

Post-proceedings of the 5th Italian Research Conference on Digital Libraries - IRCDL 2009

*A Conference of the DELOS Association and the
Department of Information Engineering of the
University of Padova*

Padova, 29-30 January 2009

Editors: M. Agosti, F. Esposito and C. Thanos
July 2009



DELOS: an Association for Digital Libraries

Maristella Agosti Floriana Esposito Costantino Thanos
(Editors)

**Post-proceedings of the
5th Italian Research Conference on
Digital Libraries - IRCDL 2009**

*A Conference of the DELOS Association and the
Department of Information Engineering of the University
of Padova*

Padova, Italy, 29-30 January 2009
Revised Selected Papers

Volume Editors

Maristella Agosti
Department of Information Engineering
University of Padua
Via Gradenigo, 6/a
35131 Padova - Italy
Email: agosti@dei.unipd.it

Floriana Esposito
Department of Computer Science
University of Bari
Via Orabona, 4
70126 Bari - Italy
Email: esposito@di.uniba.it

Costantino Thanos
Istituto di Scienza e Tecnologie dell'Informazione "Alessandro Faedo"
Area della Ricerca CNR di Pisa
Via G. Moruzzi, 1
56124 Pisa - Italy
Email: Costantino.Thanos@isti.cnr.it

Preface

This volume contains the revised accepted papers presented at the *5th Italian Research Conference on Digital Libraries* (IRCDL 2009), which was held in the Department of Information Engineering of the University of Padova, Italy, from the 29th to the 30th January 2009. The well-established aim of IRCDL is to bring together the Italian researchers interested in the different methods and techniques that allow to build and operate Digital Libraries and to present ongoing research projects. A national Programme Committee was set up composed of 12 members, with representatives of the most active Italian research groups on digital libraries. Among the papers presented at the conference, 12 were accepted for inclusion in these proceedings together with 9 project presentations.

The IRCDL series of national conferences was originally conceived and organized in the context of the activities of DELOS, the Network of Excellence on Digital Libraries (<http://www.delos.info/>), partially funded by the European Union under the Sixth Framework Program from 2004 to 2007. The first IRCDL conference took place in 2005, as an opportunity for the Italian researchers to present recent results on their research activities related to the wide world of Digital Libraries. In particular, young researchers were (and still are) invited to submit the results of their on going research, to be presented in a friendly and relaxed atmosphere, to facilitate constructive discussions and exchange of opinions.

Thanks to the initial support of DELOS, and later on thanks to the support of both the DELOS Association and the Department of Information Engineering of the University of Padua, IRCDL has continued in the subsequent years and is confirming itself as a yearly meeting point for the Italian researchers on digital libraries and related topics. Detailed information about IRCDL can be found at the conference home page (<http://ims.dei.unipd.it/ircdl/home.html>), which contains links to the past editions, and also at the DBLP web site, which contains information on proceedings and presented papers (<http://www.informatik.uni-trier.de/~ley/db/conf/ircdl/>).

We should like here to thank those institutions and individuals who have made this conference possible: the Programme Committee members, the Department of Information Engineering of the University of Padova, the members of the same department who have contributed to the organization of the event, namely Maria Bernini, Emanuele Di Buccio, Marco Dussin, Riccardo Miotto, and Gianmaria Silvello, and the members of the University of Padova Library Centre who have contributed to the organization of the registration and management of the on-line presentations and publications, namely Yuri Carrer, and Ornella Volpato. Finally, we take this opportunity to thank also the additional reviewers who have helped in the revision of the final papers, and who have contributed to the improvement of the papers presented in this volume.

To conclude, we would like to point out that, in addition to the enthusiastic participation of the young researchers and the good will of the members of the various committees, much of the credit for having today the IRCDL series of conferences goes to DELOS. As a matter of fact, DELOS started its activities more than ten years ago as a working group under the ESPRIT Program, then continued as a Thematic Network under the 5th Framework Program and after that went on as a Network of Excellence under the 6th Framework Program. It is generally recognized that during these years DELOS has given a substantial contribution to the establishment in Europe of a research community on Digital Libraries. At the end of 2007 the funding of the DELOS Network of Excellence has come to an end. In order to keep the “DELOS spirit” alive, a DELOS Association has been established as a not-for-profit organization, with the main aim of continuing as much as possible the DELOS activities by promoting research activities in the field of digital libraries. In this vein, there is also the commitment of supporting the new edition of IRCDL, which, as customary, will be held in Padova in January 2010. A Call for Participation will be circulated, but meanwhile we invite all researchers (Italian or non Italian, young or not so young), having research interests in Digital Libraries, to start thinking about possible contributions to the next year conference.

