middle Ages to the fifteenth century, that are of interest of botany, medicine, astronomy, and ancient astrology. The digital library software application [4] was envisaged and implemented by the University of Padua to study the medieval science and the scientific image in its tradition and evolution, in particular in relation to the studies conducted within the University. IPSA constitutes a valuable aid for people interested in the genesis of modern science also through the use of the visual transmission of knowledge.

The main goals of IPSA are: spreading the knowledge of ancient images both for their scientific and historical importance; creating links between images to relate them to different cultural areas of interest; and showing the importance that the University of Padua has played since the end of the Middle Ages in the spread and development of sciences and culture. Taking into account the main objectives of the project, it is clear that IPSA aimed at being used by professional researchers, i.e. scholars in history of medieval art specialised in history of illumination. It has to be noted that the text of an illuminated manuscript can be copied verbatim from older manuscripts, because the most relevant part of the illuminated manuscript is the iconographic part, so the text is accompanied by illustrations that can be copied from or inspired by older manuscripts, or taken directly from nature.

In the actual version, access to IPSA is given only to authorised users, because the manuscripts that are represented in the digital library application through metadata and digital version of pages are the property of different institutions spread throughout the world. In fact the archive includes 56 manuscripts belonging to some of the most important libraries in Europe and in the world. In order to grant access to a wider public, the research group responsible of maintaining the





Fig. 1. Examples of images from Egerton 2020, London, British Library

digital archive is in the process signing a license agreement with the institutions that own the manuscripts. A visible watermark has been added to each image, with a reference to the owner of the copyright.

A relevant example is the manuscript entitled *Liber Agregà de Serapion*, which is now property of the British Library (London, British Library, ms. Egerton 2020<sup>1</sup>) and the digital version of the pages of the manuscript is inserted and managed by IPSA thanks to an agreement between those two institutions. Therefore, an example of the iconographic content of IPSA can be found in the British Library catalogue of illuminated manuscripts; some examples are reported in Figure 1 which shows the digital representation of three pages of Egerton 2020.

### 1.1 Long-Term Objectives for a New Digital Library Application

Taking into account that IPSA is a combination of digitised images of manuscripts and related metadata information, and that its content can be of interest to a much larger group of users in respect with the one that was the initial target of the work, the IPSA application was selected to contribute to the design, development and evaluation of the innovative research environment that is under design and development in the context of CULTURA (Cultivating Understanding and Research through Adaptivity<sup>2</sup>), a EU funded STREP project<sup>3</sup>.

CULTURA aims at personalisation and community-aware adaptivity for digital humanities through the implementation of innovative adaptive services in an interactive environment. This goal is motivated by the desire to provide a fundamental change in the way digital cultural heritage is experienced, analysed and contributed to by communities of interested individuals. These communities typically comprise a diverse mixture of professional and apprentice researchers, informed users and interested members of the general public.

<sup>1</sup> The detailed record for Egerton 2020 is at the URL: <http://www.bl.uk/catalogues/illuminatedmanuscripts/record.asp?MSID=8320&CollID=28&NStart=2020>

<sup>2</sup> CULTURA Project Website, URL: <http://www.cultura-strep.eu/>

In line with the CULTURA objectives, an effort has been initiated to redesign the IPSA application to prepare an innovative digital library application able to face the challenge of supporting the different user groups of interest. In order to carry out an effective recollection of user requirements for a novel adaptive and interactive digital library system, we decided to carry out a first round of evaluations with students at the university level. The goal was to address the main problems that main arise while novel users were interactive with the digital archive. The presence of macroscopic issues, that can be due to a difference in the levels of expertise and motivation in using the system, may hide more subtle requirements that are more related to the development adaptive systems.

This paper reports the way this redesign has been addressed together with the initial results of user evaluation that are incorporated in the new version of the application now available<sup>3</sup>. This version will be the starting point for a second round of evaluation with a second cohort of students at the university level. Having addressed the most evident issues that arose during the initial evaluation, we are confident that a new recollection of user requirements will provide additional insights to achieve the aims of CULTURA project.

## 2 Requirements of Professional Users

As mentioned in the introduction, IPSA was developed as a tool for professional users, and, instead of limiting the requirements analysis to a number of interviews, the design approach was to create a research team where computer scientists and professional users (i.e. researchers in history of art specialized in history of illumination) collaborated together. Additional contributions from scholars in related disciplines, such as history of science, botany and astronomy, were integrated as well and formalized in a draft proposal that was presented and discussed with professional users. A similar approach was maintained during the development of the prototype system, because all the novel functions were directly tested by members of the research team.

The requirements for carrying out scientific research are in general more complex and articulated than the requirements of final users. Final users access an image digital archive to acquire information in a given field, researchers access the same archive to disclose knowledge and discover new relations between digital objects. IPSA was designed and developed taking into account the requirements of professional users in history of illumination. This means that IPSA is the outcome of the effort of producing an original and innovative system for a specialised group of professional users.

To understand the effort that has been recently started, in the context of the CULTURA project, aimed at re-designing IPSA to add new functions to the original ones to face the characteristics of interest of the new user groups of interest, it is necessary to know the inspiring requirements that pervade the

---

<sup>3</sup> Authorised users can use the new IPSA digital library application at the URL:

<http://ipsa.ipsa-project.org/>

original IPSA system. For this reason, those relevant characteristics are briefly presented in the rest of this section.

## 2.1 Disclosure of Relations between Images

It is of primary importance for professional users in history of illumination to discover whether illustrations have been copied from images of other manuscripts, or they have been merely inspired by previous works, or if they are directly inspired by nature. A major IPSA function thus regards the possibility of enriching the digital archive by highlighting explicit relations that have been discovered by a domain professional user. In particular, the user should be able to create *links* that connect one image to another that is related to it in some way. The analysis of user requirements on link management highlighted a number of advisable features that needed to be implemented, these are link authorship, link typology, and paths [6].

The analysis of user requirements also suggested the use of typed annotations connecting two manuscripts, two images, or even two parts of different images. These annotations, which have been called *linking annotations*, have a type that describes the kind of relation between the two objects and provides a semantic to the link. For this reason, we proposed a taxonomy for linking annotations [2] which is divided in two classes, including annotations that express either hierarchical or relatedness links. Annotations have been developed and integrated within the digital archive according to the formal model described in [1].

Researchers did not show any interest towards content-based image retrieval tools to ease their work. Apart from a possible lack of trust on automatic tools, they motivated this choice considering that general visual similarity is not particularly useful for their research work, and in most cases the relation between images is due to stylistic reasons, like the way small details are drawn. For this reason, content-based image retrieval was not considered as a relevant feature.

## 2.2 Dynamic Records and Intellectual Rights

Almost every digital archive dynamically changes over the years, mainly because of new acquisitions that increase the number of documents that are stored and managed by the archive system. This is also true for a digital archive of illuminated manuscripts, but there are other reasons that produce changes on the archive over time. These include the creation of records describing the documents and the images of an illuminated manuscript, which is part of the scientific research itself as for any collection of historical works. Some examples of changes to records are that new relations with other works have been discovered, or that the attribution to a given author became less certain.

Because creating a new record or modifying an existing one is part of the scientific work of researchers, the data management has to deal with intellectual rights. A researcher may prefer that some of the newly created records are not accessible by other users, at least until the results of his research have been

checked and afterwards published. This situation implies that users may decide which information can be shared with other users and which cannot.

This novel knowledge, which is due to original results, should be stored in the digital archive at a different level than the information that is based on a general consensus. To this end, the use of annotations, both classical textual annotations and linking annotations, can be a viable tool providing that a user may state which annotations can be shared with the community or with his research group, and which ones have to remain private. Such a mechanism allows researchers both to use the digital archive as an advanced research tool and to protect their intellectual rights. Moreover, linking annotations add a hypertextual structure to the archive, which is different for each user and reflects his personal knowledge in the field.

### 2.3 Presentation of Digital Images

A digital archive of illuminated manuscripts has the double role of preserving cultural heritage and giving access to users in a networked environment. As always happens in this situation, there is a trade-off between the high quality required for preservation and the small size needed for transfer over the network of the image files. Moreover, it has to be considered that research users should be able to perform comparisons among images belonging to different manuscripts that, in principle, may differ in their original size. According to professional users involved in the original design of IPSA, the number of images that should be presented on the computer screen varies from one to a maximum of six.

This last requirement implies that the image size, and hence its resolution, can dynamically vary depending on the context, because in principle a link can be created between any pair of images. The image files transfer load can be reduced through the use of thumbnails, at least for the first presentation of images. Thumbnails may also be a viable solution when the comparison between images is not part of the scientific research but can be used for dissemination to students or, if future releases of IPSA will be available on the Web, to casual users without controlled access.

Image acquisition is another important issue, because researchers should be able to analyze even small details of images. At the same time, researchers also need to see the image of the complete page of a manuscript, because it gives the context in which a particular object is presented. Moreover, many manuscripts have more than a single image for each page, with images surrounding or overlapping with text. For these reasons, it is advisable to carry out multiple acquisitions of the same page, with different resolutions depending on the level of detail needed for the analysis by researchers.

## 3 IPSA Digital Library System

The IPSA prototype implementing the requirements briefly recalled in Section 2 was developed. The close collaboration within a single team of researchers and

scholars of all the disciplines involved allowed us to create a closed loop for evaluation, testing and refinement of the different functions of the evolving prototype. Once the underlying database structure had been designed and developed, the organization of the user interface and the development of the novel functions highlighted by the user requirements were done incrementally, with scholars in history of art starting to populate the archive with the initial collection of images while the refinement of the software tools was taking place.

The IPSA digital archive system was made available on its stabilised form in 2005 and from then on it has been used for research purposes by history of art researchers. Over the years the collection of manuscripts and images has been incremented. Due to the launch of the CULTURA project, the IPSA digital archive system has been reconsidered for use by different categories of users, and a re-engineering of the system has taken place to bring the system up-to-date with the new technologies that in the meanwhile have been made available, while the underlying model of content management has been kept. Taking into consideration that users mainly focus their attention on the graphical interface when interacting with a digital resource, the system interface has been re-designed to bring it more in line with recent advancements.

The new IPSA user interface aimed at simplicity and easy user accessibility. The main layout is designed for optimal visualisation with a screen resolution of 1024x768 pixels and up, horizontally centered and filling the vertical space. The layout contains three zones: the top header, the main area and the bottom footer.

The main header is as thin as possible. It contains the main starting points to the IPSA functionalities: a small IPSA logo which links to the home page, the login/logout button, a structured multi-level menu and a form for searching the IPSA illustrations. When users are logged in, their name is shown in the header, linked to their profile for editing, if necessary. Near the search form there is a link to the advanced search function. The menu adapts itself following the user permissions, and it guides the user in the navigation, showing the most common functionalities in its first or at maximum its second level. The footer is designed for containing secondary menus and non critical information for the user. At present it contains the copyright information and the language selectors. The IPSA user interface is fully localized in Italian and English.

Most of the screen is occupied by the main area. The layout of this zone is strictly related to each functionality, and is designed and implemented following the user needs of usability. It is designed for showing the main information on the left, with a small sidebar on the right containing the links to the operations on the currently displayed object, and the related information.

A screenshot of the present Web interface presenting an image and related metadata of the IPSA collection is shown in Figure 2. The small image on the left is a tool that allows user to zoom in relevant details, which are presented in the central part of the screen. The image on the right is a link to an images that has been considered in a relevant relation with the image under analysis by a researcher, and it can be directly accessed for further comparison.

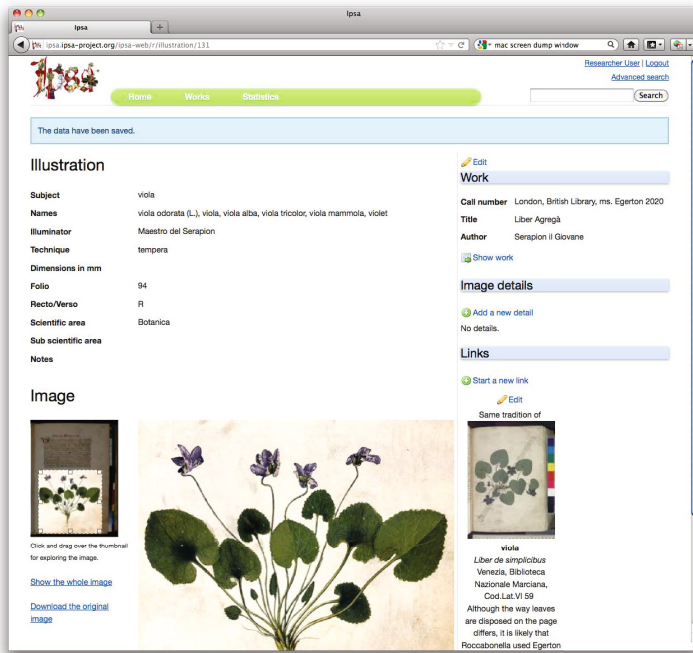


Fig. 2. Screenshot of the IPSA Web application

## 4 Accessing IPSA by Non-domain Users

Since 2005, IPSA has been used by professional users to carry out their studies on illuminated manuscripts. Starting from 2011, the new challenge is to investigate whether IPSA features can be of interest to non-domain users as well. So the goal is to use IPSA as a case study to compare the approach of different kinds of users to the same digital content. We conducted two subsequent evaluations: the initial one with different perspective users belonging to the class of non-domain professional research experts and to the class of the student community (in this case master students in archival science), the second one with the other groups that have to be taken into account in the context of the CULTURA environment and that were mentioned in Section 1.

The initial evaluation took place and completed with the main goal of highlighting possible overlaps between the requirements of domain professional users and the two considered groups. The main results were reported in 3 and for that are only summarised in the following. After having generalised the findings, the IPSA system has been re-engineered, as reported in Section 5. The second evaluation is still under development and possibly will give further useful insights.

*Interaction with Digital Systems.* As expected, computer skills play a central role in the way users interact with the system. This becomes particularly relevant in the case of specialists in other research areas who were asked to directly study multimedia content instead of bibliographic values. In contrast, master students had no problems interacting with multimedia content. Due to their habit of interacting with large multimedia collections, their requirements regarded search facilities, such as recommendations based on user-generated tags.

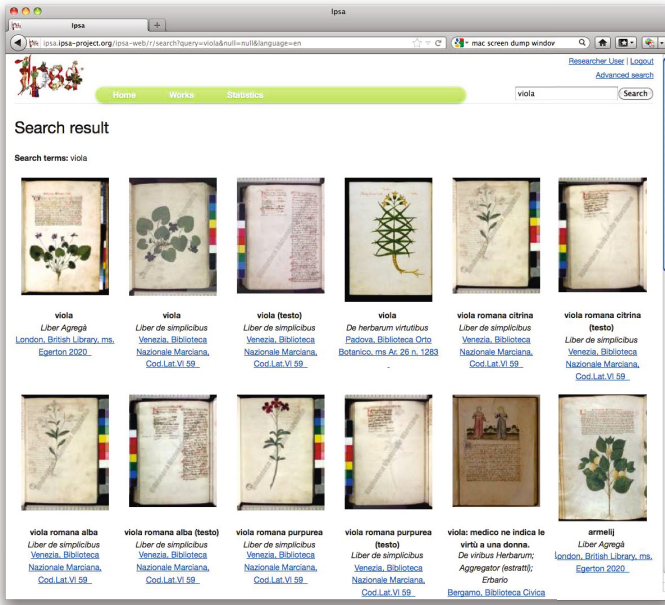
*Hypertextual Structure.* Probably because of its web interface, the archive was perceived as a hypertext and additional links towards external resources were considered an important improvement for IPSA. The possibility of using the digital archive as the starting point to retrieve other digital collections was considered highly relevant, maybe because users did not have a specific research interest towards the IPSA collection. The presence of navigation tools to browse the archive was considered important as well, because a lack of knowledge about archive content may prevent users from retrieving information when only direct search is made available. The presence of links between related images induces an hypertextual structure also to the collection of digital images and manuscripts. The exploitation of these links to improve research has already been proposed in [2], where an approach to mine the linking structure to discover novel relations is described. Moreover, the current evaluation highlighted that this feature will be useful also for non specialists as an alternative to direct search, in order to partially overcome the drawbacks of an imprecise knowledge of the domain.

*Textual Descriptors.* Although considered visually appealing, the digital images of the IPSA collections were not sufficient to raise interest when they are not paired by accompanying textual information. Analytic descriptions were suggested to improve user understanding of image characteristics, while additional bibliographic descriptors were suggested to make users aware of the cataloguing process. The approach to IPSA content depended on the particular field of interest of non-domain users.

## 5 Extending the Digital Library System

Although all non-domain users showed an interest towards IPSA digital archive, they highlighted a number of directions on how to improve interaction with the multimedia content. Part of these suggestions have already been implemented, in order to carry out a more effective evaluation with additional user groups. It is likely that the final outcomes of our evaluation will require a reengineering of the system.

First of all, a novel interface to display on screen a number of images has been developed. The interface now presents the images as a “wall” of thumbnails of the illustrations, with tools for incrementally loading additional slots of images at user request. A link can be followed from each image to its detailed description, where an image inspection tool is available to allow users to analyze its content in detail, and to follow a link to the manuscripts where it is contained.



**Fig. 3.** Results of a search using the term “viola” (violet)

This novel interface, which is shown in Figure 3 for the results of an image search using the term “viola” (violet), is used consistently each time a number of images has to be presented on screen at the same time. In the particular case of images belonging to a given work, images are shown in the order as they appear in the manuscripts (given the focus on digital images, most of the pages that contain only text are not part of the IPSA collection) and the interface allows the user to select the central image of the wall of images.

Another improvement regarded the rendering of individual images, which was designed for expert users that might be interested in very small details. The initial version of the interface allowed a personalized rendering, because the full version of the image was processed at server-side each time a request was made by a user. The actual version is based on a pre-rendering of the magnified image, using a predetermined fixed-screen maximum resolution. The user can still interact with the image by dragging the mouse over its thumbnail, but the image is rescaled at client-side, obtaining a more fluid navigation and improving user experience.

The interface to create links between images has been re-designed as well. Now it is possible to always have available the thumbnail of the image that is used to start a link, and to select the second image from a wall of thumbnails, which is the results of an image search. The starting page of a link creation is shown in Figure 4.



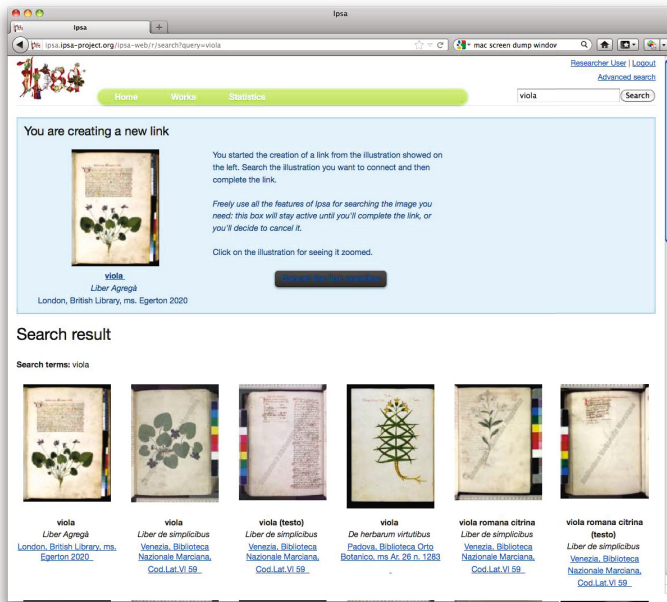


Fig. 4. Screenshot of the IPSA interface to create a link between two images

## 6 Conclusions and Future Developments

The IPSA digital library system was developed to help scholars carrying out their scientific research and is in the process of being extended to contribute to the dissemination of cultural heritage to a wider public. To this end, we carried out a first round of evaluation with non-domain users, in order to highlight a number of directions for improving the interaction with the digital collection of manuscripts. In this paper we describe how these requirements were translated in additional features of the IPSA digital archive.

We are organizing a second round of evaluation with other user groups, among the ones taken into account by the CULTURA project. It is likely that additional insights will be produced by the comments of new users, indicating that a digital library for cultural heritage should be a continuously growing system that has to evolve to adapt to new requirements in order to maintain its role of disseminating cultural heritage.

**Acknowledgements.** The authors would like to thank Giordana Mariani Canova and Chiara Ponchia of the Department of Cultural Heritage of the University of Padua and Nicola Ferro of the Department of Information Engineering for the useful discussions on aspects related to the reported effort.

The work reported has been partially supported by the CULTURA project, as part of the Seventh Framework Programme of the European Commission, Area “Digital Libraries and Digital Preservation” (ICT-2009.4.1), grant agreement no. 269973.

## References

1. Agosti, M., Ferro, N.: A Formal Model of Annotations of Digital Content. *ACM Transactions on Information Systems (TOIS)* 26(1), 3–57 (2008)
2. Agosti, M., Ferro, N., Orio, N.: Graph-based Automatic Suggestion of Relationships among Images of Illuminated Manuscripts. In: Haddad, H. (ed.) *Proc. of the ACM Symposium on Applied Computing*, pp. 1063–1067. ACM Press, New York (2006)
3. Agosti, M., Orio, N.: To envisage and design the transition from a digital archive system developed for domain experts to one for non-domain users. In: Nelson, M., Van de Sompel, H., Sølvsberg, I. (eds.) *Proc. 12th ACM/IEEE Joint Conference on Digital Libraries (JCDL 2012)*. ACM Press, New York (2012)
4. Agosti, M.: Digital libraries. In: Melucci, M., Baeza-Yates, R., Croft, W.B. (eds.) *Advanced Topics in Information Retrieval. The Information Retrieval Series*, vol. 33, pp. 1–26. Springer, Heidelberg (2011)
5. Agosti, M., Benfante, L., Orio, N.: IPISA: 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)
6. Agosti, M., Ferro, N., Orio, N.: Annotating Illuminated Manuscripts: an Effective Tool for Research and Education. In: Marilino, M., Sumner, T., Shipman III, F.M. (eds.) *Proc. 5th ACM/IEEE Joint Conf. on Digital Libraries*, pp. 121–130. ACM Press, New York (2005)
7. Agosti, M., Orio, N.: The CULTURA project: CULTivating understanding and research through adaptivity. In: Agosti, M., Esposito, F., Meghini, C., Orio, N. (eds.) *IRCDL 2011. CCIS*, vol. 249, pp. 111–114. Springer, Heidelberg (2011)
8. Borghesani, D., Grana, C., Cucchiara, R.: Surfing on artistic documents with visually assisted tagging. In: *Proc. of ACM Multimedia*, pp. 1343–1352 (2011)
9. Lebourgeois, F., Emptoz, H.: DEBORA: Digital Access to Books of the Renaissance. *Int. Jour. on Document Analysis and Recognition* 9(2-4), 193–221 (2007)
10. Mariani Canova, G.: La cultura universitaria padovana e la nascita del realismo nell’immagine botanica. *Atti e memorie dell’Accademia di Storia della Farmacia* XX(3), 198–212 (2002)
11. Ogier, J.M., Tombre, K.: Madonne: Document Image Analysis Techniques for Cultural Heritage Documents. In: *Proc. of the Int. Conf. on Digital Cultural Heritage*, pp. 107–114 (2006)
12. Stenzer, A., Woller, C., Freitag, B.: MonArch: Digital Archives for Cultural Heritage. In: *Proc. of ACM Int. Conf. on Information Integration and Web-Based Applications & Services*, pp. 144–151 (2011)

# Engaging the User: Elaboration and Execution of Trials with a Database of Illuminated Images

Chiara Ponchia

Department of Cultural Heritage – University of Padua  
Piazza Capitaniato, 7 – 35139 Padua – Italy  
[chiara.ponchia.1@studenti.unipd.it](mailto:chiara.ponchia.1@studenti.unipd.it)

**Abstract.** Currently one of the most important challenges for curators and providers of digital cultural heritage is to increase and enhance the engagement of users and communities with digital humanities collections. The reflections and efforts made to open up the IPSA database to new user categories is an ongoing process able to offer useful suggestions and contributions to this field of investigation. The considerations taken into account to elaborate the IPSA database trials engaging non-domain users are presented and the design of the trials is described.

## 1 Introduction

IPSA (*Imaginum Patavinae Scientiae Archivum*) is a digital archive of illuminated manuscripts which includes both astrological and botanical codices produced mainly in the Veneto region during the 14th and 15th centuries<sup>[1]</sup>. The database was created specifically for professional researchers of History of Art and History of Illumination to allow them to compare the illuminated images held in the database and verify the development of a new scientific mentality in the 14th century at the University of Padua and a new realistic way of painting closely associated with the new scientific studies<sup>[1]</sup>. Disclosing new relationships between images is one of the main purposes of art historical research, because it brings further knowledge on a specific artistic period, on a painter or an illuminator, on a work of art, and so on. According to this specific user requirement, in IPSA professional researchers are provided with tools that allow images to be linked and annotated, thus sharing knowledge in a collaborative environment<sup>[2]</sup>.

Due to involvement in the CULTURA project<sup>[3]</sup>, it was decided to open the database to other categories of users, such as non-domain professional researchers, the student community and the general public. This new task required the identification of the needs, wishes and preferences of these categories in order to define the required changes and improvements to IPSA<sup>[3]</sup>. User requirements were elicited in different ways. Firstly, thorough interviews with professional researchers were held, both with the domain professional researchers involved in the creation of IPSA from the very beginning, and with non-domain researchers

---

<sup>1</sup> <http://www.ipsa-project.org/>

<sup>2</sup> <http://www.cultura-strep.eu/>

expert in the field of History of Medieval Art but not acquainted with the IPSA collection and the History of Illumination in general. All the interviews were held on an individual basis.

The difficulty founded in eliciting professional researchers requirements highlighted the need to elaborate new ways to carry out user requirements elicitation, also for other user categories. For that reason, when it was decided to involve the student community, the elaboration of task-oriented trials was thought to be the best solution, as explained in Section 2. In Section 3 a description of the trials is presented, while in Section 4 can be found an overview of the first outcomes obtained.

This paper focusses on the elaboration and the description of the trials; information on the design and implementation of the computer system for user access can be found in [4].

## 2 The Design of the Trials: Preliminary Considerations

The IPSA database trials were developed specifically for two groups of students of the Faculty of Humanities of the University of Padua, Italy:

- Group 1 was formed by 24 first year Bachelor students in History of Artistic and Musical Heritage attending the first semester course on Foundations of Computer Science (Oct-Dec 2011);
- Group 2 was formed by 51 Master degree students in Communication Strategies, attending the first semester course on Design of Websites (Oct-Dec 2011).

The starting point for developing the trials were the outcomes reached in the first eight months of work of the CULTURA project. Most of this period was dedicated to drawing a profile and to identifying needs and wishes of the new categories of IPSA users. In this period we focussed on two user categories: professional researchers, namely established academics experienced in the general area covered by the resource, but not necessarily with the specific content of the resource, and the student community, particularly the university students. From these evaluation experiences, especially from the interviews with the professional researchers and the non-domain professional researchers, it was noted that the user should not only be presented the system and its functionalities, but also be provided with a task-oriented hands-on experience. Actually, interactions with professional researchers and non-domain professional researchers were two-fold: firstly, the interviewees were shown IPSA and its functionalities, and then they were asked about their impressions and their suggestions for improvement. When it came to this point, generally the interviewees showed a certain lack of imagination, concentrating on poorly relevant details, like the font or the colour of the text. We noticed that this happened because lack of motivation in using the system may reduce the effort put into learning how to interact with it, and this inevitably affects the quality of the interaction and of the reflection on the experience. When we decided to involve the student community, we hence knew

it was necessary to work out at least two tasks that would require the students to interact with the system in different ways. Actually, task-oriented experience is generally acknowledged as the best way to carry out systems evaluation because of several reasons. For instance, it allows for measuring effectiveness of systems such as how well the system enables a user to find an information needed or to answer a question, as can be seen in [5].

In line with these considerations, two tasks were developed to be carried out by students in two different trials. After in-depth discussions on the issue, it was decided that the tasks needed to be:

- *The same for all the students involved in the trials*, in order to obtain easily comparable data;
- *Specifically tailored to the groups of students chosen for the exercises*: considering that the first group of students had just begun their University career and that the second group of students is not attending a degree in History of Art, the development of some simple tasks that would not require a thorough knowledge of History of Art and History of Illumination was preferred;
- *One task related to the botanical codices collection, and the other related to the astrological codices collection*, in order to allow students to work with both the collections of the IPSA database.

In order to obtain further feedback, after each trial the students had to answer an evaluation questionnaire developed specifically by a team of psychologists from the University of Graz. The questionnaire aimed mainly at evaluating interaction with the system and user acceptance.

### 3 The Trials

For both trials, students were given a researcher account so they could enter the digital archive using the same tools as professional researchers. Hence they were able to set links between images and annotate them.

#### 3.1 First Trial

**Task 1.** This task is related to the botanical codices and proposes a guided comparison between the *Liber Agregà de Serapion* (London, British Library, ms. Egerton 2020) and the *Erbario Roccabonella* (Venice, Biblioteca Marciana, ms. Lat.VI.59). *Liber Agregà* is a remarkably important manuscript made in Padua at the end of the 14th century and commissioned by the prince of Padua, Francesco II da Carrara [6]. It shows the realistic representations of many different plants, with a short text explaining their therapeutic virtues [7]. *Erbario Roccabonella* is a Renaissance illuminated botanical codex written by the Medician Nicol Roccabonella in the 15th century [8]. It includes representations of hundreds of plants, some of which are also described in the *Liber Agregà*. Art historians understood that Roccabonella must have studied and partially copied images

from the *Liber Agregà* because of the similarities of many images in his book with those in the Paduan manuscript [9].

The students were required to verify this relation as well as find out which plants in the *Erbario Roccabonella* manuscript are copied from the *Liber Agregà* and which are not copied from this model but from other sources. So every student was assigned a page number belonging to the *Liber Agregà*. They had to check which plant was painted in the assigned page, and search whether the *Erbario Roccabonella* had an image of the same plant. Once they had found a second image, they had to analyse the two illuminations and decide whether the plant looked the same in both images, and if this was the case, set a link between the two illuminations, specifying the kind of link between them. They could choose between the following options:

- **Copied in:** the subject of the older image is quite faithfully re-proposed in the newer image;
- **Not related to:** the two illuminations show subjects belonging to different iconographic traditions;
- **Same tradition of:** the two illuminations show subjects belonging to the same iconographic tradition; this kind of relation is valid both for images markedly distant in time and for images closer in time;
- **Siblings:** the two illuminations were copied from the same model;
- **Similar to:** the two illuminations show some analogies, but it is not possible to further specify the kind of relation existing between them.

Afterwards, students could annotate the link, specifying the reason why they had chosen that link.

Clearly this is an “art historian task”, since it requires the comparison and analysis of two different images to discover the kind of link existing between them, so it was a good exercise for the Bachelor students in History of Artistic and Musical Heritage to become acquainted with the History of Art methodology.

Furthermore, this task points out one of the most valuable aspects of IPSA: the art historian is given the possibility to compare two different images and understand the relation between them simply by sitting at a computer. In the specific case of the *Liber Agregà* and the *Erbario Roccabonella*, the art historian need not travel to Venice and London to study these manuscripts. This is a great help for scholars, and perhaps not so immediately evident for young students who have no research experience: the task aims to show them the enormous potential of IPSA.

**Task 2.** This task is related to astrological manuscripts. The objective of this task was to have the students read the catalogue files and mine information from the database. Each student was given an astrological subject, namely:

- representations of constellations, i.e. *Ursa major* (Great bear);
- astrological signs, i.e. *Sagittarius*.

They were required to do a search by the subject assigned and analyse the first or the last five images in the results list. Then they were required to put them in chronological order. In this way, not only did they have to compare images, but they also had to read the catalogue files of five different manuscripts. Once the chronological order was set, they had the possibility of following the iconographic development of the subject.

### 3.2 Second Trial

Both tasks were planned to have a further development in the second trial (30 November for Group 1; 21 December for Group 2). On this occasion a short explanation of the IPSA functionalities was given before the trial in order to check whether the results changed and to what degree.

**Task 1.** Each student was re-assigned the same illuminated page from the first trial. This was the starting point for another kind of search: students were required to find out whether there were other images of the same plant in other botanical codices of the collection. Since plant names were not precisely codified in the Middle Ages, the students had to pay attention not only to the images, but also to the name variables. For example, the plant represented in f. 14v of the *Liber Agregà* is called *Stichados*, but in other botanical manuscripts held in the IPSA database the same plant is spelled *Sticados*, so the student working on this subject needed to search by every name variable, and to verify whether the plant was the same by carefully analysing the illuminations found. Once the student had verified that the represented plant was the same, he had to create a link between the two illuminations as in the previous trial.

**Task 2.** Each student was re-assigned the astrological subject of the previous trial. They had to make a search by this subject, and create links between the illuminations they found. So this time not only did they have to establish a chronological order, but they also had to analyse the kind of relation existing between all the images.

In the second trial the tasks were quite similar, but they presented different kinds of difficulty:

- In Task 1 the students had few manuscripts, a limited number of images, but a large number of etymological variables;
- In Task 2 the students had a larger number of manuscripts and images, but the illuminations were easier to find, since they only had two name variations or none at all.

## 4 First Outcomes

Since the analysis of the outputs of the trials is still an ongoing study, it is too early to have a comprehensive overview of the results obtained. Nonetheless, the trials seem to be a successful way of creating an useful and dynamic relation



Fig. 1. London, British Library. ms. Egerton 2020, f. 4r, *Citron*.

with users. In fact, in the first trial it was already possible to identify some necessary improvements that were immediately made to the database, in order to test them in the second trial. The most important improvement needed was to work out a more practical and faster way to present the illuminations to the users. For example, the *Erbario Roccabonella* holds hundreds of illuminations that required some minutes to be loaded. In the second trials the images were shown divided into smaller groups, and the loading was faster. This example clearly shows how the trials are bringing about a useful process of eliciting user requirements, immediately inserting changes into the database and subsequently evaluating the modifications made. The trials also prompted some preliminary considerations on how such a specialist collection is perceived by a non-specialist user. First of all, it was noted that people not used to working with images as historical documents focus their attention mainly on the text. When asked to find the images of the same plant in the *Liber Agregà* and in the *Erbario Roccabonella*, most of the students preferred to look for the images by searching with their names, rather than comparing the illuminations. This is not the best way to proceed, since in the Middle Ages names had a lot of variations, and a painted representation is normally more trustworthy. For example, the word citron means both lemon and cucumber, and some students set a link between the images of these two plants without noticing that they are far different, as can be seen comparing Figure 1 with Figure 2.





**Fig. 2.** London, British Library, ms. Egerton 2020, f. 162v, *Citron piolo*

This is a very important consideration, since it points out the need to develop a way to draw user attention to the illuminations, to make the user understand the real meaning and value of the IPSA collection and the way art historians work. Lastly, the trials brought about a reflection on what can be interesting for the student community. The students involved in the trials showed a particular attention to the Renaissance illuminations, probably because this is the best-known artistic period in Italy and the most studied in high school. This points out that users are mainly interested in something they can recognize or they mainly refer to something already known. So underlining the links between the IPSA collection and the history of Padua, of the Veneto region and of the Italian History of Art could be a good way to make the database more involving for non-specialist users. For example, underlining the connection between the botanic illuminations and the development of the scientific mentality in the University of Padua can make the database more interesting for students belonging to the same University.

## 5 Conclusions

This is a very important consideration, since it points out the need to develop a way to draw user attention to the illuminations, to make the user understand the real meaning and value of the IPSA collection and the way art historians work. Lastly, the trials brought about a reflection on what can be interesting for the student community. The students involved in the trials showed a particular

attention to the Renaissance illuminations, probably because this is the best-known artistic period in Italy and the most studied in high school. This points out that users are mainly interested in something they can recognize or they mainly refer to something already known. So underlining the links between the IPSA collection and the history of Padua, of the Veneto region and of the Italian History of Art could be a good way to make the database more involving for non-specialist users. For example, underlining the connection between the botanic illuminations and the development of the scientific mentality in the University of Padua can make the database more interesting for students belonging to the same University.

**Acknowledgements.** The author would like to thank Professor Maristella Agosti, Professor Giordana Mariani Canova and Dr. Nicola Orio of the University of Padua for the useful discussions on many aspects related to the reported investigation, and Eva Hillemann, Alexander Nussbaumer and Christina Steiner of the University of Graz for their helpful inputs and suggestions.

The work reported has been partially supported by the CULTURA project as part of the Seventh Framework Programme of the European Commission, Area “Digital Libraries and Digital Preservation” (ICT-2009.4.1), grant agreement no. 269973.

## 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., Ferro, F., Orio, N.: Annotating illuminated manuscripts: an effective tool for research and education. In: Marilino, M., Summer, T., Shipman, F. (eds.) Proc. 5th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL 2005), pp. 121–130. ACM Press, New York (2005)
3. Agosti, M., Mariani Canova, G., Orio, N., Ponchia, C.: Methods of personalising a collection of images using linking annotations. In: Proceedings of the First Workshop on Personalised Multilingual Hypertext Retrieval (PMHR 2011), pp. 10–17. ACM, New York (2011)
4. Agosti, M., Orio, N.: To envisage and design the transition from a digital archive system developed for domain experts to one for non-domain users. In: Proc. 12th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL 2012), pp. 11–14. ACM Press, New York (2012)
5. Hersh, W., Pentecost, J.: A Task-Oriented Approach to Information Retrieval Evaluation. *Journal of the American Society for Information Science* 47(1), 50–56 (1996)
6. Bettini, S.: Le miniature del Libro *Aggregà de Serapiom* nella cultura artistica del tardo Trecento. In: Grossato, L. (ed.) *Da Giotto a Mantegna*, exhibition catalogue (Padova, Palazzo della Ragione Giugno 9-Novembre 4, 1974), Milano, pp. 55–60 (1974)

7. Mariani Canova, G.: Serapion il Giovane, Liber Agregà. In: Baldissin Molli, G., Canova Mariani, G., Toniolo, F. (eds.) *La Miniatura a Padova dal Medioevo al Settecento*, Exhibition Catalogue (Padua, Palazzo della Ragione Palazzo del Monte; Rovigo, Accademia dei Concordi), Panini, Modena, March 21-June 27, vol. (54), pp. 154–157 (1999)
8. Marcon, S.: Nicolò Roccabonella. Liber de simplicibus. In: Fogliati, S., Dutto, D. (eds.) *Il giardino di Polifilo. Ricostruzione virtuale dalla Hypnerotomachia Poliphili di Francesco Colonna stampata a Venezia nel 1499 da Aldo Manuzio*, Ricci, Milano, pp. 113–115 (2002)
9. Mariani Canova, G.: La tradizione europea degli erbari miniati e la scuola veneta. In: *Di sana pianta. Erbari e taccuini di sanità*, exhibition catalogue (Praglia 1988), Panini, Modena, 21-28, pp. 25–26 (1988)

# Modeling Archives by Means of OAI-ORE

Nicola Ferro and Gianmaria Silvello

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

**Abstract.** Currently, archival practice is moving towards the definition of complex relationships between the resources of interest as well as the constitution of compound digital objects. To this end archives can take advantage of using the *Open Archives Initiative - Object Reuse and Exchange (OAI-ORE)* providing additional and flexible visualizations of archival resources.

In this paper we define a formal basis that provides a means for defining OAI-ORE instances which are consistent with the fundamental archival principles.

## 1 Motivation

Archives are composed of aggregations of interrelated material and their significance lies in their aggregate, or collective nature. Archivists work to preserve the *original order* of the documents within an archive – i.e. principle of provenance – because the context and the physical order in which the documents are held are as valuable as their content [3]. The principle of provenance leads archivists to evaluate records on the basis of the importance of the creator’s mandate and functions, and fosters the use of a hierarchical method for describing the archives. Although this practice is still vitally important for the archives, the archivists also need more powerful tools to capture the complexity of the reality of interest. Indeed, the reality of modern records creation is that documents may exist in “multiple contexts and have multiple and complex relationships that describe their significance and value” [9]. Furthermore, new archival trends encourage the adoption of “plural, provisional and interpretative perspective” [12] in the description of the archives.

The archival practice is thus experiencing a transformation process which promotes the definition of complex relationships between the resources of interest and the constitution of compound digital objects [9]. For similar reasons in the wider context of digital libraries we are experiencing a wide-ranging diffusion of the *Open Archives Initiative - Object Reuse and Exchange (OAI-ORE)* [1].

Archives as a meaningful part of the DL can take advantage of using the OAI-ORE [9]; indeed, a methodology for representing the archives in OAI-ORE would allow richer methods for modeling archival descriptions and can also provide additional and flexible visualizations of the documents that would not be restricted

---

<sup>1</sup> <http://www.openarchives.org/ore/>

to the “old linear view inspired by the paper tradition” [9]. At the same time, it is commonly agreed [12,18,9] that new approaches, such as the adoption of OAI-ORE model, should add to, but not undermine, the fundamental archival theory.

We can see an archive as a compound object composed by atoms of information which have to be identifiable and we need to define the granularity of this atoms. In this paper we adopt the NESTOR Model [1] to provide an alternative way to model archives allowing us to manipulate archival resources as atoms of information without losing their multileveled relationships. Therefore, in this work we lever on the *NEsted SeTs for Object hieRarchies (NESTOR)* Model [5] to:

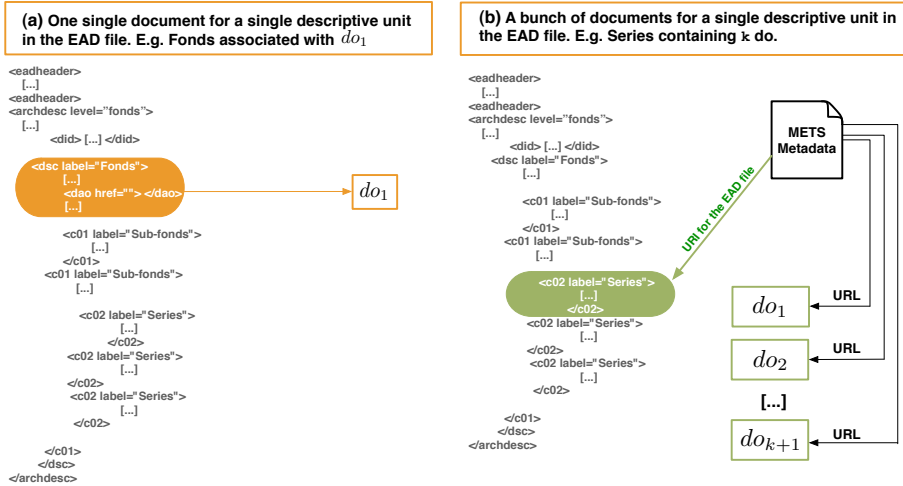
- define a formal basis that allows us to model an archive as an OAI-ORE instance while retaining its hierarchical structure;
- propose a methodology to map archival descriptions into OAI-ORE showing how it enables both the preservation of their original order and the definition of new types of relationships.