Maristella Agosti, Floriana Esposito and Costantino Thanos
Editors of the Proceedings of the 5th IRCDL

July 2009

Organization and Support

General Chair

Costantino Thanos, ISTI CNR, Pisa

Program Chair

Maristella Agosti, University of Padova

Program Committee

Giuseppe Amato, ISTI CNR, Pisa
Giorgetta Bonfiglio Dosio, University of Padova
Donatella Castelli, ISTI CNR, Pisa
Tiziana Catarci, University of Rome "La Sapienza"
Alberto Del Bimbo, University of Florence
Floriana Esposito, University of Bari
Nicola Ferro, University of Padova
Maria Guercio, University of Urbino "Carlo Bo"
Nicola Orio, University of Padova
Fausto Rabitti, ISTI CNR, Pisa
Pasquale Savino, ISTI CNR, Pisa
Carlo Tasso, University of Udine

Local Organization - University of Padova

Department of Information Engineering:

Maria Bernini
Emanuele Di Buccio
Marco Dussin
Ivano Masiero
Riccardo Miotto
Gianmaria Silvello

Library Centre - CAB:

Yuri Carrer
Ornella Volpato

Additional reviewers

Stefano Ferilli, University of Bari
Raffaele Perego, ISTI CNR, Pisa
Fabrizio Sebastiani, ISTI CNR, Pisa
Giuseppe Serra, University of Florence

Supporting Institutions

IRCDL 2009 benefited from the support of the following organizations:

- DELOS Association
- Institute for Information Science and Technologies of the Italian National Research Council (ISTI-CNR), Pisa, Italy
- Department of Information Engineering, University of Padova, Italy.

Table of Contents

Models for Digital Libraries

On Foundations of Typed Data Models for Digital Libraries	1
<i>L. Candela, D. Castelli, P. Manghi, M. Mikulicic, P. Pagano</i>	

Design and Development of the Data Model of a Distributed DLS Architecture for Archive Metadata	12
<i>N. Ferro, G. Silvello</i>	

Combining Qualitative and Quantitative Keyword Extraction Methods with Document Layout Analysis	22
<i>S. Ferilli, M. Biba, T.M.A. Basile, F. Esposito</i>	

Handling Evolution in Digital Libraries	34
<i>A. Baruzzo, P. Casoto, A. Dattolo, C. Tasso</i>	

Content Description

WibNED Wikipedia Based Named Entity Disambiguation	51
<i>A.L. Gentile, P. Basile, G. Semeraro</i>	

A Continuous Language Modelling Approach for Assessing Real-valued Attributes of Documents	60
<i>R. Bache, F. Crestani</i>	

CLEF Ad-hoc: A Perspective on the Evolution of the Cross-Language Evaluation Forum	72
<i>N. Ferro, C. Peters</i>	

Towards an Integrated Approach to Music Retrieval	80
<i>E. Di Buccio, I. Masiero, Y. Mass, M. Melucci, R. Miotti, N. Orio, B. Sznajder</i>	

Information Access

Searching 100M Images by Content Similarity	88
<i>P. Bolettieri, F. Falchi, C. Lucchese, Y. Mass, R. Perego, F. Rabitti, M. Shmueli-Scheuer</i>	

Design of an Information Retrieval System Based on the Peer-to-Peer Paradigm: An Application to Music Retrieval	100
<i>E. Di Buccio, N. Ferro, M. Melucci, R. Miotti, N. Orio</i>	

A Hybrid Strategy for Italian Word Sense Disambiguation	108
<i>P. Basile, M. de Gemmis, P. Lops, G. Semeraro</i>	

Searching and Browsing Digital Library Catalogues: A Combined Log Analysis for The European Library	120
<i>M. Agosti, F. Crivellari, G.M. Di Nunzio, Y. Ioannidis, E. Stamatogiannakis, M.L. Triantafyllidi, M. Vayanou</i>	

Relevant Projects Presentations

The On-TIME Project	136
<i>T. Catarci, A. Dix, R. Giuliano, M. Piva, A. Poggi, F. Terella, E. Tracanna</i>	