This paper is organized as follows: Section 2 presents a brief overview on archives and archival metadata, the NESTOR Model, and OAI-ORE. Section 3 describes the formal basis for modeling the archives as OAI-ORE instances while respecting the fundamental archival principles. Section 4 introduces a methodology which shows how we can represent a sample archive as an OAI-ORE instance. Lastly, Section 5 draws some final remarks.

## 2 Background

**Archives: Metadata and Digital Objects.** An archive is 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. Archives have to keep the context in which their records have been created and the network of relationships between them in order to preserve their informative content and provide understandable and useful information over time [6]. 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 then by sub-series and so on – see Figure 2a for an example; 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 enables the full informational power of the archival documents to be maintained. The archival documents are analyzed, organized, and recorded by means of *archival descriptions* [15] that have to reflect the peculiarities of the archive [3].

In a digital environment an archive and its components are described by using the metadata that have to be able to express and maintain such structure



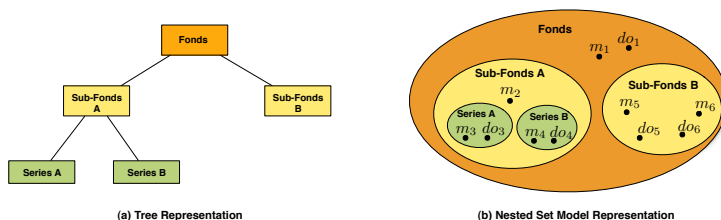
**Fig. 1.** A solution to link the EAD file with the described digital objects

and relationships. The standard metadata format for representing the hierarchical structure of the archive is the *Encoded Archival Description (EAD)*<sup>2</sup>, which reflects the archival structure and holds relations between documents in the archive [17]; an EAD file is an *eXtensible Markup Language (XML)* file with a deep hierarchical internal structure. In an EAD file the information about fonds, sub-fonds and series are mapped into several nested elements and the archival structure is maintained by a collection of nested `<cN>` tags (e.g. `<c02 label="Series">` in Figure 1). EAD describes an archive as a unique monolithic resource; indeed, it is not an aggregation of metadata each describing a single part of the archive, but a monolithic metadata where every sub-component describes a different division or document of the archive. In order to access a specific division of an archive described by EAD we may need to navigate the whole XML hierarchy; otherwise, it is also possible to define an ad-hoc solution, for instance using XPointers<sup>3</sup>, to provide direct access to frequently requested archival divisions encoded by EAD subcomponents.

Each `<cN>` tag of the EAD may contain a description of a digital object or a bunch of digital objects. These objects are usually reachable by means of an *Uniform Resource Identifier (URI)*; the link from EAD to a digital object or group of objects can be made at any level, but “*it should be made at the level where the object(s) is described or implied in EAD*” [13]. To this end EAD provides a `<dao>` tag which allows us to specify a URI to an external digital object which is part of the described material (see Figure 1a); furthermore, EAD also provides an `<extptr>` element to point to a digital object that is not part of the described materials [13]. By means of these tags we can link one external

<sup>2</sup> <http://www.loc.gov/ead/>

<sup>3</sup> <http://www.w3.org/XML/Linking/>



**Fig. 2.** The structure of a sample archive represented by: (a) a tree; (b) an Euler-Venn diagram

digital object to each archival division; if we need to link more than one digital object to a specific division we have to exploit third-party components – i.e. the so-called “digital wrappers”<sup>4</sup>; a relevant example is the *Metadata Encoding and Transmission Standard (METS)* metadata that is used as an in-between component for relating a bunch of digital objects to an EAD component [19,14] – see Figure 1b.

**The NESTOR Model.** The NESTOR Model relies on two set data models called *Nested Set Model* (NS-M) and *Inverse Nested Set Model* (INS-M) [1]. Both these set data models, formally defined in the context of axiomatic set theory [8], can be used to model an archive [5]. Indeed, we can represent the archival structure by means of a collection of nested sets where each set represents an archival division and contains the metadata describing the resources belonging to that division [4]. An extensive analysis of the NESTOR Model and its applications in the context of DL and archives can be found in [1]; in this paper we exploit the NS-M and thus we focus our presentation on this model.

The most intuitive way of understanding how the NS-M works is to see how a sample tree is mapped into an organization of nested sets based on the NS-M. An organization of sets in the NS-M is a collection of sets in which any pair of sets is either disjoint or one contains the other. In Figure 2 we can see how a sample tree representing an archive is mapped into an organization of nested sets based on the NS-M – for the moment please ignore the elements belonging to the sets. We can see that each node of the tree is mapped into a set, where child nodes become *proper subsets* of the set created from the parent node. Every set is subset of at least one set; the set corresponding to the tree root is the only set without any supersets and every set in the hierarchy is subset of the root set. The external nodes are sets with no subsets. The tree structure is maintained thanks to the nested organization and the relationships between the sets are expressed by the set inclusion order. Even the disjunction between two sets brings information; indeed, the disjunction of two sets means that these belong to two different branches of the same tree.

<sup>4</sup> Digital wrappers “are pieces of software for binding digital content files and their metadata together and for specifying the logical relationships among the content files” [14].

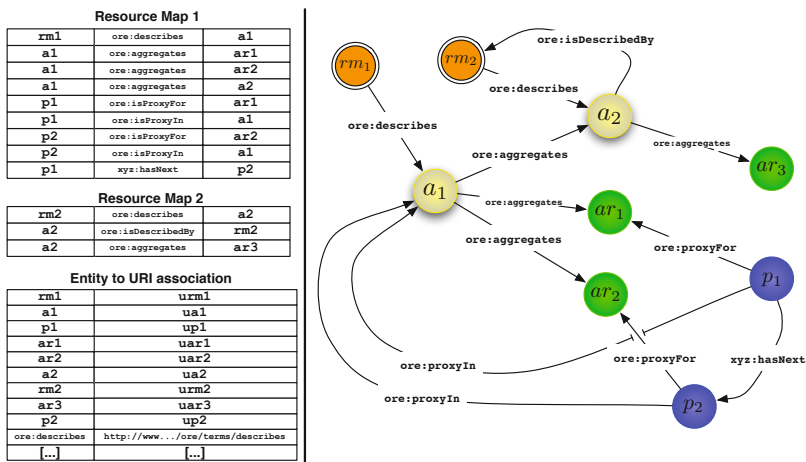


Fig. 3. An instance of the OAI-ORE data model represented by an RDF graph

In [4] a methodology is described for mapping an EAD file into the NESTOR Model which preserves the full informative power of the metadata. [4] shows that the EAD is mapped into a NS-C which retains the EAD structure and a collection of lightweight metadata – e.g. Dublin Core Application Profile<sup>5</sup> – which contains the content of archival descriptions. In this way, the NESTOR Model can be used as a model to describe an archive from scratch as well as a mapping component that allows us to manipulate and transform the EAD files while respecting archival principles [5].

**OAI-ORE.** The OAI-ORE defines a machine-readable and standard mechanism for defining aggregations of resources on the Web. By means of OAI-ORE we can identify a bunch of resources related to each other as a single entity enabling the access and exchange of them at an aggregation level of granularity. The OAI refers these aggregations as “*compound objects*”. Compound units are aggregations of distinct information units that, when combined, form a logical whole. Some examples [20] of these are a digitized book that is an aggregation of chapters, where each chapter is an aggregation of scanned pages, and a scholarly publication that is an aggregation of text and supporting materials such as datasets, software tools, and video recordings of an experiment; also the archives can be seen as aggregations of archival metadata describing archival objects which in turn can have a digital form.

The OAI-ORE data model is based on three main kinds of resources: *Aggregation*, *Aggregated Resources* and *Resource Map*. An Aggregation is defined as a resource representing a logical collection of other resources. An Aggregation is a logical construct and thus it has no representation; it is described by a Resource Map which can be seen as a materialization of the Aggregation. A Resource

<sup>5</sup> <http://www.dublincore.org/>



Map must describe a single Aggregation and must enumerate the constituent Aggregated Resources; a resource is an “Aggregated Resource” in an Aggregation only if it is asserted in a Resource Map. Each resource in the OAI-ORE data model is identified by a URI. The OAI-ORE data model is expressed by the *Resource Description Framework (RDF)*<sup>6</sup>, so its instances are expressed as RDF graphs as we can see in Figure 3. An RDF graph is defined by a set of triples  $(s, p, o)$  expressing the relationship defined by a predicate  $p$  between a subject  $s$  and an object  $o$ ;  $s$  and  $o$  may be a URI with an optional fragment identifier, a literal or a blank (having no separate form of identification). Properties  $p$  are URI references<sup>7</sup>. In Figure 3, we can see a set of subject-property-object triples represented as an RDF graph.

Although OAI-ORE is a relatively young specification, it has been becoming a standard reference in the context of digital libraries and its use is widespread in many systems and applications that deal with aggregations of digital objects. The use of OAI-ORE was adopted firstly for the management, access, and curation of scholarly publications and now it is spreading into the management and representation of scientific data [16] and of complex cultural objects [2].

### 3 A Formal Basis for Modeling Archives by Means of OAI-ORE

The aim of this work is to define a way to model archives by means of the OAI-ORE data model and the formal basis we propose provides a means to produce OAI-ORE instances which are consistent with the fundamental archival principles.

In order to explain how an archive can be properly modeled as an instance of the OAI-ORE data model we need to introduce several formal definitions. First-of-all, we present the definition of the NS-M which is based on the basic set-theoretical concept of “collection of subsets” [8].

**Definition 1.** *Let  $B$  be a set and let  $\mathcal{C}$  be a collection of subsets of  $B$ . Then  $\mathcal{C}$  is a **Nested Set Collection** if:*

$$B \in \mathcal{C}, \quad (3.1)$$

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

Thus, we define a *Nested Set Collection* (NS-C) as a collection of subsets where two conditions must hold. The first condition (3.1) states that set  $B$  which contains all the subsets of the collection must belong to the NS-C. 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 subset of the other one. This formulation of the NS-C follows the original definition of “nested sets representation” of a tree given by [10] and that we informally explained in the background section.

<sup>6</sup> <http://www.w3.org/RDF/>

<sup>7</sup> <http://www.w3.org/TR/rdf-concepts/>

Now, we can introduce a compact representation of the OAI-ORE data model in order to clarify the relationships between the entities and to manipulate them in a formal environment. We express OAI-ORE in terms of sets and functions in order to establish a direct connection with the NS-M by using the same mathematical formalism. We define with  $R$  the set of all the resources<sup>8</sup> we take into account, with  $U$  the sets of all possible URIs identifying the resources and with  $\eta : U \rightarrow R$  the bijective function<sup>9</sup> which associates a URI in  $U$  with one resource in  $R$ .

We indicate with  $UA \subset U = \{ua_1, \dots, ua_k, \dots, ua_n\}$  the set of URI identifying the Aggregations and with  $\eta_A : UA \rightarrow R$  the restriction of  $\eta$  ( $\eta|_A$ ) to  $UA$ ; the image of  $\eta_A$  is the set of Aggregations  $A \subset R = \{a_1, \dots, a_k, \dots, a_n\}$ . In the same way, we indicate with  $URM \subset U$  the set of URI identifying the Resource Maps and we define  $\eta_{RM} : URM \rightarrow R$  to be the restriction  $\eta|_{RM}$  where  $RM \subset R$  is the set of Resource Maps. Finally, we indicate with  $UAR \subset U$  the set of URI identifying the Aggregated Resources<sup>10</sup>. We define  $\eta_{AR} : UAR \rightarrow R$  to be the restriction  $\eta|_{AR}$  where  $AR \subset R$  is the set of Aggregated Resources. Every  $rm_i \in RM$  must describe one and only one  $a_j \in A$ , but  $a_j$  may be described by more than one Resource Map; thus, we indicate with  $\varphi_{RMA} : RM \rightarrow A$  a function which maps a Resource Map to the Aggregation it materializes. Every  $ar_i \in AR$  may be aggregated by more than one  $a_j \in A$ . An example of the use of these URIs is shown in the tables in Figure 3.

In Figure 3 we can see the set of triples constituting two Resource Maps ( $rm_1$  and  $rm_2$ ) materializing two Aggregations ( $a_1$  and  $a_2$ ). This triple states that the Resource Map  $rm_i$  identified by  $urm_i$  describes the Aggregation  $a_i$  identified by  $ua_i$ .

OAI-ORE comes with another two important features: *Proxy* and *Nested Aggregations*. A Proxy is a resource that indicates an Aggregated Resource in the context of a specific Aggregation; a Proxy is associated with an Aggregated Resource via an assertion in a Resource Map describing the Aggregation that is the context of the Proxy [11]. We indicate with  $UP \subset U = \{up_1, \dots, up_k, \dots, up_z\}$  the set of URI identifying the Proxies. We define  $\eta_P : UP \rightarrow R$  to be the restriction  $\eta|_P$  where  $P \subset R$  is the set of Proxies. Proxies allow us to define relationships between Aggregated Resources; in Figure 3 we can see two Proxies  $p_1$  and  $p_2$  defining an order of precedence between the Aggregated Resources  $ar_1$  and  $ar_2$  in the context of Aggregation  $A_1$ . We indicate with  $\varphi_{PAR} : P \rightarrow AR$  a function which maps a Proxy to the Aggregated Resource *for which* it is a Proxy and with  $\varphi_{PA} : P \rightarrow A$  a function which maps a Proxy to the Aggregation *in which* it is a Proxy.

<sup>8</sup> In this context a *resource* can be a metadata or a digital object.

<sup>9</sup> We choose to define  $\eta$  as bijective function to keep the problem as straightforward as possible; in a different context, a resource could be identified by more than one URI.

<sup>10</sup> Please note that the definition of the sets  $UA, URM, UAR$  is a mere convention to indicate URIs pointing to different kind of resources in OAI-ORE and they do not stand for different kind of URIs [20].

The *Nested Aggregations* feature enables the definition of Aggregations of Aggregations; this is consistent in the OAI-ORE data model because an Aggregation is a Resource which can also be seen as an Aggregated Resource of another Aggregation. Thanks to this feature, an order exists between Aggregations, call it  $\prec_a$ ; more formally: for all  $a_i, a_j \in A$  we say that  $a_i \prec_a a_j$  if and only if the Aggregation  $a_i$  is aggregated by  $a_j$ ; in Figure 3 we show two nested Aggregations  $a_1, a_2 \in A$  where  $a_2 \prec_a a_1$ . It is important to notice that  $\prec_a$  cannot define any orders between any OAI-ORE entities other than Aggregations; in fact, to define an order between Aggregated Resources we must use Proxies. Now, we can summarize the concept of *OAI-ORE Data Model* thanks to the next definition.

**Definition 2.** Let  $\mathcal{E} = \{A, R, AR, P, UA, UR, UAR, UP\}$  be the collection of OAI-ORE entity sets and  $\Phi = \{\eta_A, \eta_{RM}, \eta_{AR}, \eta_P, \varphi_{RMA}, \varphi_{PAR}, \varphi_{PA}\}$  be the set of OAI-ORE functions. We define  $\mathcal{O} = \langle \mathcal{E}, \Phi \rangle$  to be an OAI-ORE Data Model.

In order to model an archive by means of OAI-ORE we need a methodology to identify the archival resources and to express the relationships between them. We have seen that we can represent a tree by means of the NS-M and that an archive can be modeled by means of a tree as well as by a NS-C. Therefore, we can model an archive throughout OAI-ORE by starting from its representation in the NS-M. We need to define a mapping between a NS-C  $\mathcal{C}$  and an OAI-ORE model  $\mathcal{O} = \langle \mathcal{E}, \Phi \rangle$ ; in order to do this we have to take into account the two main entities of the NESTOR Model which are: the sets and the resources belonging to them.

The intuitive idea is that every set  $H \in \mathcal{C}$  becomes an Aggregation  $a_h \in A$  and consequently, every resource  $r_t \in R$  belonging to  $H$  becomes an aggregated resource  $ar_t \in AR$  aggregated by  $a_h$ . Furthermore, for every pair of sets  $\{H, K\} \in \mathcal{C} \mid H \subseteq K$  it is possible to create a pair of aggregations  $\{a_h, a_k\} \in A$  such that  $a_h \prec_a a_k$  where  $\prec_a$  is a binary relation between aggregations.

Every set in a collection of subsets can be mapped into an Aggregation in the OAI-ORE model; the inclusion order between the sets is maintained by the binary relation defined between the nested Aggregations of OAI-ORE. Then, by the means of the function  $\varphi_{RMA}$  a Resource Map is associated with each Aggregation. Every resource belonging to a set  $H$  in the NS-C is mapped into an Aggregated Resources belonging to the Aggregation mapped from  $H$ . Thus, we can map a NS-C into a correspondent OAI-ORE model being sure that the hierarchical dependencies are properly retained. This means that if we model an archive through a NS-C then we define an OAI-ORE instance of the archive which retains the original hierarchical structure of the archive.

## 4 How to Model an Archive as an OAI-ORE Instance

The presented formal basis guarantees that an archive modeled by means of the NS-M can be mapped into an instance of the OAI-ORE Data Model retaining the

Sets	Aggregations
fonds	$a_1$
subFondsA	$a_2$
subFondsB	$a_3$
seriesA	$a_4$
seriesB	$a_5$

**Table A**  
Mapping of sets into aggregations

Nested Sets		Nested Aggregations	
subFondsA	$\subset$ fonds	$a_2$	$\prec_a a_1$
subFondsB	$\subset$ fonds	$a_3$	$\prec_a a_1$
seriesA	$\subset$ subFondsA	$a_4$	$\prec_a a_2$
seriesB	$\subset$ subFondsA	$a_5$	$\prec_a a_2$

**Table B**  
Mapping of nested sets into nested aggregations

Elements	Aggregated Resources
$m_1$	$ar_a$
$do_1$	$ar_b$
$m_2$	$ar_c$
$m_3$	$ar_d$
[...]	[...]
$do_6$	$ar_m$

**Table C**  
Mapping of elements into aggregated resources

Elements and Sets	Aggregations and Aggregated Resources
$m_1 \in$ fonds	$a_1$ aggregates $ar_a$
$do_1 \in$ fonds	$a_1$ aggregates $ar_b$
$m_2 \in$ subFondsA	$a_2$ aggregates $ar_c$
$m_3 \in$ seriesA	$a_4$ aggregates $ar_d$
[...]	[...]
$do_6 \in$ subFondsB	$a_3$ aggregates $ar_m$

**Table D**  
Mapping of the elements belonging to sets into aggregated resources belonging to aggregations

Aggregated Resources	Proxies
$ar_a$	$p_a$
$ar_b$	$p_b$
$ar_d$	$p_d$
$ar_e$	$p_e$
[...]	[...]
$ar_m$	$p_m$

**Table E**  
Proxies for the aggregated resources

$p_a$ isMetadataOf $p_b$
$p_d$ isMetadataOf $p_e$
[...]
$p_l$ isMetadataOf $p_m$

**Table F**  
The use of property "isMetadataOf"

fundamental archival hierarchy. In this section we show how we can define different kinds of relationships between the resources; furthermore, we show how a proper use of Proxies can preserve the order between the resources within the same archival division. It is worthwhile to provide a concrete example of how this formal basis can be applied to a sample archive modeled by the NS-M; we describe the mapping methodology step-by-step with the help of some mapping tables.

Let us take into account the sample archive represented in Figure 2b; this archive is composed by five archival divisions – i.e. one fonds, two sub-fonds and two series – each containing metadata and digital objects. In NS-M these divisions are represented by means of five sets and the hierarchical relationships are retained by means of the inclusion dependencies between the sets. In “Table A” we can see the mapping of the sets into the OAI-ORE Aggregations and in “Table B” we can see how the inclusion dependencies are mapped into Nested Aggregations. These two mappings show us how to represent the structure of a sample archive into an instance of the OAI-ORE data model.

Each set in the NS-C contains several elements which are metadata or digital objects. For instance, the set “fonds” contains two elements: a metadata (i.e.  $m_1$ ) and an associated digital object (i.e.  $do_1$ ). The set “sub-fondsA” contains only a metadata (i.e.  $m_2$ ), the set “seriesA” contains a metadata (i.e.  $m_3$ ) and an associated digital object (i.e.  $do_3$ ), and so on and so forth. In “Table C” we can see how the elements are mapped into Aggregated Resources and in “Table D” we can see how the Aggregated Resources are associated with the correct Aggregations. We can see that an element belonging to a set – e.g.  $m_i \in H$  – is mapped into an Aggregated Resource – e.g.  $ar_i$  – aggregated by the Aggregation  $a_h$  which corresponds to the set  $H$ . “Table E” and “Table F” show how we can use Proxies to associate the metadata with the digital objects they describe. OAI-ORE allows us to define different kinds of relationships between the Aggregated Resources using the Proxies. For instance, in Table F we can see that two Proxies  $p_a$  and  $p_b$  associated to  $ar_a$  and  $ar_b$  respectively are related by the relationship “isMetadataOf”; thus, throughout  $p_a$  and  $p_b$  we can say that

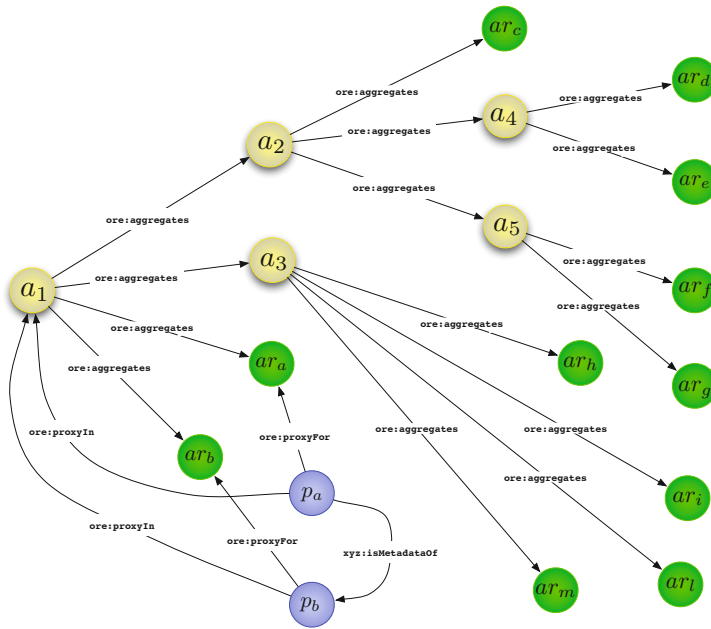


Fig. 4. An instance of OAI-ORE which models a sample archive

the Aggregated Resource  $ar_a$  is a metadata describing the digital object  $ar_b$ . In the same way we can define a linear order between the Aggregated Resources as we have shown in Figure 3 where we defined a “hasNext” relationship between the Proxies “ $p_1$ ” and “ $p_2$ ”. The relationships between the Aggregated Resources can reflect the order between the archival descriptions within a common archival division; in this way, we are sure that the OAI-ORE representation of the archive respects the original order principle. We can see that within this methodology it is quite simple to extend the range of the relationships connecting the Aggregated Resources and to define in this way new semantic associations between the archival resources.

In Figure 4 we can see the RDF graph representing the OAI-ORE instance of the sample archive in Figure 2b. In this figure we represent the Aggregations, the Aggregated Resources and the Proxies associated to  $a_1$ ; for space reasons we have omitted showing the other Proxies and the Resource Maps. This methodology makes it possible to model and describe the archives from scratch by means of OAI-ORE while allowing archivists to easily express relationships between archival metadata and digital objects. Archival principles are preserved and still have primary importance for understanding archival resources; at the same time, OAI-ORE offers the possibility of defining new relationships between the resources enabling the definition of new services over the archives. Moreover, this methodology provides a means to define archival compound objects that can be shared with the systems which already employ OAI-ORE and related technologies.

On the other hand, this methodology and the described formal basis guarantee the backward compatibility with other archival descriptive standards; for instance, a methodology to map the archival descriptions modeled by OAI-ORE into EAD can be easily defined. Indeed, we know how to map EAD into a NS-C and a NS-C into an instance of the OAI-ORE data model. In the same way, we can map the archival descriptions modeled by OAI-ORE into an EAD file by reversing the presented methodology<sup>[1]</sup>. In this context, the NESTOR Model can act as an interoperability layer between EAD and OAI-ORE and guarantee the possibility of going from one model to the other.

## 5 Final Remarks

In this paper we present a formal basis and a methodology to model an archive by means of the OAI-ORE data model consistent with the fundamental archival principles. OAI-ORE is widely-employed in the context of Digital Libraries but is still not completely exploited within archives; the formal basis reported in this paper can settle the ground for further investigations about the adoption of OAI-ORE in the archival context. This research direction can bring into archival practice the expressive power of OAI-ORE to allow for a multitude of non-linear relationships, providing richer and more powerful access and descriptions.

Furthermore, the use of OAI-ORE is increasing in several systems and digital library federations such as Europeana<sup>[2]</sup> the aim of which is to collect and make available resources from a wide spectrum of cultural institutions including the archives. A further step toward this direction will be to investigate how the NESTOR Model may allow different ways of modeling archival resources easing the integration of these resources with the Europeana Data Model (EDM). It will be interesting to consider the proposed methodology under the lens of other approaches trying to map archival resources into EDM <sup>[7]</sup>.

**Acknowledgments.** CULTURA<sup>[3]</sup> (Grant agreement no. 269973) and the PROMISE network of excellence<sup>[4]</sup> (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)

<sup>11</sup> Please note that the backward compatibility can be limited by the fact that the EAD expressive power is inferior to that of OAI-ORE.

<sup>12</sup> <http://www.europeana.eu/>

<sup>13</sup> <http://www.cultura-strep.eu/>

<sup>14</sup> <http://www.promise-noe.eu/>

2. Doerr, M., Gradmann, S., Henicke, S., Isaac, A., Meghini, C., Van de Sompel, H.: The Europeana Data Model (EDM). In: IFLA 2011: World Library and Information Congress: 76th IFLA General Conference and Assembly, Gothenburg, Sweden (2010)
3. Duranti, L.: Diplomatics: New Uses for an Old Science. In: American Archivists and Association of Canadian Archivists. Association with Scarecrow Press, Lanham (1998)
4. 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)
5. Ferro, N., Silvello, G.: The NESTOR Framework: How to Handle Hierarchical Data Structures. In: Agosti, M., Borbinha, J., Kapidakis, S., Papatheodorou, C., Tsakonas, G. (eds.) ECDL 2009. LNCS, vol. 5714, pp. 215–226. Springer, Heidelberg (2009)
6. 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)
7. Henicke, S., de Boer, V., Isaac, A., Olensky, M., Wielemaker, J.: Conversion of EAD into EDM Linked Data. In: Proceedings of the First International Workshop on Semantic Digital Archives, SDA 2011 (2011)
8. Jech, T.: Set Theory. Springer, Berlin (2003)
9. Kaplan, D., Sauer, A., Wilczek, E.: Archival Description in OAI-ORE. In: OR 2010: The 5th Int. Conf. on Open Repositories (2010)
10. Knuth, D.E.: The Art of Computer Programming, 3rd edn., vol. 1. Addison Wesley, Reading (1997)
11. Lagoze, C., Van de Sompel, H., Johnston, P., Nelson, M., Sanderson, R., Warner, S.: ORE Specification-Abstract Data Model. Technical report, OAI (2008)
12. Light, M., Hyry, T.: Colophons and Annotations: New Directions for the Finding Aids. *The American Archivists* 65, 216–230 (2002)
13. Metadata Standards Subcommittee OAC Working Group. OAC Best Practice Guidelines for EAD. Version 2.0. Technical report, 61 pages (April 2005)
14. University of California. California Digital Library. CDL Guidelines for Digital Objects. Version 2.0. Technical report, 34 pages (January 2011)
15. Pearce-Moses, R.: Glossary of Archival and Records Terminology. Society of American Archivists (2005)
16. Pepe, A., Mayernik, M., Borgman, C., Van de Sompel, H.: From Artifacts to Aggregations: Modeling Scientific Life Cycles on the Semantic Web. *JASIST (J. of the American Society for Inf. Science and Technology)* 61, 567–582 (2010)
17. Pitti, D.V.: Encoded Archival Description. An Introduction and Overview. *D-Lib Magazine* 5(11) (1999)
18. Ross, S.: Digital Preservation, Archival Science and Methodological Foundations for Digital Libraries. In: Keynote Address at the 11th European Conf. on Digital Libraries (ECDL), Budapest (2007)
19. Sugimoto, G., van Dongen, W.: Archival Digital Object Ingestion into Europeana (ESE-EAD Harmonization). Technical report, Europeana v1.0 (August 2009)
20. Van de Sompel, H., Lagoze, C.: Interoperability for the Discovery, Use, and Re-Use of Units of Scholarly Communication. *CT Watch Quarterly* 3(3) (August 2007)

# Reflecting on the Europeana Data Model

Silvio Peroni<sup>1</sup>, Francesca Tomasi<sup>2</sup>, and Fabio Vitali<sup>1</sup>

<sup>1</sup> Department of Computer Science, University of Bologna, Italy  
{essepuntato, fabio}@cs.unibo.it

<sup>2</sup> Department of Classical Philology and Italian Studies, University of Bologna, Italy  
francesca.tomasi@unibo.it

**Abstract.** We describe some issues arising while using Europeana, and analyze some features of the Europeana Data Model (EDM), starting from the rationale of the project. Some aspects of the theoretical model, derived mostly from the mapping between the provided Cultural Heritage Object (CHO) and the EDM, prevent useful results in users' queries. The concept of media type, the multi-layer description and the relation between roles and values are some issues about which we reflected. The aim of Europeana to make records available as Linked Open Data on the Web could require moreover a redefinition of the implementation techniques.

**Keywords:** EDM, Linked Data, RDF, DC, CIDOC CRM, FRBR.

## 1 Introduction

Europeana<sup>1</sup> is the European Digital Library, a distributed access point to Europe's multilingual cultural heritage in a digital form. The main aim of the project is to collect metadata from a large number of providers, mainly cultural institutions, across Europe, and to enable search and discovery of cultural items described therein.

The metadata aggregation is based on a mapping between the providers' data description and the Europeana model. The Europeana v1.0 project<sup>2</sup> [1] proposes the Europeana Data Model (EDM)<sup>3</sup> that defines a set of classes and properties to be used in Europeana for describing cultural objects. The EDM [2] is a clear improvement over the earlier data model, the Europeana Semantic Elements (ESE) [3]. ESE was meant to express the providers' datasets using the Dublin Core (DC) standard as its "lowest common denominator", while EDM, based on the DCMI Metadata Terms and a number of more advanced metadata models, adopts "an open, cross-domain Semantic Web-based framework" leaving each provider free to use their preferred metadata standard with regard to the element sets and the vocabularies of values [4].

---

<sup>1</sup> User access: <http://europeana.eu/portal>

<sup>2</sup> <http://pro.europeana.eu/web/europeana-v1.0>.

In: Europeana Professional: <http://pro.europeana.eu>

<sup>3</sup> The family of technical documents about EDM (in particular Definition, Primer, Guidelines) could be found at:

<http://pro.europeana.eu/web/guest/edm-documentation>



Given this, an extreme heterogeneity can be observed in the descriptions of cultural objects, a situation determined by the differences in existing collections that Europeana involves (museums, archives, audiovisual collections and libraries), and by the different kinds of objects described in the cultural repositories that were harvested (i.e. manuscripts, documents, paintings, art and architecture objects, photos, videos, etc.). Also, the variety of descriptive situations depends both on the reference models for the metadata element sets (not only DC, SKOS, and the CIDOC CRM, but also e.g., EAD for archives, METS for digital libraries, TEI for literary texts<sup>4</sup>) and from the vocabularies of values (authority lists, thesauri or controlled vocabularies) used by data providers (such as Geonames, Art&Architecture Thesaurus, Iconclass, WordNet, Dewey Decimal Classification, DBPedia, etc.)<sup>5</sup>.

Furthermore, even if the EDM has been introduced, a majority of records in Europeana still seems to follow the old ESE data model and the users' query interface is based on just a few categories of Dublin Core. The main problems in using the current release of Europeana derives in part from the above-mentioned issues: the original object descriptions are often lost and the ESE, the first proposed model, was not sufficient to describe the complexity of many cultural objects, as it was based on just a few DC categories. Providers that had complex and structured descriptions have had to force them into to a much simpler model and providers that created descriptions in a model not compliant with ESE have lost data or have forced them into incorrect properties, leading to aggregations of imprecise information. Consequently, many user queries cannot be satisfied completely. Probably the complete integration of EDM in Europeana records will help towards completeness and correctness of the contained information. But some other improvements could solve many situations that not even EDM takes into complete account.

The whole point of this effort clearly is to allow users to enact better queries and to obtain better results through the use of a sophisticated and increasingly ontological metadata model on which to let the search engine work. Europeana is now working on improving the quality of the responses to users' queries, but much is still to be done.

For example at the moment there is no possibility to perform a multilingual search, although the vocabulary alignment is a problem being studied currently by the Europeana group, since one of the aims of subproject EuropeanaConnect<sup>6</sup> is in fact to solve this gap.

---

<sup>4</sup> DC, SKOS and CIDOC CRM will be discussed in section 2. As regards to the other schemas we just mention here the official sites: EAD (Encoded Archival Description), <http://www.loc.gov/ead/>; METS (Metadata Encoding & Transmission Standard), <http://www.loc.gov/standards/mets/>; TEI (Text Encoding Initiative), <http://www.tei-c.org>

<sup>5</sup> A comprehensive list of metadata and vocabularies published as Linked Data sets could be found in: W3C Incubator Group Report 25 October 2011. Library Linked Data Incubator Group: Datasets, Value Vocabularies, and Metadata Element Sets at:

<http://www.w3.org/2005/Incubator/11d/XGR-11d-vocabdataset-20111025/>

<sup>6</sup> <http://www.europeanaconnect.eu/>

Some other features are planned to be implemented using Semantic Web technologies<sup>7</sup>. For example, Europeana lacks a semantic network for the subjects, that could help users in finding records by specifying either the exact word or any of its synonyms, hyponyms, hypernyms, related terms, etc. WordNet, for example, has been published as a RDF vocabulary<sup>8</sup> and could be used after a vocabulary alignment. Another issue that Semantic Web technologies could solve is to enrich the existing metadata sets, contextualising objects and using authoritative sources, including controlled vocabularies, concretely shared in an integrated environment. The “related content”, presented in each object description in Europeana, is now mostly limited to other objects from the same data provider. More complex and structured relationships between what Europeana calls the provided “Cultural Heritage Object (CHO)” [2] and other pertinent resources, internal (other data providers) or external (other digital libraries), could be solved in a Linked Data perspective.

Although some features are now being studied as an experimental increment to the existing feature set, some discrepancies could be noticed between the EDM and the objects described, for which we do not know of any on-going work.

In this paper we analyze the query results of the current implementation and propose some reflections on the EDM in this phase of implementation. We focus here on three aspects, derived from the adoption of DC as the end-user property set: the “media type” concept, the multi-layer description of subjects and the connection between roles and values (Section 3). Even if the original adoption of DC is the main reason for the limitations in user queries, a redefinition of some classes and properties of EDM could solve some issues. Finally, given that one of the main aims of Europeana project is to expose metadata as Linked Open Data we verified that the current implementation is fairly good for the data that has been converted, but is still lacking in handling properly error cases. For this reason we analysed the “data.europeana.eu” Linked Data responses and the issues with empty resources (Section 4).

## 2 Related Works

In its specifications (EDM), Europeana mentions vocabularies, models and ontologies adopted in its data model [2]. The aim is to represent metadata for cultural heritage objects and to give access to digital representations of these objects. The EDM moves in the context of data aggregation, where objects can be complex, and several data providers may entertain different views on them.

The basis of the metadata description is RDF statements. An XML Schema has been defined for describing classes and properties. Some classes and properties are re-used from public models: DC, DCterms, SKOS, OAI-ORE, CIDOC-CRM, FRBR. Some other classes and properties are specifically created for the EDM and are mostly equivalent to predicates used in the most common ontologies.

---

<sup>7</sup> Library Linked Data Incubator Group wiki. Use Case Europeana:

[http://www.w3.org/2005/Incubator/1ld/wiki/Use\\_Case\\_Europeana](http://www.w3.org/2005/Incubator/1ld/wiki/Use_Case_Europeana)

<sup>8</sup> Wordnet 3.0 in RDF: <http://semanticweb.cs.vu.nl/lod/wm30/>

In addition to classes and properties, Europeana is defining also controlled vocabularies useful for CHO interoperability (such as AAT, DDC, DBpedia, Iconclass). The main aim of Europeana is to work on Linked Data both exposing record sets [5] and using Linked Data resources [6] in order to augment Europeana content.

In the following sub-sections we introduce the main external models adopted by EDM, highlighting which part of them are effectively used.

## 2.1 Dublin Core

The current versions of the Dublin Core (DC) Metadata Elements [7] and of the DC Metadata Terms [8] are the most widely used vocabularies for describing and cataloguing resources. These vocabularies have become particularly important and relevant for sharing metadata about documents among different repositories and digital libraries. While very useful for creating basic metadata that permit bibliographic resource descriptions (e.g., *creator*, *contributor*, *publisher*, *format*), the main limitation of DC is a consequence of the generic nature of its terms. In fact, its classes are organised without a strong hierarchical structure and their properties often lack in clear domain/range definitions. EDM makes extensive use of DC Elements and DC Terms entities, such as the properties *dc:subject*, *dc:contributor*, *dcterms:created* and *dcterms:alternative*.

## 2.2 SKOS

Data providers, publishers and aggregators, such as Europeana, need to classify the resources they publish according to discipline-specific thesauri and classification schemes. The Simple Knowledge Organization System (SKOS) [9] is an RDFS ontology to support the use of knowledge organization systems (KOS). A large number of well-known thesauri and classification systems have started to convert their specifications into SKOS documents, such as the “Nuovo Soggettario” of the National Central Library in Florence<sup>9</sup>. This makes SKOS the de facto standard for encoding controlled vocabularies for the Semantic Web. EDM uses the main SKOS class, i.e. *skos:Concept*, defined as a particular kind of *edm:NonInformationResource* for introducing an idea or notion.

## 2.3 FRBR

The Functional Requirements for Bibliographic Record (FRBR) [10] is a general model for describing bibliographic entities, such as documents and artistic works. FRBR specifies four basic concepts – work, expression, manifestation and item – used for characterising a particular bibliographic entity from different perspectives. In particular:

- A *work* is the *abstract essence* of an intellectual or artistic creation, e.g. the ideas in Shakespeare’s head concerning the *Macbeth*. A work is realised in one or more expressions;

---

<sup>9</sup> <http://thes.bncf.firenze.sbn.it/>

- An *expression* is the *content* of a particular work at a specific point in time, e.g. the final text of the *Macbeth* written by Shakespeare or its Italian translation made by Andrea Maffei. An expression is embodied in one or more manifestations;
- A *manifestation* is the particular *format* in which an expression is stored, such as a printed object or a digital document, e.g. the 2005 edition of *Macbeth* published by Penguin Books or its HTML Italian version published by. A manifestation is exemplified in one or more item;
- An *item* is a particular *physical or electronic copy* of the *Macbeth* that a person can own, e.g. the printed version of that book you have in your bookcase or the specific HTML document of its Italian version you are visualising in your browser.

Overall, EDM makes only limited use of FRBR concepts, although it declares explicitly their adoption. The only specific references to FRBR are:

- the class *edm:InformationResource*, defined as union of *FRBR Work*, *FRBR Expression* and *FRBR Manifestation* that results in collapsing completely the hierarchy of the FRBR model;
- the classes *edm:Event* and *edm:Place*, defined as equivalent to *FRBR Event* and *FRBR Place* respectively.

## 2.4 ORE

The Open Reuse and Exchange specification (ORE specification) [11] is a standard defined by the Open Archives Initiative for describing and exchanging aggregations of Web resources. Europeana uses two terms from this model:

- *Aggregation*, i.e. a particular resource that aggregates, either logically or physically, other resources;
- *Proxy*, used to refer to a specific aggregated resource in a context of a particular aggregation.

EDM uses all the main classes and properties of the ORE specification. For instance, it allows one to describe a “cultural heritage object” (i.e., *edm:providedCHO*) and its digital representations (i.e., *edm:WebResource*) as a particular aggregation (*ore:Aggregation*) representing the results of the activity of a particular data provider (i.e., *edm:Agent*).

## 2.5 CIDOC CRM

CIDOC Conceptual Reference Model (CRM) [12] is an ISO standard defining a model for describing and sharing cultural heritage information. It provides entity definitions and a formal multi-level structure to link physical objects to related events and

agents (i.e., people and organisations), so as to represent a mediator between different sources of cultural heritage information (e.g., museums, libraries and archives).

EDM aligns some of its classes and properties to the CIDOC CRM specification, for instance the class *edm:Event* as equivalent to *E4 Period*, the class *edm:InformationResource* as subclass of *E73 Information Object*, and the property *edm:wasPresentAt* as equivalent to *P12 occurred in the presence of (was present at)*<sup>10</sup>.

### 3 Issues Arising While Using Europeana

The EDM rationale is based on some principles [14]:

1. distinction between “provided objects” (painting, book, movie, etc.) and their digital representations;
2. distinction between objects and metadata records describing an object;
3. allow for multiple records for a same object, containing potentially contradictory statements about it;
4. support for objects that are composed of other objects;
5. compatibility with different levels of description;
6. standard metadata format that can be specialized;
7. support for contextual resources, including concepts from controlled vocabularies.

Although the Europeana core classes stress the difference between the provided object (*edm:ProvidedCHO*), i.e., the “real object”, and its digital representation (*edm:WebResource*), i.e., its Web resource, sometimes this difference is not evident at all in the aggregated metadata exposed to the final user, generating confusion. Sometimes the description seems to be addressed to the electronic version, some other to the original work, without a clear distinction (see 3.1). Additionally, the Europeana contextual classes, which are designed to answer to the four fundamental questions of the *who* (the Agent), the *where* (the Place), the *when* (the Time), and the *what* (the Concept) of the object, sometimes are not correctly represented in the metadata description because of a potential multi-layer representation issue derived from the stratification of object and subject (see 3.2). Therefore, the rationale of the EDM appears not always respected and the application of the listed properties is not totally functional (see 3.3). Here we describe some examples of these limits.

#### 3.1 The Media Type

The first and most evident source of confusion is the concept of media type found as the topmost choice in the filter section after every query (*edm:type* = text, image, audio, video). The media type is sometimes congruent with the type of the provided

---

<sup>10</sup> While in the Europeana Data Model the CIDOC CRM property *was present at* has the identifier *P121*, in [12] that property is defined as *P12 occurred in presence of (was present at)*.