Major Preservation Projects under the 6 th Framework Program	143
<i>V. Casarosa</i>	
Digital Repository Infrastructure Vision for European Research – DRIVER	150
<i>S. Jones, P. Manghi</i>	
Europeana: Towards The European Digital Library	154
<i>N. Aloia, C. Concordia, C. Meghini</i>	
TrebleCLEF: Evaluation, Best Practices and Collaboration for Multilingual Access	158
<i>C. Peters</i>	
MultiMatch: Multilingual/Multimedia Access to Cultural Heritage	162
<i>G. Amato, F. Debole, C. Peters, P. Savino</i>	
D4Science: an e-Infrastructure for Supporting Virtual Research Environments	166
<i>L. Candela, D. Castelli, P. Pagano</i>	
TELplus: Aimed at Strengthening, Extending and Improving The European Library Service	170
<i>M. Agosti</i>	
Author Index	175

On Foundations of Typed Data Models for Digital Libraries

Leonardo Candela, Donatella Castelli, Paolo Manghi,
Marko Mikulicic, and Pasquale Pagano

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

Abstract. Digital Library Systems (DLS) are software applications implementing the functionalities to operate over the (possibly compound) objects of a Digital Library. In the past, the development of DLSs has been mainly characterised by a *from-scratch* approach. Only in the recent period Digital Library System developers started adopting *Digital Library Management Systems* (DLMS), special software systems easing the development process by supporting facilities common to DLSs. The Digital Library community has not yet reached a formal agreement on the detailed functionality these systems must implement in terms of content management support. This work focuses on the foundational aspects of content management for DLMSs by discussing a data model whose novelty is that of (*i*) identifying modeling primitives capable of expressing the nature of the compound objects of any DLS and (*ii*) re-introducing the notion of *object type* as a mean of supporting safe, optimized and efficient DLS implementations. The model will inspire the realization of the first DLMS supporting the expressiveness of compound objects and the traditional capabilities of static typing.

1 Introduction

According to the DELOS Digital Library Reference Model [1] a *Digital Library* (DL) is “*an organisation, which might be virtual, that comprehensively 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 operation of any DL is guaranteed by an operational software system, the *Digital Library System* (DLS). A DLS is a software system developed and deployed to implement every aspect of the DL, i.e. to manage the content the DL is devised for, to take care of the users the DL is conceived to serve, to realise the functionality the DL has been designed for, to put in place the policies governing every aspect of the DL, to guarantee the DL expected quality of service.

The development of such kind of systems as well as their deployment and operation has been performed *from scratch* for many years, i.e. different approaches and technologies have been developed to satisfy the same needs arising in the various application scenarios. In the recent years, the notion of *Digital Library Management Systems*

(DLMS) [2,1], i.e. software system specially devised to assist and ease the implementation of DLSs, has been introduced in the DLSs development arena. DLMSs range from software systems implementing specific aspects of the DLs (e.g. content management functionality in the case of Repository systems like DSpace [3] and Fedora [4]) that once enriched with a set of *extensions* lead to the software needed to implement the expected DLS, to Extensible Digital Library Systems (e.g. DelosDLMS [5]), i.e. a component-oriented software systems implementing complete DLSs that thanks to the openness of the architecture can be easily enriched with additional capabilities, and Digital Library Systems Generators (e.g. gCube [6] and MARIAN [7]), i.e. open systems promoting a development model based on (*i*) a declarative definition and configuration of the expected system and (*ii*) an autonomic deployment and management of the needed constituents.

Beyond the internal architecture and the DLS construction facilities they support, DLMSs are characterized by the data models they implement, also called *compound object data models*. In this paper, we focus on such data models, discussing the modeling primitives they should minimally comprise to describe the *information object model* of any DL. According to the Reference Model, Information objects conceptually represent DL content in terms of a “graph” of digital objects associated with each other through relationships whose “label”, i.e. name, expresses the nature of their association. In this respect, DLMSs differ by the *modeling primitives* they offer, that is the way they characterise DLS digital objects, and thus by the kind of information objects they are able to represent.

Information object models can range from *rigid data models*, where the model basically expresses “one” information object model allowing for light customizations (e.g. DSpace [3], Greenstone [8], Eprints [9] data models), to *flexible models*, where the model can potentially describe “any” information object model (e.g. Fedora data model [4]). DLMSs equipped with “rigid” and “flexible” models tend to offer different kind of modeling primitives, for defining personalised information object models and for managing or querying their digital objects. DLMSs supporting rigid models focus on efficient storage of pre-defined information object models that developers may only configure in some aspects, e.g. label vocabularies. In such DLMSs, modeling primitives are provided directly as ready-to-install DLS components, with pre-defined set of ingestion, access and search functionalities, whose graphical appearance can be customized to the user community needs. DLMSs implementing flexible models expose to developers the modeling primitives they need to create, store and search digital objects according to the structure of the information object model desired by the user community. Because of this genericity of the data model, the rest of facilities needed to produce the specific DLS have to be implemented by developers as extensions of the selected DLMS.