CHO, and sometimes to its web resource. Yet, the rationale of Europeana is to distinguish between the description of the CHO and its digital representation. If we search for any object called “illuminated manuscript” we receive different answers: sometimes it has type “text”, sometimes type “image” some other times “physical object”. In general, resources classified as IMAGE are in fact image files (regardless of whether they represent pictures or physical objects such as buildings or statues or manuscripts), but resources classified as TEXT are sometimes texts, and sometimes images of texts (e.g., photographs of old volumes or manuscripts). One may wonder which would be more useful for searches, i.e., for the media type to refer to the web resource, providing a description of a computer-specific object, or to the cultural heritage object, which is what the user would be actually searching for. In both cases, it would be quite important to provide subtypes: they could be either subtypes of the relevant Internet MIME type<sup>11</sup> in the first case, or a selection of the values found in the *dc:type* facet of the records as specified in the collections in the other case.

### 3.2 Multi-layer Descriptions

The issue of the separation between web resource and cultural heritage object can be subsumed in the issue of separating objects and subjects in record descriptions. In fact, what does exactly a Europeana record describe? Is it an image, the content of the image, or the object represented in the content of the image? Sometimes this is easy to understand, and sometimes it creates interpretation issues, and the problem of distinguishing between object and subject in a record can go several layers deep. For instance, consider the 1756 publication by Giambattista Piranesi called “Le antichità romane”, containing prints of famous Roman monuments, including the Coliseum. A best seller of the time, the volume appears in several of the collections of Europeana. We analysed 3 items, alla 3 form the Bildarchiv Foto Marburg as data provider. According to one item<sup>12</sup>, the page representing the Coliseum is of *dc:type* *druck* (print), *dc:creator* Giambattista Piranesi, *dc:date* 1756. According to another<sup>13</sup>, the same page is of *dc:type* *amphitheatre*, *dc:description* *Location:Rome* and its *dc:date* is 70/80 a.D. (and no *dc:creator*), but reports (in the *dc:description* field) that the actual photo was taken in 1956, and that the content is an extract of the Piranesi’s book of 1756. This is coherent with several colour photographs of the Coliseum<sup>14</sup> present in the same collection, whose *dc:type* is also *amphitheatre*, *dc:date* is 72/80 a.D., further adding that *dc:format* is *travertine*, and *dc:contributor* is *Vespasianus* (as contractor).

In cases such as Piranesi’s, the number of layers of subjects is multiple: the CHO being described is a 1956 b/w photograph of unknown creator, whose subject is a

---

<sup>11</sup> <http://www.iana.org/assignments/media-types/text/index.html>

<sup>12</sup> <http://www.europeana.eu/portal/record/08501/7B74073B6E9E90F5B572EF6DF20426AF0135202E.html>

<sup>13</sup> <http://www.europeana.eu/portal/record/08501/EA45A0B5F838ABDDD1956DE3BE636A70F1B8EA8A.html>

<sup>14</sup> <http://www.europeana.eu/portal/record/08501/43E4B1EF54983567EC92DDCDD57B3DBD2D4CC013.html>

1756 print whose creator is Giambattista Piranesi, whose subject is a 70 a.D. travertine amphitheatre whose creator (as contractor) was Vespasianus. If we add the issue of the media type of the web resource, as introduced in 3.1, an additional level becomes manifest: we are describing a 21<sup>st</sup> century JPEG image of a 20<sup>th</sup> century photograph of a 18<sup>th</sup> century print of a 1<sup>st</sup> century building.

One of the most frequent dilemmas for a provider of metadata about an object is deciding whether interesting information for which no natural facet is available should be omitted, forced into an inappropriate facet (e.g., the *dc:type* or *dc:date* in the above examples), or dumped into a generic container (e.g., the *dc:description* above). A better solution would be to use a metadata model whose characteristic naturally accommodates the interesting information. As such, a simple solution exists already for the layers of subjects: while in DC the subject is “the topic of the resource [that is] typically [...] represented using keywords, key phrases, or classification codes”, in FRBR “the «has as subject» relationship indicates that any of the entities in the model, including work itself, may be the subject of a work”.

Thus the example of the print by Piranesi could be expressed more precisely and with fewer misunderstanding as a record for a JPEG image whose *frbr:subject* is a 1956 photo whose *frbr:subject* is a 1756 print whose *frbr:subject* is a roman building, for instance as in figure 1<sup>15</sup>:

```
ontology:Photography rdfs:subClassOf frbr:Work .
ontology:Print rdfs:subClassOf frbr:Work .
ontology:Amphitheatre rdfs:subClassOf frbr:Work .

resource:jpeg-photo a ontology:JPEGImage
; frbr:subject resource:photo .

resource:photo a ontology:Photography
; dc:date "1956"
; frbr:subject resource:antichità-romane .
```

**Fig. 1.** An OWL rendering of the correct relationships between subject layers using FRBR (all other facets are expressed as in the original examples for simplicity)

<sup>15</sup> In this and all subsequent examples, we use the following prefixes (please note that the prefixes with the Europeana domain are fictitious, are present in these example only as a suggestion and do not correspond to existing ontologies):

```
@prefix resource: <http://data.europeana.eu/resource/>
@prefix ontology: <http://data.europeana.eu/ontology/>
@prefix foaf: <http://xmlns.com/foaf/0.1/>
@prefix pro: <http://purl.org/spar/pro/>
@prefix dc: <http://purl.org/dc/elements/1.1/>
@prefix frbr: <http://purl.org/vocab/frbr/core#>
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#>
```

```

resource:antichità-romane a ontology:Print
; dc:date "1756"
; dc:creator "Giovanbattista Piranesi"
; frbr:subject resource:colosseum .
resource:colosseum a ontology:Amphitheatre
; dc:date "70/80"
; dc:creator "Vespasianus [Auftrag]" .

```

**Fig. 1.** (Continued)

Using FRBR in its true meaning, so as to distinguish the stratification of layers resulted from describing an object (the idea, the content, the format and the specific item), it is also possible to better distinguish between the different levels (the work, the expression, the manifestation and the item).

EDM, in fact, defines *dc:subject* more precisely than Dublin Core itself, explicitly specifying that its value is either a string or a reference (thus allowing references to other CHOs), and even defines a subproperty of *dc:subject*, called *edm:isRepresentationOf*, to precisely specify the relationship between representations and represented entities (e.g., a statue and a painting of the statue). Yet, the *edm:isRepresentationOf* property is currently greyed out (meaning that it “will not be used in the first implementation so any values provided for them will not be used”), and the current number of subjects specified as strings will make turning them into references a conspicuous and non-trivial job<sup>16</sup>.

### 3.3 Roles and Values

Many of DC properties (e.g., *dc:creator* and *dc:contributor*) are considered insufficient so often that in most Europeana resources that we have checked many actual values are composed of the indication of a role or other contextual information as well as a name (e.g. of the creator and/or contributor). For example consider:

**Creator:** Morel, Francois (Radierer)<sup>17</sup>

**Creator:** Cartographer : Ryther, Augustus<sup>18</sup>

<sup>16</sup> ...not to mention the fundamental problem that in OWL a property can either be a data property (i.e., allowing strings) *or* an object property (i.e., allowing references) but not both, so that any OWL ontology based on EDM will have to choose one representation for the values of all the properties, including *dc:subject*, that allow either strings or references as their values. This, in and by itself, will be a major exercise in reconvension and qualification of existing data sets.

<sup>17</sup> <http://www.europeana.eu/portal/record/08547/DC2A5E3DB3A0675D12DA7699647D1D8FA1B9293D.html>

<sup>18</sup> <http://www.europeana.eu/portal/record/92037/25F9104787668C4B5148BE8E5AB8DBEF5BE5FE03.html>



**Creator:** Friedrich, J. C. F. [Production]<sup>19</sup>

**Subject:** AUTN=Piranesi Giovanni Battista; AUTA=1720/ 1778<sup>20</sup>

The universality of this approach is evident, and similarly evident is the need to provide more information than a mere name, although the syntaxes, the metadata models and the provided information differ.

A better solution could be obtained by promoting strings into first-class objects – e.g., converting people names into individuals of the class *edm:Agent* or *foaf:Person*<sup>21</sup> – and dealing with people names and people’s roles separately. There exist two alternative ways to address efficiently and effectively this issue.

On the one hand, we can create explicit sub-properties of properties such as *dc:creator* or *dc:contributor*. For instance, by allowing “cartographer” to become an explicit sub-property (e.g., property *ontology:cartographer*) of *dc:creator*, the identification of name and roles becomes possible and easy, and consequently the queries become more powerful, as shown in the following excerpt related to *The Cittie of London 31* by Augustus Ryther:

```
resource:cittie-of-london-31 ontology:cartographer resource:ryther .

resource:ryther a foaf:Person
    ; foaf:givenName "Augustus"
    ; foaf:familyName "Ryther" .
```

A problem with this approach is that the TBox of the ontology needs to be modified every time a new role is defined as a new subproperty of *dc:creator*, which is not a good design principle in general.

An alternative is to define people’s roles as individuals of a class. In theory, CIDOC CRM already implements this behaviour by using the meta-property *P14.1 in the role of* [13] (a property of property *P14 carried out by*) so as to specify the role that an agent has in the context of a particular event (e.g., the creation of an artistic work) through an instance of the class *E55 Type*. However, the official RDFS ontology of CIDOC CRM<sup>22</sup> does not implement any meta-property and in reality RDF lacks the expressive power needed to define meta-properties altogether. To simulate a meta-property in RDFS/OWL we may define an additional class that associates the person to his/her role. For instance, the *Publishing Roles Ontology (PRO)*<sup>23</sup> has this behaviour by means of the class *pro:RoleInTime*,:

<sup>19</sup> <http://www.europeana.eu/portal/record/08547/AD78BAEF3D932EF43765BAD78FE8E707EA4AFF85.html>

<sup>20</sup> <http://www.europeana.eu/portal/record/08504/3F54955AB672A4DA3C5A9C268D659A0075170C7F.html>

<sup>21</sup> [http://xmlns.com/foaf/spec/#term\\_Person](http://xmlns.com/foaf/spec/#term_Person)

<sup>22</sup> [http://www.cidoc-crm.org/rdfs/cidoc\\_crm\\_v5.0.4\\_english\\_label.rdfs](http://www.cidoc-crm.org/rdfs/cidoc_crm_v5.0.4_english_label.rdfs)

<sup>23</sup> <http://purl.org/spar/pro>

```

resource:ryther a foaf:Person
    ; foaf:givenName "Augustus"
    ; foaf:familyName "Ryther"
    ; pro:holdsRoleInTime [ a pro:RoleInTime
        ; pro:withRole resource:cartographer
        ; pro:relatesTo resource:cittie-of-london-31 ] .

resource:cartographer a pro:Role .

```

This approach has the advantage of not requiring the modification of the TBox of the ontology whenever a new role is needed: we have just to add a new individual of the class *pro:Role*.

## 4 Experimenting on “data.europeana.eu”

The current implementation of the web site <http://data.europeana.eu> already contains a first selection of the full library of items as RDF statements and they are already queryable via Linked Data aggregators<sup>24</sup>. However some limits can be observed in how the site handles non-existing and non-translated resources, which prevents this implementation from being fully compliant with the Linked Data architecture.

According to [15], one of the most important principles of Linked Data is that all the published resources must be deferenceable. *Content negotiation* is necessary, since information about a resource should be always returned according to the format requested by the user who is navigating the Linked Data, e.g., HTML for humans and Turtle for computer agents. Content negotiation usually has the form of a “303 redirect”: the client asks for a resource in a specific format, the server answers with a “303 See Other” HTTP status code indicating the URL where that requested representation is available to the client, and finally the client gets the content from the specified URL.

Europeana does in fact correctly implement the “303 redirect” approach for the resources it makes available in RDF, but does not behave correctly for non-existing or non-available resources. Regardless of whether the resource exists or not, in fact, the server always returns a 303 redirect, and then, after the client restates the query to the new URL, it returns an error if the resource is non-existent. Good Linked Data policy, on the other hand, is that 303 is only returned on existing resources, and an immediate error is returned for non-existing or non-available resources.

Two different approaches can be adopted for the return code, depending on which perspective is adopted: in an *Open World perspective*, we cannot state whether a resource exists, but we can only say whether we have data about it, while in a *Closed World perspective*, if no data is available about a resource, then the resource itself does not exist.

---

<sup>24</sup> Although the Europeana Linked Data project is still ongoing, we hope that what we describe in this section can be seen as valuable and meaningful suggestions for future modifications of the Linked Data infrastructure of Europeana.

Depending on which of the above views the server adopts, the client should expect a different reply than a “303 See Other” when its initial request cannot be satisfied, as shown in table 1.

**Table 1.** HTTP status code for non-existing or non-available resources in Linked Data

	Open World view	Closed World view
<b>Resource is not available in the requested format</b>	<i>406 Not Acceptable</i> Information about the resource exist but they cannot be returned in the requested format	<i>406 Not Acceptable</i> There are no information about the resource in the requested format
<b>There is no resource with that URI</b>	<i>204 No Content</i> (no body specified) or <i>303 See Other</i> and the indication of a new location that contains empty content in the format requested, e.g. an empty RDF/XML string “<rdf:RDF />” There are no information about the resource	<i>404 Not Found</i> The resource does not exist

## 5 Conclusion

The final question is: how can we improve user queries? How do we work in the direction of an effective enrichment of metadata, in order to address the information needs of the end users? Europeana currently misleads in the object descriptions mainly because of the imprecise mapping of the original metadata set onto the Europeana specific model. The variety of metadata vocabularies, ontologies and models makes things difficult to manage. Approaches towards a better integration of the different metadata sources that feed Europeana could be helped by existing works on the creation, extension and alignment of OWL ontologies, but much work in the mapping of richer models still needs to be dealt with by hand. The EDM Mapping Guidelines [13] should lead content providers to create descriptions compliant to EDM, at the same time leaving them free to use metadata models and value vocabularies most appropriate to their own internal uses. But although by correctly using the Guidelines many of the existing problems would be solved, some aspects of the EDM could be improved, reflecting on the different levels of description of the objects. Full and correct Linked Data compliancy, furthermore, is the right direction for the future and will help Europeana in giving more complete and structured descriptions. Yet, the techniques have to be refined. We wait for the announced Europeana v2.0 at the end of 2014<sup>25</sup>.

## References

1. Aloia, N., Concordia, C., Meghini, C.: Europeana v1.0. In: Agosti, M., Esposito, F., Meghini, C., Orio, N. (eds.) IRCDL 2011. CCIS, vol. 249, pp. 127–129. Springer, Heidelberg (2011)
2. Europeana Project. Definition of the Europeana Data Model. Version 5.2.3 (February 24, 2012), <http://pro.europeana.eu/documents/900548/bb6b51df-ad11-a78-8d8a-44cc41810f22>

<sup>25</sup> <http://pro.europeana.eu/web/europeana-v2.0>

3. Europeana Project. Europeana Semantic Elements Specification, Version 3.4 (March 31, 2011), <http://pro.europeana.eu/documents/900548/4968d0bd-416b-48ed-bc67-6a4a47f09098>
4. Europeana Project. Europeana Data Model Primer (October 26, 2011), <http://pro.europeana.eu/documents/900548/770bdb58-c60e-4beb-a687-874639312ba5>
5. Haslhofer, B., Isaac, A.: data.europeana.eu: The Europeana Linked Open Data Pilot. In: DCMI Proceedings of the International Conference on Dublin Core and Metadata Applications, pp. 94–104 (2011)
6. Haslhofer, B., Roochi, E.M., Gay, M., Simon, R.: Augmenting Europeana Content with Linked Data Resources. In: Paschke, A., et al. (eds.) Proceedings of the I-Semantics, 6th International Conference on Semantic Systems. ACM (2010)
7. Dublin Core Metadata Initiative. Dublin Core Metadata Element Set, Version 1.1. DCMI Recommendation (2010), <http://dublincore.org/documents/dces/>
8. Dublin Core Metadata Initiative. DCMI Metadata Terms, DCMI Recommendation (2010), <http://dublincore.org/documents/dcmi-terms/>
9. Miles, A., Bechhofer, S.: SKOS Simple Knowledge Organization System Reference. W3C Recommendation (August 18, 2009), <http://www.w3.org/TR/skos-reference/>
10. International Federation of Library Associations and Institutions. Functional Requirements for Bibliographic Records Final Report (2009), [http://www.ifla.org/files/cataloguing/frbr/frbr\\_2008.pdf](http://www.ifla.org/files/cataloguing/frbr/frbr_2008.pdf)
11. Lagoze, C., Van de Sompel, H., Johnston, P., Nelson, M., Sanderson, R., Warner, S.: Abstract Data Model. ORE Specification. Open Archives Initiative (October 17, 2008), <http://www.openarchives.org/ore/1.0/datamodel>
12. TC 46/SC 4. ISO 21127:2006. Information and documentation—A reference ontology for the interchange of cultural heritage information. International Organization for Standardization (2006)
13. Crofts, N., Doerr, M., Gill, T., Stead, S., Stiff, M.: Definition of the CIDOC Conceptual Reference Model. Version 5.0.4, ICOM/CIDOC CRM Special Interest Group (November 2011), [http://www.cidoc-crm.org/docs/cidoc\\_crm\\_version\\_5.0.4.pdf](http://www.cidoc-crm.org/docs/cidoc_crm_version_5.0.4.pdf)
14. Europeana Project. Europeana Data Model Mapping Guidelines (February 24, 2012), <http://pro.europeana.eu/documents/900548/ea68f42d-32f6-4900-91e9-ef18006d652e>
15. Heath, T., Bizer, C.: Linked Data: Evolving the Web into a Global Data Space. Morgan & Claypool Publishers (2011), doi:10.2200/S00334ED1V01Y201102WBE001

# The Europeana Linked Open Data Pilot Server

Nicola Aloia, Cesare Concordia, and Carlo Meghini

Istituto di Scienza e Tecnologie dell'Informazione,  
National Research Council, Pisa, Italy

(nicola.aloia, cesare.concordia, carlo.meghini)@isti.cnr.it

**Abstract.** The Linked Data is a set of principles and technologies providing a publishing paradigm for sharing and reusing RDF data on the Web. The Linked Data Cloud is expanding at a very high speed since 2007, when the Linked Data Project was launched. Europeana, the European Digital Library, subscribes to the view of a web of data, and the distribution of cultural heritage data is one of the main objectives established by the Europeana Strategic Plan. The paper illustrates how Europeana publishes Linked Data, with focus on the technological approach adopted.

**Keywords:** Linked Data, Linked Data Server, Europeana.

## 1 Introduction

The Linked Data is a set of principles and technologies providing a publishing paradigm for sharing and reusing data on the Web [1]. In a well-known paper [3] Tim Berners-Lee, coined the term Semantic Web which advocated to extend the web of documents as "a web of data that can be processed directly and indirectly by machines", with the ability of discovering new resources through the interconnection of similar data. The Europeana project goal is to provide integrated access to digital objects from the cultural heritage organizations of all the nations of the European Union. To achieve this objective, Europeana provides a set of tools, such as the portal, a set of APIs for programmatic access to its resources, etc. Having the ability to provide metadata as Linked Open Data, is very important for Europeana to attract new users and new providers because the linked data paradigm enables the use of digital representations of cultural artifacts for generating knowledge [7]. For this reason, the implementation of Linked Open Data Pilot Server (LODPS) is an important step for Europeana, its partners and third parties. It paves the way towards achieving two crucial Europeana targets: enable connecting related data and makes them easily accessible using common Web technologies and enable everyone to access, reuse, enrich and share data.

Distributing the whole Europeana datasets as Linked Open Data (LOD) requires to process the existing Europeana metadata, coded according to the Europeana Semantic Elements (ESE), to obtain RDF descriptions as required by Linked Data approach (ESE enrichment and transformation), and to define an agreement with every data provider to publish their data as open data.

We decided to focus on finding solutions for the first issue, by creating a Linked Open Data Pilot server that exposes as Linked Data a subset of the Europeana content belonging to those providers, who want to make their data available on the web. Note that the Linked Open Data Pilot server is technically separated from the Europeana production server.

The approaches and technical solutions adopted for transforming ESE metadata into a richer and more flexible format and to link the Europeana data with other sources are described in [2].

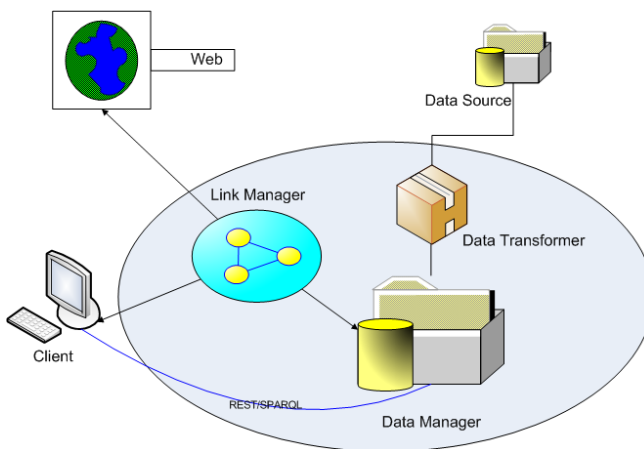
This paper will describe in details the server built to publish Europeana Linked Data, showing its architecture, and the technical solutions adopted.

## 2 Linked Data Server

In [4] a number of best practices, known as the Linked Data Principles, are proposed. The basic idea is to use the architecture of the World Wide Web to share data on a large scale:

1. Use URIs as names for things. That is, use the URIs to report not only documents, but also objects and concepts of the real world.
2. Use HTTP URIs so that people can look up those names.
3. When someone looks up a URI, provide useful information, using the standards (RDF\*, SPARQL). That is, all URIs must be dereferenceable, i.e. client applications use the HTTP protocol to look up the URI and to obtain a description of the resource identified by the URI, using standard notations.
4. Include links to other URIs. so that they can discover more things.

A Linked Data Server is an HTTP server application that offers the ability of discovering new resources through the interconnection of *similar* data and complies with the *Linked Data Principles*.



**Fig. 1.** Linked Data Server

A LD server can be logically divided in three components: a *Link Manager*, a *Data Transformer* and a *Data Manager* (Fig. 1). The role of the *Data Transformer* is to process the original dataset formatting and enriching it in order to publish it as linked data, and store it by means of the *Data Manager*. Generally speaking the *Data Manager* provides functionalities to index, search, access and maintain data. The *Link Manager* is the front end for the client application. It usually provides functionalities to process requests and format responses according to Linked Data specifications [1].

### 3 Web of Data: Making URIs Dereferenceable

The HTTP protocol was originally designed to manage HTML documents, i.e. compound hypermedia resources formatted according to a common rendering language. In Linked Data, instead, resources are not only documents; they can also be real world objects or abstract concepts. Moreover in the web of documents hypertext links are simply a way to access documents and don't contain information on the resource accessed. In the Linked Data paradigm, every resource is identified by a URI, and a Linked Data (LD) server must be able to dereference the URI i.e. to propose a description of the resource identified by the URI if this resource is not a document.

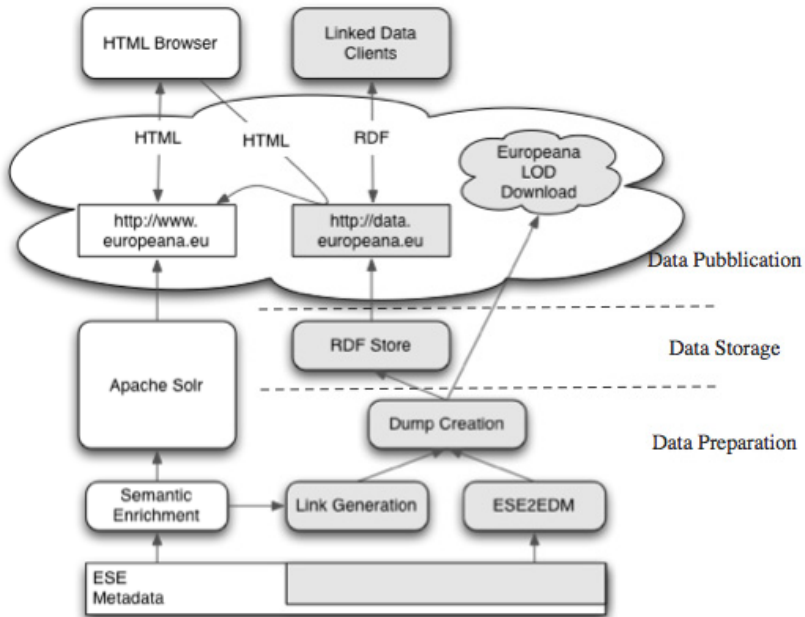


Fig. 2. LD server technical architecture

There are two main strategies to implement the URI dereference mechanism: 303 URIs and hash URIs, both are described in details in [6]. In summary:

- In the 303 URIs strategy if the server recognizes that the URI identify a real object or an abstract concept , it sends to the client a HTTP response code “303 See Other” and a link to a web document describing the resource, the client then asks for this document and the server returns it with HTTP code “200”
- In the hash URIs strategy the *fragment identifier* of a URI (the part of a URI that follows the # symbol) is used to identify real-world objects and abstract concepts, without creating ambiguity.

Advantages and disadvantages of both strategies are discussed in [6]; in essence: hash URIs have the advantage of reducing the number of necessary HTTP connections, which, in turn, reduces access latency, 303 URIs, on the other hand, is more flexible because the redirection target can be configured separately for each resource.

For the Europeana LD server we decided to adopt the 303 URIs strategy, the main reasons for this choice are explained in the following paragraph.

## 4 The Europeana Linked Data server

The following picture, taken from [2], shows the overall architecture of the Europeana Linked Data server.

According to the publishing steps individuated in [1] we can describe the server as follows:

- **Data preparation:** this step is executed by three server components. The ESE2EDM component that downloads the data from the Europeana dataset and maps the ESE metadata records into EDM information objects, the Link Generation component that enrich the EDM objects and creates the links to other Linked Data Sources and the Dump creation component that merges the results of the above components into a set of dump files. A detailed description of the algorithms and tools implementing these components can be found in [2].

The output of this step consists of a number of RDF triples (currently: 115.769.306) that are stored a) in a set of dump files, b) in an RDF-Store. Every dump file contains RDF triples belonging to a specific collection; dump files are published and can be downloaded.

- **Data Storage:** The data storage is implemented using an RDF store. It is important to note that the dataset of the Europeana Linked Open Data Pilot is loaded in the RDF Store using a batch procedure and it does not change, this means that data manipulation is not critical in the Europeana Linked Data Server. On the contrary the response time for queries to the RDF Store is very critical since every data resource in the store is accessed via query.

Results presented in [8], where performances and features the main RDF Store are compared, shows that the Virtuoso server is a good solution from the performance point of view. Moreover Virtuoso provides also a REST web service to perform SPARQL queries over HTTP, this feature is used to publish Europeana Linked Data via SPARQL.



- **Data Publication:** the component publishing Linked Data is implemented by a Web Server and by a library of Java servlets. The Web Server receive every request and redirect it to i) the download area if a dump file is requested, ii) the servlets library if, instead, a resource is requested. The servlets implement the 303 URIs dereference strategy. The implementation algorithm is based on the HTTP server-driven content negotiation mechanism [5], which enables HTTP clients and servers to negotiate a possible answer to a specific request. When a client requests a resource the LOD server checks the expected media type, if the request is for an HTML document a ‘303 redirection’ is issued to the document describing the resource in the official Europeana Web server (www.europeana.eu). In case the client expects an RDF media type then the request URI is parsed to individuate the type of resource requested (a resource map, a proxy, an aggregation or an item) and the resource ID. Using these information a new URI is created and used as 303 redirect target. If the client accepts this redirection the URI is used as query parameter for a SPARQL Describe query. An example of interaction is shown in Fig. 3



Fig. 3. Dereferencing URI for an RDF resource

## 5 Accessing Linked Data

The Europeana LOD server provides three way to access linked data (see Table 1): via file transfer it is possible to download the whole dataset or specific collections, using the SPARQL GUI it is possible to query the LOD dataset to obtain single resources or collections of resources according to defined query filters, the HTTP/GET protocol allow to access resources via URI dereferencing.

As described in the previous chapter the implementation of the URI dereference strategy is based on the analysis of the “Accept” field value in the HTTP request header.

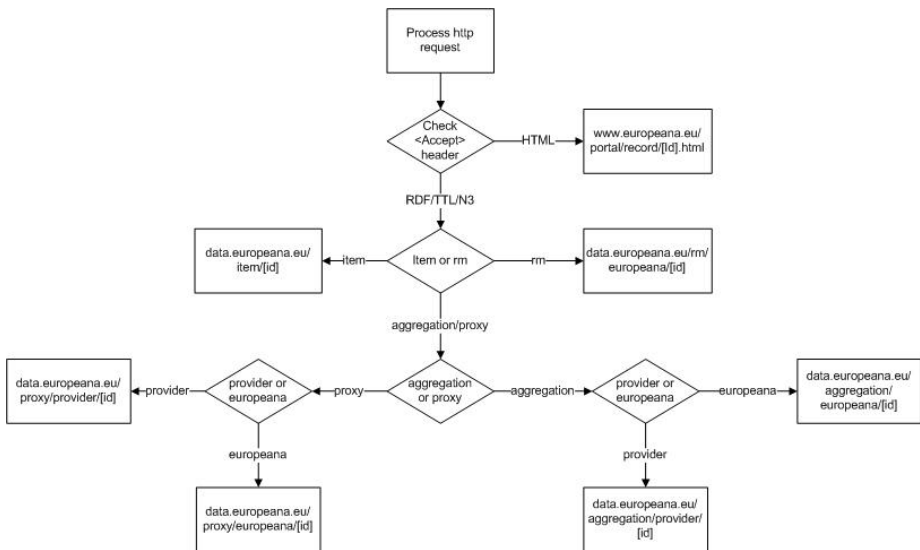
The role of this field in an HTTP request is to specify the acceptable media for the response, the value consists of a comma separated list of media types with the associated quality factor (a ‘q:’ followed by number in scale 0 1) i.e. the degree of preference indicated by the client for the specific media type, if the quality factor is not defined for a media type it is considered as 1. An example is the following:

Accept: application/rdf, text/html;q=0.9, text/plain;q=0.8

**Table 1.** Europeana Linked Data publishing methods

<b>Publishing method</b> \ <b>Data published</b>	<b>File Transfer</b>	<b>REST/SPARQL</b>	<b>HTTP/GET</b>
<b>Complete dataset</b>	Download dataset dump	N.A.	N.A.
<b>Collection of resources</b>	Download collection(s) dump	SPARQL ‘Select’ query	N.A.
<b>Single resource</b>	N.A.	SPARQL ‘Describe’ query	URI dereference

The Accept header value is parsed to check if the request asks for an HTML document or if an RDF resource is needed. When an html document is requested, the client request is redirected (303 redirection) toward the document describing the resource in the Europeana server: [www.europeana.eu](http://www.europeana.eu).



**Fig. 4.** HTTP request parsing

If instead the client asks for an RDF/TTL/N3 media type then the requested URI is parsed to individuate (i) the actual category of the resource requested (a resource map, a proxy, an aggregation or an item) and (ii) the resource ID.

The different categories of resources served by [data.europeana.eu](http://data.europeana.eu) are [2]:

- **Item** ([http://data.europeana.eu/item/\[id\]](http://data.europeana.eu/item/[id])), a real-world object for which digital resources are available through Europeana
- **Resource Map** ([http://data.europeana.eu/rm/europeana/\[id\]](http://data.europeana.eu/rm/europeana/[id])), a OAI-ORE resource map [10] indicating meta-level statements about the creation and publication of ORE data (ORE aggregations and their aggregated resources)

- **Provider aggregator** ([http://data.europeana.eu/aggregation/provider/\[id\]](http://data.europeana.eu/aggregation/provider/[id])) the digital resources submitted on an object by its provider, it also gives meta-information on the digital resource aggregation process, e.g., the name of the data provider
- **Europeana aggregator** ([http://data.europeana.eu/aggregation/europeana/\[id\]](http://data.europeana.eu/aggregation/europeana/[id])) the digital resources maintained by Europeana for the object, it also gives meta-information on the data aggregation process, which is created by Europeana
- **Provider proxy** ([http://data.europeana.eu/proxy/provider/\[id\]](http://data.europeana.eu/proxy/provider/[id])) gives all the data that applies to the real-world object, from the perspective of the data provider
- **Europeana proxy** ([http://data.europeana.eu/proxy/europeana/\[id\]](http://data.europeana.eu/proxy/europeana/[id])) gives all the data that applies to the real-world object, from the perspective of Europeana.

The LOD server gets a resource via a SPARQL ‘DESCRIBE’ query (i.e. a specific form of SPARQL query that returns a RDF graph describing the resource). The query is executed in the Europeana dataset stored in the Virtuoso triple-store.

The query result is parsed by the LOD server, formatted according to the requested media type and sent back to the client.

## 6 Conclusions and Next Steps

The Linked Open Data Pilot server publishes a subset of the Europeana dataset as Linked Data. It offers three different ways to clients for getting Europeana Linked Data: URIs dereferencing via the server located at <http://data.europeana.eu>, SPARQL queries via Web Service and data dump file downloading. The technology adopted and the code developed is open source [9]. The current activity on the Europeana Linked Data pilot has three main goals: to increase the number of the Europeana content providers contributing to the Europeana Linked Data dataset, to refine the dataset quality by adding links to other Linked Data sets and to improve the implementation of the server functionalities. Concerning this last activity probably the biggest challenge is to improve the Data Store performances. Even if the Virtuoso Server query response time is acceptable for a pilot server, we’re working to identify a solution applicable in a ‘production’ server, when potentially the whole Europeana Dataset could be published as Open Data. Another activity is to investigate other technical solutions adopted for the data publication to improve technical interoperability with other Linked Data servers.

## References

1. Heath, T., Bizer, C.: *Linked Data: Evolving the Web into a Global Data Space*. Morgan & Claypol Publishers (2011)
2. Haslhofer, B., Isaac, A.: *data.europeana.eu: The Europeana Linked Open Data Pilot*. In: DCMi Meetings and Conferences, DC 2011, The Hague (2011)
3. Berners-Lee, T., Hendler, J., Lassila, O.: *The Semantic Web*. Scientific American Magazine (May 17, 2001)
4. Berners-Lee, T.: *Linked Data - Design Issues* (2006),  
<http://www.w3.org/DesignIssues/LinkedData.html>

5. RFC 2616 – Hypertext Transfer Protocol – HTTP/1.1 – (Section 12: Content Negotiation)
6. Sauermann, L., Cyganiak, R.: Cool uris for the semantic web - w3c interest group note (2008), <http://www.w3.org/TR/cooluris/>
7. Gradmann, S.: Knowledge = Information in Context: on the Importance of Semantic Contextualiation in Europeana. Technical report, Berlin School of Library and Information Science, Humboldt University (April 2010), <http://www.scribd.com/doc/32110457/Europeana-White-Paper-1> (retrieved April 30, 2011)
8. Haslhofer, B., Roochi, E.M., Schandl, B., Zander, S.: Europeana RDF Store Report. Technical report, University of Vienna (retrieved April 30, (2011), <http://eprints.cs.univie.ac.at/2833/>
9. <https://github.com/behass/slodr>
10. <http://www.openarchives.org/ore/>

# Managing Authenticity through the Digital Resource Lifecycle\*

Maria Guercio and Silvio Salza

Università degli studi di Roma “La Sapienza”  
maria.guercio@uniroma1.it,  
salza@dis.uniroma1.it

**Abstract.** On the basis of principles and methodologies developed by the major projects on digital preservation, the paper addresses the fundamental problem of authenticity management, and specifically of defining appropriate mechanisms and tools to transform the presumption of authenticity into the capacity of its verification. The approach we propose is to concentrate on the digital resource lifecycle, since, in order to make a proper assessment, one must be able to trace back all the transformations the digital resource has undergone since its creation, and that may have affected its authenticity. For these transformations one needs to collect and preserve the appropriate evidence that would allow, at a later time, to make the assessment. We have therefore developed a model of the digital resource lifecycle in order to identify the main events that impact on authenticity and to define precise operational guidelines to specify which evidence should be collected and how to organize it. A case study analysis is currently being performed to check the validity of the model and to see how it specializes on several specific environments. Preliminary results are already available and confirm that the model is sound and that the implementation of the guidelines can be worked out effectively and with a fairly reasonable amount of effort.

**Keywords:** authenticity, curation, long-term preservation, repository.

## 1 Introduction

Authenticity is considered in the literature and by all major projects on digital preservation in the area of digital libraries and institutional repositories as one of the most crucial characteristics to be maintained over time and consistently documented for evidence and future use [1-4]. In the last decade the scientific community has developed robust principles and a basic methodological approach to this issue, with the aim of establishing among different communities a common understanding of the concepts involved and on the specific tools to be implemented.

The InterPARES projects (1999-2012) [5] have addressed the creation, maintenance and preservation of digital records, with specific reference to authenticity. A

---

\* Work partially supported by European Community under the Information Society Technologies (IST) program of the 7th FP for RTD - project APARSEN, ref. 269977.

major finding is that, to preserve *trustworthy digital records* (i.e., records that can be demonstrated to be *reliable, accurate* and *authentic*), records creators must create them in such a way and in such a form that it is possible to maintain and preserve them. This entails that a relationship between a records creator and its designated preserver must begin at the time the records are created.

The CASPAR project (2006-2009) [6] has dedicated specific effort in developing a methodology for the management of authenticity in the digital environment. In particular the Caspar Authenticity Team has identified a set of attributes that allow to capture information relevant for the authenticity as it can be collected along the lifecycle of the digital resources; the Team has also developed tools and procedures to manage this information.

On the basis of this analysis, a well stated terminology has been included in the new version of the OAIS reference model <sup>1</sup> [8], and considerable agreement has been reached in the digital preservation community on some basic principles:

- it is not possible (feasible) to preserve electronic resources as *original unchanged resources*: one may have only the ability to reproduce them in the form of *authentic copies* thanks to the preservation of authentic copies of digital components;
- authenticity *cannot be recognized as given once and for ever* within a digital environment: a clear distinction should be made between the authenticity of the preserved record/resource (not necessarily the same objects as those originally deposited) and the procedure of *evaluating* and *validating* the same object;
- the profile of the authenticity has to be considered as a *process* aimed at gathering, protecting and/or evaluating information/set of attributes mainly about identity and integrity of the digital resource.

As a consequence of this, and because digital resources curation is increasingly and dynamically based on the concept of *trust* <sup>2</sup>, the heart of the problem has become how to support the unavoidable principle of trustworthiness and, even more, how to transform these general concepts and assumptions into a series of concrete, measurable and well interconnected steps to sustain the presumption of digital authenticity both in the pre-ingest phase and in the repository itself.

In other words, the fundamental question is how to transform *presumption* into *evidence*, and how to define a multilayer approach able to provide a convincing structured series of events, agents and information, related to the interconnected phases of the digital resources lifecycle, in order to verify their integrity and authenticity conveniently to the various levels of analysis and according to the specific needs of consumers.

---

<sup>1</sup> In the draft of the new standard authenticity is defined as: “the degree to which a person (or system) may regard an object as what it is purported to be. The degree of authenticity is judged on the basis of evidence”.

<sup>2</sup> In the Merriam-Webster dictionary trust is identified as “a charge or duty imposed in faith or confidence or as a condition of some relationship”, a sort of “glue which binds that relationship together”, whose ingredients have to be identified and described for effectiveness of the custody.

These questions have been thoroughly addressed by the APARSEN project [9], a NOE funded by the EU (2011-2014) with the goal of overcoming the fragmentation of the research and of the development in the digital preservation area by bringing together major European players, to combine and integrate these efforts into a shared program of work, thereby creating a pre-eminent virtual research centre in digital preservation in Europe.

The activity carried on in APARSEN has concentrated on establishing operational guidelines to properly manage the digital resource lifecycle in order to:

- conveniently trace (for future verification) all the transformations the digital resource has undergone since its creation that may have affected its authenticity and provenance,
- collect and preserve for each of these transformations the appropriate evidence that would allow, at a later time, to make the assessment and, more precisely,
- develop a model of the digital resource lifecycle, which identifies the main events that impact on authenticity and provenance and investigate in detail, for each of them the evidence that has to be gathered in order to conveniently document the history of the digital resource.

The final target of this effort is, of course, trying to achieve the interoperability among the systems where the digital resource is kept or preserved along its lifecycle, since there may be several changes of custody, and therefore very often the evidence about authenticity needs to be managed and interpreted by systems that are different from the ones that gathered it.

Indeed the model we present in this paper is based on a broad analysis of the main standards developed or supported by the major research projects in the preservation area [12-16]. Because the focus is set on the events and the responsibilities in the various phases of lifecycle (creation, keeping, preservation), the main standards that have been considered are (apart from OAIS [7], which is the basis of our common understanding in building an open framework for digital preservation) those concerning the creation and keeping of accurate, complete and reliable records in the e-government field. Even if intended for a specific domain, these rules are relevant for the preservation of any type of resources.

Achieving such an ambitious goal requires time, consensus and a thorough discussion. For this reason the methodology we propose in the following sections, although we already have some encouraging feedback from test applications, should only be considered as a preliminary step, and a basis to derive more complete operational guidelines to improve the current (too often very poor) practices in managing digital authenticity and providing evidence in preservation systems.

## **2 Authenticity and the Digital Resource Lifecycle**

In order to properly assess the authenticity of a Digital Resource (DR) we must be able to trace back, along the whole extent of its lifecycle since its creation, all the transformations the DR has undergone and that may have affected its authenticity and

provenance. For each of these transformations [10] one needs then to collect and preserve the appropriate evidence that would allow, at a later time, to make the assessment, and that we shall call therefore *authenticity evidence*.

Under quite general assumptions, we may consider the DR lifecycle as divided in three phases:

- **Pre-ingest.** This phase begins when the DR is delivered for the first time to a keeping system and goes on until the DR is submitted to a *Long Time Digital Preservation (LTDP) system*. During the pre-ingest phase, the DR may be transferred between several keeping systems and may undergo several transformations.
- **Ingestion.** This phase encompasses both the transfer of the DR from the producer to the LTDP system, and the subsequent control and transformations the DR undergoes during the ingest, which, referring to the OAIS terminology, marks the passage between the SIP (Submission Information Package) and the AIP (Archival Information Package).
- **Long term preservation.** This phase begins after the DR is ingested by a LTDP system and goes on as long as the DR is preserved. As for the pre-ingest and the ingest phases, also during the LTDP phase the DR may undergo several transformations, notably format migrations, aggregations etc. Moreover it may get moved from a LTDP system to another one.

The pre-ingest phase has been introduced as a separate phase from the ingest to represent the part of the lifecycle that occurs before the delivery to the DR of a LTDP system. Collecting evidence for all the transformations the DR undergoes during this phase is of the utmost importance to assess the its authenticity.

Each transformation a DR undergoes during its lifecycle is connected to an *event*, which occurs under the responsibility of one or more people, whom we shall call *agents*. A transformation may involve one or several DRs and one or several agents, and produces as a result a set of DRs, possibly new versions of the ones that were the object of the transformations.

A very ambitious goal would be to try to determine 'all' possible events that are relevant with regard to the authenticity of a DR, and to draw precise guidelines to specify which authenticity evidence should be collected for each of these events, and how to organize it.

This would be indeed a very interesting result since, as we have seen, the DR moves along its lifecycle from system to system, and therefore these systems, when they exchange the DR, need to interoperate in order to exchange also the related authenticity evidence. Interoperability means agreeing on a common ground, and therefore common guidelines would form the basis that would allow such systems to interoperate.

### 3 The Core Set of Lifecycle Events

Unfortunately, the variety of events that may occur during the DR lifecycle is very large and depends, at least in part, from the specific environment. Nevertheless, it is



possible to consider at least a minimal *core set of events*, that includes the most important ones, as well as the ones which are likely to occur in most of the environments in which DRs are produced and managed. The core set should be considered as a sort of common basis on which different keeping and preservation systems may agree, thus achieving at least a basic degree of interoperability in the exchange and management of authenticity evidence.

In our investigation we have considered a reasonable variety of environments, notably natural science data, health care data, social science data and administrative data repositories. As a result of our analysis, we have proposed the core set of events that we briefly outline in the following subsections. For a more complete description one should refer directly to the APARSEN project documentation [10].

### 3.1 Pre-ingest Phase

The author of a DR is the person who, individually or as the representative of an institution, takes the responsibility of the content of the DR and of the descriptive information associated to it, when the DR is created in the pre-ingest phase i.e. delivered for the first time to a keeping system, a term by which we mean any kind of system where the DR is kept, once it has been created, until it is submitted to a LTDP system.

This definition encompasses a large variety of situations. For instance in a scientific experimental environment, where a DR is a collection of experimental data, the author is the scientist in charge of the experimental measures, who certifies the authenticity and the integrity of the data and of the associated descriptive information, and the keeping system is the computer system used to store and managed the experimental data, for instance a data base centered system. Similarly, in a document management environment, where the DR is an electronic document, the author is the person who prepares the final version of the document, and the keeping system is the Electronic Record Management System (ERMS) where the document is kept.

During its stay in the keeping system the DR may undergo a series of transformations that may affect both its content of the DR and the descriptive information associated to it. For instance the DR may go through format migrations (even before it enters the LTDP custody), or it may get integrations of its content and/or of its metadata, or it may eventually be aggregated with other DRs to form a new DR. Moreover, before getting to LTDP, the DR may be transferred, one or several times, between different keeping systems.

In the model, the core set for the pre-ingest phase comprises the following events:

- **CAPTURE:** the DR is delivered by its author to a keeping system;
- **INTEGRATE:** new information is added to a DR already stored in the keeping system;
- **AGGREGATE:** several DR, already stored in the keeping system, are aggregated to form a new DR;
- **DELETE:** a DR, stored in the keeping system is deleted, after its preservation time has expired, according to a stated policy;
- **MIGRATE:** one or several components of the DR are converted to a new format;
- **TRANSFER:** a DR is transferred between two keeping systems.

### 3.2 Ingest Phase

In the model, the ingest phase includes also the submission of the DR to the preservation repository. It involves therefore both the system where the DR was kept and the LTDP system to which is delivered.

The content and the structure of the SIP (Submission Information Package) through which the DR is delivered must comply with a submission agreement established between the system where the DR was kept (i.e. the Producer in the OAIS reference model) and the LTDP system (the OAIS). After the submission, the DR may eventually be deleted in the origin system, but this action should be considered a separate event. The DR identity is maintained in the keeping system, but a new identity may be given to the DR in the LTDP system.

Altogether it is a crucial phase, since during the ingestion all the authenticity evidence about the pre-ingest life of the DR must be collected, accepted and checked by the LTDP system, and becomes, according to the OAIS reference model, part of the PDI (Preservation Description Information) of the AIP (Archival Information Package).

In the model the following two events are considered in this phase:

- **SUBMIT:** a DR is delivered by the keeping system where it is stored (producer) to a LTDP system;
- **INGEST:** a DR delivered from a producer is ingested by the LTDP system and stored as an AIP.

Even in a minimal situation, as long as a clear distinction between keeping and preserving is done, as it should be, both the above events occur. Thus providing precise guidelines on which evidence should be included in the SIP and how it should be structured is a crucial requirement to ensure interoperability.

### 3.3 Long Term Digital Preservation (LTDP) Phase

This phase begins when the DR is delivered to a LTDP (Long Term Digital Preservation) system and goes on as long as the DR is preserved. During this phase, the DR may undergo several kinds of transformations, that range from format migrations to changes of physical support, to transfers between different preservation systems.

The OAIS is here considered the reference model. According to the OAIS, many activities are carried out in connection with each of these events, but the model will focus here on the sole aspects related to authenticity and provenance of the DR and on the information (authenticity evidence) that has to be gathered and preserved in the PDI (Preservation Description Information), and more specifically in the Provenance, Context and Fixity components.

Analyzing this phase many possibilities have to be considered: the possibility of transfers between LTDP systems, which is very likely to happen in the long run, and the possibility of changes in the structure of the preserved DRs (integration, aggregation etc.), that routinely happens in the health care sector, since records must enter

preservation as soon they are created and still there may be later the need to introduce corrections. The resulting set of events is then:

- **LTDP-AGGREGATE:** one or several DRs stored in different AIPs, are aggregated in a single AIC;
- **LTDP-EXTRACT:** one or several DRs which are extracted from an AIC to form an individual AIPs;
- **LTDP-INTEGRATE:** new information is added to a DR already stored in the LTDP system;
- **LTDP-MIGRATE:** one or several components of a DR are converted to a new format;
- **LTDP-DELETE:** one or several DR, preserved in the LTDP system and stored as part of an AIP are deleted, after their stated preservation time has expired;
- **LTDP-TRANSFER:** a DR stored in a LTDP system is transferred to another LTDP system.

#### 4 Authenticity Evidence Records

When giving the guidelines that should be followed to ensure interoperability on authenticity among keeping and LTDP systems, beside providing a precise definition of the event, the crucial point is to specify which controls should be performed, which evidence should be collected and how it should be structured.

In the model each event is represented according to an uniform schema:

- the *agent*, i.e. the person(s) under whose responsibility the transformation occurs;
- the *input*, i.e. the preexisting DR(s) that are the object of the transformation, if any;
- the *output*, i.e. the new DR(s) that are the result of the transformation (possibly new versions of input DR(s));
- the *authenticity evidence record*, i.e. the information that must be gathered in connection with the event to support the tracking of its authenticity and provenance.

As the DR progresses along its lifecycle through a sequence of events, an incremental sequence of *authenticity evidence records* is collected by the systems where the DR is kept or preserved, and strictly associated to it. From a practical point of view, an authenticity evidence record is a structured set of information, according to our proposal an XML file of predefined structure, which is strictly related to a given event. At any given stage of its lifecycle a DR brings with it, as part of its metadata, a (temporally) *ordered sequence* of such records, to document all the transformations the DR has undergone and to allow to assess its authenticity and provenance.

Authenticity evidence will follow the DR when it is transferred between different systems, and will accompany it along all its lifecycle. Thus, to ensure interoperability, it is necessary to standardize the way the authenticity evidence is collected and structured. To this purpose existing standards should be accurately considered, as for instance the Open Provenance Model (<http://openprovenance.org>).

At the moment – as already mentioned in the introduction – the model developed in the framework of the APARSEN project [10], and here presented, should be considered only as a preliminary step in that direction. Nevertheless, as it turned out from some preliminary practical experiences, it provides a sound basis to derive more detailed operational guidelines and to improve in a significant way the current (and often very limited) practices in managing authenticity and provenance in keeping and preservation systems.

The following subsections discuss a few examples from some events from the core set discussed in sect. 3. For a more detailed discussion one should refer directly to the project documentation.

#### 4.1 SUBMIT

A submit occurs when a DR is moved from a keeping system to a LTDP system. The submit needs to be authorized by the owner of the DR, and involves also the responsibility of the administrator of the keeping system and of the administrator of the LTDP system.

A submission may be considered as the sequence of two steps: i) preparing in the keeping system the DR for shipping; ii) receiving and accepting the DR in the LTDP system. As a consequence, two distinct new versions of the DR are produced: DR' which is kept in the keeping system, and DR'', that is accepted in the LTDP system.

As two different and independent systems are involved in the submission, the keeping system and the LTDP system, the corresponding authenticity evidence record must contain the evidence produced, and conveniently authenticated, by the administrators of both systems. Accordingly there will be two distinct authenticity evidence records, generated and preserved in the two systems.

- **Agents:**
  - owner: the physical or juridical person who originally created the DR;
  - keeping system administrator: the person who submits the DR.
  - LTDP system administrator: the person who accepts the submitted DR.
- **Input:** any DR in the keeping system
- **Output:**
  - DR': the new version of the DR which is kept in the origin system
  - DR'': the new version of the DR, accepted and ready for ingestion.
- **Authenticity evidence record:**
  - Keeping system
    - Event type: submit
    - Identification data of the LTDP system
    - Date and time the DR has been prepared for submission
    - Identification and authentication data of the owner of the DR who has given the authorization for the submission
    - Identification and authentication data of the keeping system administrator
    - Evidence that the DR has been received and accepted by the LTDP system
    - Digest of the DR authenticated by the keeping system administrator

- *LTDP system*
  - Event type: submit
  - Identification data of the keeping system
  - Identification data of the LTDP system
  - Date and time the DR has been received from the origin system
  - Identification and authentication data of the LTDP system administrator
  - Assessment by the LTDP system administrator on the delivery of the DR:
    - Identification and authentication of the keeping system
    - Trustworthiness of the data channel used for the transfer
    - Integrity check performed on the digest received from the keeping system
  - Digest of the DR authenticated by the LTDP system administrator

## 4.2 LTDP-MIGRATE

To migrate an AIP or an AIC means to change the data format of one or several of their components. This is generally triggered by technical obsolescence, but may be as well the result of new policies adopted by the LTDP system on accepted formats. As a result of the migration a new version of the DR(s) is generated, which should preserve its intellectual content, despite the format migration. The most delicate part of this transformation, is to verify that the integrity of the individual DR has been maintained, i.e. that its intellectual content has not changed. Migration may occur both in the pre-ingest and in the LTDP phase, we are considering here the latter case.

- **Agents:**
  - *LTDP system administrator*: the person responsible of performing the migration
- **Input:** one or several DRs contained in an AIP or in an AIC
- **Output:** a new version of the AIP or AIC
- **Authenticity evidence record:**
  - Event type: migration
  - Date and time the migration has taken place
  - Identification data of the LTDP system
  - Identification and authentication data of the system administrator
  - Digest of the new version of each affected DR after the migration
  - Statement, for each migrated DR, that the intellectual content of the DR has not changed, specifying also the criteria adopted to make the assessment
  - Digest of the new version of the AIP produced by the migration

## 5 Case Study Analysis

As part of the activities carried on in the APARSEN project, a case study analysis has been performed to check the validity of the model and to see how it specializes on several specific environments [11].

Four case studies have been selected, to cover a reasonable variety of situations:

- a health care data repository,
- two repositories of experimental scientific data,
- a repository of social science data.

Each case study is organized in two parts:

- *What is done right now.* A description and analysis of the *current practices* adopted in the management of the specific repository. The first step is understanding the meaning of authenticity and provenance for the designated community, identifying the main *events* in the DR lifecycle, the transformations the DRs undergo and their impact on authenticity and integrity. Next step is to analyze what is currently done about that, i.e. how DRs are delivered by producers, which controls are performed, which authenticity evidence is collected, etc.
- *What should be done.* That means applying the methodology and the guidelines we propose to the results of the analysis of the current practices, i.e. fitting the lifecycle events into the *core set* of events we propose, identifying the controls that should be done and the authenticity evidence that should be collected, and sketching the improvements one should introduce to correctly manage authenticity and provenance.

We briefly discuss in the following subsections two of the case studies. Due to space limitations the presentation is restricted to the main highlights. For a more detailed account one should refer directly to the complete report that has been published as a deliverable of the APARSEN project [11].

In both cases our model has proved to be effective, since the events in the current situation have clearly mapped into our core set of events, and the structure we propose to represent the events has shown to be adequate. Moreover it has been helpful in formally documenting the workflow and in identifying deficiencies in the management of authenticity evidence.

## 5.1 Repository of the Public Health Care System in Vicenza, Italy

This study deals with several types of DRs, mainly test results (files in DICOM format and more) and medical reports (digitally signed by physicians), each type of DR being handled by a separate workflow. All records are sent to the repository shortly after their creation and managed according to the Italian rules on LTDP, which are very specific and mostly centered on digital signatures and certified timestamps, and mandate to collect many DRs in a single large batch (called Preservation Volume).

We refer here about the workflow of *studies* (i.e. sets of diagnostic images), which involves in the pre-ingest phase several systems under different responsibilities: Modalities (imaging devices) and local and central PACS (Picture Archiving and Communication Systems) that act as keeping systems in the medical structures. The ingest phase involves a preservation system called Scryba which is compliant with the Italian regulations and the OAIS model. According to our model we could clearly identify in the lifecycle the following events:

- **CAPTURE:** studies are generated by modalities and captured by local PACS;
- **TRANSFER:** studies are transferred from a local PACS to the central PACS;
- **SUBMIT:** a SIP is prepared for each study and is moved from the central PACS to the preservation system Scryba;
- **INGEST:** an AIP is generated for each SIP; the process includes some controls on provenance and integrity, generating the PDI from metadata (both explicit and extracted from the DICOM file) and adding a certified timestamp;
- **AGGREGATE:** several AIPs are aggregated in a single AIC (Archival Information Collection) which corresponds to a Preservation Volume.

According to our analysis the management of the authenticity along the lifecycle is rather reasonable, due to the compliancy to the quite detailed national regulations, but a few improvements have been suggested:

- in the pre-ingest and ingest phases the responsibilities for local and central PACS should be explicitly documented in the authenticity evidence records (AER);
- further controls should be introduced in the ingest phase (integrity checks in the transfers) and the outcome of all controls should be recorded in the AERs.

## 5.2 Social Science Data Repository at UK Data Archives

The Archive acquires data from a variety of producers in the academic, public, and commercial sectors, providing continuous access to these data, in a relationship which is based on a network of confidence with the stakeholders. The DR lifecycle is substantially different from the previous case and is concentrated on the ingest and the LTDP phases. According to our model we could clearly identify in the lifecycle the following events:

- **SUBMIT:** a SIP, prepared according to the submission agreement, but with a very large degree of variety in its structure, is submitted by the producer to the Archive;
- **INGEST:** a complex transformation that may require the manual intervention of specialized teams to normalize the structure of the information package (and the data themselves) to meet the Archive standards;
- **MIGRATE** and **DELETE:** two additional events, that correspond to transformations in the process to be implemented.

Although the workflow is currently based on well devised and well documented procedures, and complies with international standards, referring to our model during the analysis has proved helpful in identifying a few problems that should be addressed:

- part of the authenticity evidence is currently not included in the SIP, but derived from data deposit forms and agreements: it should instead be collected by the Producers, structured according to detailed specifications and incorporated in the SIP;
- some of the transformations that are currently performed by specialized teams during the ingestion may affect the authenticity of the preserved DRs since the responsibility of the producers cannot be properly documented; according to the

OAIS model the only clean way to fix the problem could be to require the producers to normalize the data themselves before preparing the SIP, possibly providing them with assistance from the specialized teams, if they need it.

## References

1. Giaretta, D.: Advanced Digital Preservation. Springer, Heidelberg (2011)
2. Giaretta, D., Matthews, B., Bicarregui, J., Lambert, S., Guercio, M., Michetti, G., Sawyer, D.: Significant Properties, Authenticity, Provenance, Representation Information and OAIS. In: IPRES 2009: Proc. of the Sixth Int. Conference on the Preservation of Digital Objects, California Digital Library (2009), <http://www.escholarship.org/uc/item/0wf3j9cw>
3. InterPARES Project, Authenticity Task Force: Authenticity Task Force Final Report (2001), <http://www.interpares.org>
4. Factor, M., Henis, E., Naor, D., Rabinovici-Cohen, S., Reshef, P., Ronen, S., Michetti, G., Guercio, M.: Authenticity and Provenance in Long Term Digital Preservation: Modeling and Implementation in Preservation Aware Storage. In: First Workshop on the Theory and Practice of Provenance, TaPP 2009, San Francisco (2009), [http://www.usenix.org/event/tapp09/tech/full\\_papers/factor/factor.pdf](http://www.usenix.org/event/tapp09/tech/full_papers/factor/factor.pdf)
5. InterPARES (International Research on Permanent Authentic Records in Electronic Systems), <http://www.interpares.org>
6. CASPAR Project - Cultural, Artistic and Scientific Knowledge for Preservation (2006-2009), <http://www.casparpreserves.eu> (access and retrieval)
7. Open Archival Information System (OAIS) – Reference Model, ISO 14721:2003 (2003), <http://public.ccsds.org/publications/archive/650x0b1.pdf>
8. CCSDS: Reference Model for an Archival Information System–OAIS. Draft Recommended Standard, 650.0-P-1.1 (Pink Book), Issue 1.1 (2009), <http://public.ccsds.org/sites/cwe/rids/Lists/CCSDS%206500P11/CCSDSAgency.aspx>
9. APARSEN Project – Alliance Permanent Access to the Records of Science in European Network (2011-2014), <http://www.alliancepermanentaccess.org/index.php/current-projects/aparsen>
10. APARSEN Project: Deliverable D24.1. Report on Authenticity and Plan for Interoperable Authenticity Evaluation System (2011)
11. APARSEN Project: Deliverable 24.2. Implementation and testing of an Authenticity Protocol on a Specific Domain (2011)
12. ISO RM 15489-1:2001 Information and documentation – Records management. Part 1: General
13. ISO 23081-1:2006 Information and documentation – Records Management Processes – Metadata for Records
14. UN/CEFACT: Business Requirements Specification. Transfer of Digital Records, Version 1.0 (2008)
15. DLM Forum: MoReq2 – Model Requirements for the Management of Electronic Records (2008), <http://www.moreq2.eu>
16. DLM Forum: MoReq2010 - Modular Requirements for Records Systems (2011), <http://moreq2010.eu>



# An Innovative Character Recognition for Ancient Book and Archival Materials: A Segmentation and Self-learning Based Approach

Nicola Barbuti<sup>1</sup> and Tommaso Caldarola<sup>2</sup>

<sup>1</sup>Department of Classical and Late Antiquity Studies, University of Bari Aldo Moro  
n.barbuti@ateneo.uniba.it

<sup>2</sup>D.A.BI.MUS. L.L.C., Spin Off of University of Bari Aldo Moro, Italy  
t.caldarola@dabimus.com

**Abstract.** The paper illustrates the invention of a method and an apparatus able to recognize the text in a set of digital images referring to pages of ancient manuscripts or printed books. It includes the following macro steps: identifying and connecting in sequence regions containing words in a subset of the images; structuring a thesaurus of fonts used in those regions; performing the character recognition of one or more images belonging to the set, associating to this recognition a first value of efficiency. The prototype is patent pending (National Pat. Pend. n. BA2011A000038 – Intern. Pat. Pend. n. I116-PCT).

**Keywords:** Intelligent Character Recognition (ICR), Manuscripts, Ancient printed Books, Digital Library, Digital Database of ancient Heritage.

## 1 Introduction

Existing digital libraries, containing digital collections of ancient and valuable handwritten and printed documents and books dating up to the second half of XIXth century, show a level of interactivity still extremely low. For these specific digital contents, indeed, has not been yet possible to develop optical-digital recognition systems and/or text recognition of virtual pages, able to provide an efficient indexing of databases content either already accessible or to constitute over the web 2.0.

None of the latest and most important projects of digital libraries currently available on the web 2.0 (Europeana, World Digital Library, The European Library, etc.) has accessibility and usability features that allow users to see the text content of the reproduced digital objects without having to scroll them through in full. Excluding common cataloguing research (author, title, release notes), in these databases it is not possible to develop any indexing that allows in-depth studies based on the analysis of the recurrence of words, inference about different texts, etc.

This difficulty arises from the nature of the artifacts in question. The complexity and divergence of ancient manuscript spellings, even those paleographic more linear and regular; the kind of old inks used; the obsolescence of the materials, in most cases with damages caused by biological or biochemical factors, mechanical accidents and

human carelessness: all these factors have so far prevented all attempts to go beyond the simple reproduction of these digital artifacts.

Neither the current OCR, ICR and IWR available on the market can be applied to solve the problem of text recognition in ancient documents[1].

If this situation seems almost obvious in the case of manuscripts, because of their nature, it should be less understandable for the printed books. Instead, even for this kind of artifact, in particular for books produced by handprinting (and therefore the entire print production from 1456 up to 1850), the situation is very similar to that of the manuscripts.

The problems, in fact, are not different, even if they affect to a lesser extent. The techniques of composition of the printing plates, the inks used, the alignment of stamps within words, and of the words within lines, the different graphic fonts representative of certain letters, as compared to those commonly used (eg., the "s" represented by a printing font very similar to "f"), different linguistic conventions, the various noise of the images (background noise caused by the press on the reverse side page, smudges and breakage of stamps, ink stains and some other varied cause due to time and men) are all factors that, today, frustrate any attempt to index the contents of digital images of ancient materials through application of recognition systems with satisfying results.

## 2 State of the Art

**Research.** Concurrently with the research and prior to the implementation of the prototype, a survey was carried out both in research on intelligent recognition systems, and among international patents relating to existing applications for recognition of the content of digital images.

It became apparent that the research on intelligent recognition systems, which is able to operate effectively on images of handwritten or printed ancient materials, especially before 19<sup>th</sup> century, has not yet produced significant results despite the efforts made for several years.

*Shape prior model – Ben-Gurion University.* In our opinion, the only interesting research about recognition was carried out by a team of Ben-Gurion University in Israel, whose first results were published in 2008[2].

The paper describes a method of segmentation and recognition of characters which aren't perfectly legible in damaged ancient manuscripts. The process is based on the manual construction of *shape models* representing the possible variability of the characters previously segmented from images of damaged manuscripts. On these is performed a training set that, by matching with the reference models, progressively reduces them to a core, generating for each segmented character a *shape prior* which is the essential reference for the reconstruction of damaged characters and not legible to human eye.

The system, which works on grayscale images, implies a preliminary long and laborious step of manual construction of models of reference, and requires more

progressive training set. Despite the complex laboriousness of the process, the result is certainly interesting.

Even so, the *ratio* of the system has completely different requirements than those of the application object of this work, which will be discussed in the following paragraphs.

**Patents.** The survey among international patents produced no most important results. Faced with an astounding amount of existing applications, it's easy to detect the almost complete identity of functions and applications they exhibit in the output. There are very few exceptions, however always conditioned by elaborative processes that require high manual skills, thus making not very user friendly. We describe briefly some of those that seem to be a useful paradigm to better illustrate the newness of the process that we developed.

*reCAPTCHA*[3]. An interesting patent is that developed in 2008 by researchers from Carnegie Mellon University, USA. They have revised the existing CAPTCHA systems, enabling them to interpret the doubtful words identified by OCR programs, according to a simple, but efficient, system.

When two OCR systems identify differently a word, this word is associated with a known word and sent to a user who has to pass a CAPTCHA test to access a service. It is assumed that if a user is able to identify the known word correctly, then there is an high probability he/she also identify the unknown word. When three users give the same answer, the system stores the word as correct.

In September 2009 the project has been acquired by Google, who uses it to correct errors resulting from OCR scanning of texts. It should however be noted that, for images of books printed prior to the second half of the 19<sup>th</sup> century, the results are not at the level of expectations created at the moment of the discovery and distribution of the system. In fact, the rates of return are still quite low, as it oscillates between 30% and 60% for ancient printed documents, with the highest percentage obtained exclusively on printed texts from the late nineteenth century, while for manuscripts the system did not show any noteworthy working.

*Multifont Optical Character Recognition Using a Box Connectivity Approach (EP 0649113 A2)*[4]. The approach of the system is based on a pattern recognition obtained setting a minimal bounding rectangle around the pattern, sharing out the pattern into a grid and comparing a partitioned vector derived from this grid with other vectors obtained in a similar way starting from known pattern.

Finally, you choose a set of pattern according to Pareto and select one of the patterns thus obtained. The process is laborious, and it is not able to operate effectively on images of ancient documents.

*Document Digitization (Fr 2768825 A1)*[5]. The system is based on the digitization of generic documents, acquiring the image with a scanner connected to one of two computers linked to a network. Scanned images are stored in a high capacity data storage system.

This storage system is also used to save text files "searchable" produced from the second computer with an OCR process on document images. The application does not present any significant functionality able to operate on images of ancient documents.

*Method and Apparatus for Isolating an Area Corresponding to a Character or Word (Us 5144682 A)*[6]. This system works in order to isolate an area corresponding to a character or a word in an OCR device. The main technical problem that must be solved is to recognize characters and words disposed on lines considerably inclined.

The method works on black and white images. Although it is designed to solve a noise level, which is one of the most problematic factors of the images of ancient documents, the system is not able to operate effectively on this kind of images.

*Technique for Correcting Character-Recognition Errors (Gb 2463577 A)*[7]. The system is structured on a method for identifying and correcting failures in the information extracted from images using character-recognition software like OCR or ICR. However, the level of operation on which it is effectively able to act is strictly limited to current documentation concerning financial transactions.

*A2iA's Proprietary IWR, Intelligent Word Recognition*[8]. Some systems, while using more sophisticated methods than aforesaid, base the recognition on the segmentation into words of the text regions.

Such approach is used by the *A2iA Proprietary IWR, Intelligent Word Recognition*, developed by the A2iA, USA. Although this IWR has been successfully used in projects for recognition of handwritten documents, the system is able to operate only if interfaced with specific semantic thesauri structured prior to recognition phase, otherwise it is not capable to make any refund of text. Once again, it is assumed the necessity of a preliminary laborious manual work.

**Some Remarks.** As can be easily inferred from what we have above outlined, nowadays there is not a method or system able to recognize and index images of ancient documents in either automatic or semi-automatic way.

The models and the systems able to work on such kind of document have in common test on ideographic script. Some questions remain about working on alphabetical script. Furthermore, in order to work satisfactorily on such documents, all of them require either a complex manual transcription in electronic format of the content of documents to index, or to structure specific semantic thesauri on which to match the images, followed by an equally laborious training process.

They need too the planning of complex algorithm to extract information to use as models for the matching of digital images, but the output is incomplete and unsatisfactory. And all the scientific and research papers about the digital recognition have the same limit: they purpose not systems, but pure models without sufficient certainty about their working on digital database of paleographic materials.

We consider the reason why the existing systems for optical and/or intelligent recognition of digital images don't work on ancient documents is the methodological approach used in the structuring of such systems[9-15].

This approach maps the words on the scanned images of document pages by associating them in their entirety either to an electronic text inserted manually by an operator, or to thesauri of reference preliminarily structured still manually. This approach, just as it requires a long and complex preliminary manual work, seriously limits the possibility to electronically recognize a large amount of historical texts.

In addition, if this method works fine on certain more recent materials (from the late 19<sup>th</sup> century onwards) because of the linearity of the typographic and graphics composition of the pages, it cannot be the solution to the problem of opening to scholars and mankind an interactive access to the enormous amount of older works both "more" and "minor" still unknown, stored in thousands of historical libraries in the world, whose reproduction presents graphical, typographical, and noise complex and unsolved issues.

### **3 The Method and Apparatus to Recognize Text in Digital Images Reproducing Pages of an Ancient Document (Pat. Pend. Nat. n. BA2011A000038 – Intern. n. I166-PCT).**

#### **3.1 A New System for Recognizing Text in Digital Database of Ancient Manuscripts and Printed Books**

Considering the above, the purpose of the following research has been to set up a method and an apparatus able to recognize and to transcribe full text a percentage rate greater than or equal to 50% of content in a set of digital images, each of which depicts a page of an ancient manuscript or printed document, without requiring a laborious and long preliminary manual work.

The methodological approach used has been different from those previously used for similar systems, as it has had its own premises in the characteristics of discrepancies and noise peculiar to the digital reproductions of ancient artifacts.

The process aims no longer the regions/words of text (regions that contain a word), but the regions/fonts (regions that contain a font), each of which is associated with a sample of corresponding electronic font transcribed manually by an operator.

#### **3.2 Training Stages**

The process has been tested on samples of images of printed and manuscripts documents, different in dating, typographic and graphic characteristics and noise index, calculated over the entire of intrinsic and extrinsic factors of each sample (*intrinsic*: printed books: stamps set used, cleaning of pages, presence of spots or dirt, handwritten gloss, etc.; manuscripts: handwriting readable to naked eye, non homogeneous graphic sign, irregular text lines; *extrinsic*: image quality, resolution, background noise, etc.).

The amount of images to be used for the training set has been calculated as a sample of 100 for the printed documents, 30 for manuscripts, selected according to the following characteristics:

- *ancient printed books*:
  - o 16<sup>th</sup> century; font: italic; noise rating: 80% (very high)
  - o 17<sup>th</sup> century; character: round; noise rating: 60% (high)
  - o 18<sup>th</sup> century; character: round; noise rating: 30% (average)
- *ancient manuscripts*:
  - o letters: handwriting cursive; noise rating: 90% (significantly high)
  - o census: handwriting chancery cursive; noise rating: 75% (very high).

Before running the training, it has been calculated the threshold of iteration of handwriting/typographical fonts used in the selected image set, that is the threshold beyond which the fonts set used to compile the document begin to be iterative and equal to the previous.

Therefore, the image percentage settled as exhaustive of the whole fonts set has been used as sample for the training. This percentage never exceeds 10% of images for each sample, and often the threshold has been reached already with very low percentage (2%-5%).

Then, has been performed in electronic the manual transcription of the content of each percentage of images, to use it as text to matching recognizing font. Obviously, the greater the amount of content transcribed at this stage and reconnected to the regions extracted from the images, the greater the precision in the return of correct semantic structures.

The whole training has been divided into following steps: a) document scanning; b) self-learning of fonts of digital document; c) image segmentation either in regions/words (each region matching one word) or in regions/handwritten or printed fonts (each region matching one handwritten or printed font) varying according to the image noise and to the hard reading of the content; d) proper recognition of text contained in each segment; e) storage of the recognition information; f) application; g) facility.

*a) Document scanning.* This step has carried out the manual scanning of the document which has been submitted to recognition. It has been used a planetary scanner with trilinear CCD 3 X 10.000 pixels rgb real (not interpolated), with real resolutions of 400 dpi up to 2xA2, 600 dpi up to 2xA3, 800 dpi up to 2xA4, 1200 dpi up to 2xA5.

*b) Self-learning of fonts.* For each digital document or whole databases has been assumed the existence either of a set or of multiple sets of fonts that the system should learn to recognize assimilating them permanently.

The learning of the fonts has been the key step of the system, on which the whole process is based: if there are errors or flaws at this stage everything that follows may result inaccurate.

This process is *iterative* and *incremental*. *Iterative* because it is based on a number  $n$  of iterations, *incremental* because at each iteration a new information is added to the set of recognized font, namely *extension of the set of the characters*. In fact, the font for the system can be provided not necessarily by the number of characters provided

for the single alphabet, but it's open, unlimited in relation to the possible variants (graphics, typographical, coloring, etc.) that each character brings with when digitalised.

The receipt of the font for the system occurs either when the index of noise in the iteration is lower than a given threshold  $\xi$  defined *noise rating*, or when during the iteration the noise rating to the *i-th* step concurs with that in the next step, i.e.:

$$\xi_{i+1} = \xi_i$$

that is, the system has finished the learning.

If at the end of the process the results are out of threshold, it may need human intervention which will analyze the noise to manually classify it and train the font to recognize the non-recognized character/s.

*c) Document segmentation.* When all the fonts of a document have been known to the system, the training proceeded with the *segmentation*. This is even an iterative process. Through a multiple temporary and in image memory processing, the segmentation produces a series of image processing that ends when the amount of contrast of character reaches a fixed threshold.

At the end of this step, depending on the settled segmentation, a series of segments has been selected, each of which contains the character set recognized. The setting of the segmentation can be variable according to character, word, line, etc., and must be defined referring to functionality that will be applied.

During this stage, through subsequent proceedings, it has been evaluated the functionality of the system by analyzing the steps of testing and acting on each step to refine the results recorded different from those contemplated.

For this specific step have been assumed different eligibility criteria of the segmentation process, possibly based on statistical values that self-refine as the number of processes increases.

This step has been closed when the segmentation of a relatively large number of digital documents matches to a very low percentage of noise.

*d) Recognition of document content.* After the segmentation step, all the segments produced by the recognition of the sets of characters contained in each segment have been processed. Before switching to the storage of information concerning the recognition, a further process has been performed in order to permanently remove any residual noise due to not properly recognized characters for many reasons (e.g., graphic rendering other than the modern, etc.).

*e) Storage of the information carried out from the recognition.* Once completed the recognition step, the storage and classification of information started. All information obtained from step b) have been developed, classified and stored: the font family, font type, text, noise ratio, as well as standard information like author, title, number of pages, segments per page, different kinds of fonts for the document, etc. All these

information are necessary for the data warehouse in order to make available to users more features as possible even through the use of facilities.

*f) Application.* During this step has been tested the functionality of the system by subjecting it to different test cases performed on different sample set of digital images. In case of further problems detection, it has been tried to refine and correct any step that came into play during the test cases. The testing stage was completed with check of full and effective functioning of the system.

*g) Facility.* The system has been meant to be as open, so it can be implemented with various and diversified facilities. The facilities are extensions to the system that allow to exhibit additional features to relate, classify, map information and then allow the end user to enjoy a richer information.

### 3.3 Percentages of Font Recognition

The basic algorithm has been used in an univocal way on each sample of images. Then it has been calibrated in relation to the feedback obtained from the stage c) and d) of the training step. In particular, for manuscripts it has been necessary to set up different modes of segmentation of the regions, per character on the census (functionality ICR), per word on the letters (functionality IWR), due to the high heterogeneity of the graphic sign in the documents of this latter sample.

At the end of the training, the percentages of fonts and text properly recognized resulted more than satisfactory, although with some differences between different samples of documents:

- *ancient printed books:*
  - o 16<sup>th</sup> century:
    - fonts: 87% exactly recognized, 13% error
    - words: 65% exactly recognized, 35% error
  - o 17<sup>th</sup> century:
    - fonts: 84% exactly recognized, 16% error
    - words: 57% exactly recognized, 43% error
  - o 19<sup>th</sup> century:
    - fonts: 98% exactly recognized, 2% error
    - words: 89% exactly recognized, 11% error
- *ancient manuscripts:*
  - o letters<sup>1</sup>:
    - words: 57% exactly recognized, 43% error
  - o census:
    - fonts: 36% exactly recognized, 64% error
    - words: 42% exactly recognized, 58% error

---

<sup>1</sup> As specified previously, it has been tested some segmentation processes variable depending on the kind of function that will be applied, and in the case of the letters we have chosen to test the function IWR instead of the ICR, due to low legibility of the documents used to the naked eye too.



The above percentages refer, of course, to execution of the process in the first and only solution, without further refinements and calibrations. Additional manual calibrations can be done to correct and eliminate the inevitable noise, achieving a recognition with a high index of accuracy.

As it should be noticed, for nearly all the samples the system was able to carry out an accurate recognition with percentages >50%. The only exception have been the samples related to census, but, whereas among the manuscripts they constituted the highest dating (first half of the 18<sup>th</sup> century), in this case too the percentage of refund can be considered fully satisfactory, especially if parameterized with the current state of the art outlined above.

Moreover, we must not forget that the presented results refer to a training first and only performed on representative samples not numerically significant. It follows that as much information the system receives at this stage, that is to say as many are the images on which carries out the training by having a minimal portion of extracted text as a reference base, the greater the percentage of information correctly identified, and consequently the less the noise, which would then further reduced and, plausibly at least for printed documents, almost entirely phased out in subsequent steps of manual correction, of course, also iterative.

## 4 Conclusions

This paper describes a new prototype of *Intelligent Character/Word Recognition* able to recognizing text in digital images of ancient manuscripts and printed documents. The system currently is patent-pending (Pat. Pend. Nat. n. BA2011A000038 – Intern. n. I166-PCT) and is named *ICRPad*.

It has been tested with features ICR, IWR and OCR on sample sets of digital images of ancient manuscripts and printed documents with positive feedback, such as to sustain right now that further trials, which is currently undergoing, will open possibilities for research, study and interactive use of digital libraries of cultural archival and book heritage, and perhaps not only, both differentiated and with a high level of innovativeness.

The algorithmic structure developed for this application will allow a level of accessibility to the digital documents that to date has not yet reached by any similar system. In fact, it allows two levels of usability applicable contextually.

The first allows the user to search through the document without the need to indexing the content: this procedure, however, would require time, because the segmentation would be contextual to the research step, so it would work effectively for the user only on documents of small capacity.

The other involves the launch in batch of the application on the entire document prior to its overflow into the database, with the consequent indexing of the textual content recognized, so that, once the document has been input in the database with keyword search options, the user can do all the searches he wants with an immediate return.

## References

1. Feldgajer, O.: Universal Character Section for Multifont (EP 0369761 (A2)), [http://worldwide.espacenet.com/publicationDetails/biblio?CC=EP&NR=0369761&KC=&FT=E&locale=en\\_EP](http://worldwide.espacenet.com/publicationDetails/biblio?CC=EP&NR=0369761&KC=&FT=E&locale=en_EP)
2. Bar-Yosef, I., Mokeichev, A., Kedem, K., Dinstein, I.: Adaptive shape prior for recognition and variational segmentation of degraded historical characters. *Pattern Recognition* 42(12), 3348–3354 (2008)
3. von Ahn, L., Maurer, B., McMillen, C., Abraham, D., Blum, M.: reCAPTCHA: Human-Based Character Recognition via Web Security Measures. *Science* 321(5895), 1465–1468 (2008)
4. Krtolica, R.V., Malitsky, S.: Multifont Optical Character Recognition Using a Box Connectivity Approach (EP0649113A2), [http://worldwide.espacenet.com/publicationDetails/biblio?CC=EP&NR=0649113&KC=&FT=E&locale=en\\_EP](http://worldwide.espacenet.com/publicationDetails/biblio?CC=EP&NR=0649113&KC=&FT=E&locale=en_EP)
5. Blondy, A.: Document Digitization (Fr 2768825 A1), <http://patent.ipexl.com/FR/FR2768825.html>
6. Nakamura, M.: Method and Apparatus for Isolating an Area Corresponding to a Character or Word (Us 5144682 A), <http://www.patentbuddy.com/Patent/5144682>
7. Masami, M.: Technique for Correcting Character-Recognition Errors (Gb 2463577), [http://worldwide.espacenet.com/publicationDetails/biblio?CC=GB&NR=2463577&KC=&FT=E&locale=en\\_EP](http://worldwide.espacenet.com/publicationDetails/biblio?CC=GB&NR=2463577&KC=&FT=E&locale=en_EP)
8. <http://www.a2ia.com>
9. Eynard, L., Leydier, Y., Emptoz, H.: Particular Words Mining and Article Spotting in Old French Gazettes. In: *Proceedings of MLDM Posters*, pp. 176–188 (2009)
10. Gordo, A., Llorenz, D., Marzal, A., Prat, F., Vilar, J.M.: State: A Multimodal Assisted Text-Transcription System for Ancient Documents. In: *DAS 2008. Proceedings of 8th IAPR International Workshop on Document Analysis Systems*, pp. 135–142 (2008)
11. Le Bourgeois, F., Emptoz, H.: DEBORA: Digital AccEss to BOoks of the RenaissAnce. *IJDAR* 9(2-4), 193–221 (2007)
12. Leydier, Y., Le Bourgeois, F., Emptoz, H.: Textual Indexation of Ancient Documents. In: *Proceedings of the 2005 ACM Symposium on Document Engineering*, pp. 111–117 (2005)
13. Leydier, Y., Le Bourgeois, F., Emptoz, H.: Towards an Omnilingual Word Retrieval System for Ancient Manuscripts. *Pattern Recognition* 42(9), 2089–2105 (2009)
14. Rawat, S., Kumar, K.S.S., Meshesha, M., Sikdar, I.D., Balasubramanian, A., Jawahar, C.V.: A Semi-automatic Adaptive OCR for Digital Libraries. In: Bunke, H., Spitz, A.L. (eds.) *DAS 2006. LNCS*, vol. 3872, pp. 13–24. Springer, Heidelberg (2006)
15. Toselli, A.H., Romero, V., Pastor, M., Vidal, E.: Multimodal Interactive Transcription of Text Images. *Pattern Recognition* 43(5), 1814–1825 (2010)

# Author Index

- Agosti, Maristella 1, 195  
Aloia, Nicola 241  
Amato, Giuseppe 163  
Artini, Michele 33
- Barbera, Michele 45  
Barbuti, Nicola 13, 261  
Bardi, Alessia 33  
Benfante, Lucio 195  
Bertazzo, Matteo 172  
Biagini, Federico 33  
Bolettieri, Paolo 163  
Buzzetti, Dino 4
- Caldarola, Tommaso 261  
Candela, Leonardo 21  
Casarosa, Vittore 153, 184  
Ceci, Michelangelo 117  
Cervone, Luca 81  
Concordia, Cesare 241  
Coro, Gianpaolo 21
- Debole, Franca 33  
Di Iorio, Angela 172  
Di Mauro, Nicola 105, 141
- Erriquez, Onofrio 11  
Esposito, Floriana 105, 141
- Ferilli, Stefano 105, 129  
Ferrara, Felice 93  
Ferro, Nicola 57, 216
- Gardasevic, Stanislava 153  
Gennaro, Claudio 163  
Guercio, Maria 249
- La Bruzzo, Sandro 33  
Leuzzi, Fabio 129
- Loglisci, Corrado 117  
Ly, Anh Tuan 69
- Macchia, Lucrezia 117  
Madrid, Melody 184  
Malerba, Donato 117  
Manghi, Paolo 33  
Meghini, Carlo 153, 241  
Meschini, Federico 45  
Mikulicic, Marko 33  
Morbidoni, Christian 45
- Orio, Nicola 195
- Pagano, Pasquale 21  
Palmirani, Monica 81  
Peroni, Silvio 228  
Ponchia, Chiara 207
- Quercia, Luciano 117
- Rabitti, Fausto 163  
Rotella, Fulvio 129
- Salza, Silvio 249  
Savino, Pasquale 33  
Schaerf, Marco 172  
Silvello, Gianmaria 57, 216  
Spyratos, Nicolas 69  
Sugibuchi, Tsuyoshi 69
- Tammaro, Anna Maria 184  
Taranto, Claudio 141  
Tasso, Carlo 93  
Tomasi, Francesca 45, 228
- Vitali, Fabio 228
- Zoppi, Franco 33