An organization willing to set up a DL requires a DLS matching the needs of its user communities, typically expressed in terms of an information object model and the relative functionalities. As a consequence, the organization has to cope with the trade-off between the benefits of realizing and sustaining the “DLS of their dreams” and the relative realization cost. In this respect, DLSs obtained from rigid DLMSs can be installed with simple configurations steps; realization costs are very low to the organizations,

which, for some DLMS platforms, may also count on free technical support and frequent code updates and patches. On the other hand, if the user community demands for information object models distant from those supported by existing rigid DLMS, organizations will likely rely on flexible DLMSs. In this case, costs are higher: developers, from local labs or from external companies, should be hired for customizing the DLMS to support the required information object model and then complete the DLS with the implementation of the missing features.

In this paper we put the bias on flexible DLMSs. The main motivations behind their realization can be found in the common requirements of developers facing the design of DLSs that match special information object model needs. Typically, such developers end up managing digital objects matching similar structural and functional patterns, independently from the peculiarities of the DLSs they target; e.g. they deal with files, metadata records, collections. Flexible DLMSs were conceived to supply developers with general-purpose functionalities for compound object management they should otherwise re-implement at their own cost. The Fedora Repository [4] is the only DLMS known to offer a flexible compound object data model, which describes compound objects as graphs of digital objects connected by relationships. In particular, the DLMS offers primitives for storing objects, each consisting of one mandatory Dublin Core record and a number of payloads. Objects are versioned and can be connected through relationships labeled with values from a controlled and extendible vocabulary; e.g. *isPartOf*, *isMetadataFor*, etc. The DLMS offers query languages to retrieve objects matching given Dublin Core record field values or objects matching a navigational query (RDF query) in the graph of relationships. DLS designers can thus develop their extensions on top of a Fedora DLMS, which is used for storing and querying the graph of objects corresponding to the DL information object model.

In this respect, Fedora's data model exploits the flexibility of labeled graphs to represent any instance of DL information object models, hence the development of any DLS. However, such graphs of objects are “unaware” of the information object model notion supported at the DL level, whose structure is embedded in the DLS business logic, i.e. in the peculiar way software extensions (e.g. user interfaces) ingest objects and relationships into the DLMS. This lack of information about the objects structure in the DLMS leads to well known software development drawbacks and does not allow the enforcement of storage optimization and efficient access techniques.

In this paper we propose a new flexible data model for compound information objects. The novelty of the model is that of being *statically typed* in the traditional database sense: the existence of an object must be preceded by the definition of its structure, called *type*, i.e. the description in a formal language of the object structure. As well as DBMSs require the definition of database tables prior the realization of database applications managing the records therein, in such a scenario, developers will first define in the DLMS the types matching the structure of the DL information object model and then construct the DLS using the customized DLMS instance. In this case, the DLMS is “aware” of the DLS information object model structure, which is declared as a type by the developers. The system can thus take advantage of the type information to support safe, optimized and efficient management of the DLS information object model at hand. To summarize, the data model proposed in this paper aims at:

- (i) Including the data abstractions necessary for describing any digital library information object model in terms of compound objects;
- (ii) Inspiring the design of DLMSs supporting safe, efficient and optimized storage, access and search of compound objects.

In the following we motivate the foundations of a typed compound object data model by means of real case scenario requirements, then conclude by illustrating our model proposal.

2 Model Requirements

As explained in the previous section, flexible DLMSs support compound object data models devised to meet the requirements of developers willing to realize DLSs. A desirable feature of such models is that of being both “fully expressive” and “minimal”, that is (i) the set of primitives they provide should be capable of describing any information object model and (ii) removing one of such primitives would compromise the expressivity of the data model language, i.e. leave a subset of DLS information object models out of our solution domain. To identify such sets of primitives, a study of common DL behavioral patterns is necessary. Consider the following DLS real-case scenarios.

Real-case 1 (Catalogues). DLSs for management of metadata record catalogues, for example in standard library administration. In this case the records are describing entities, i.e. publications, whose digital payload is not stored within the DLS. The metadata records may obey to a standard bibliographic metadata format, such as Dublin Core or MARC, or to a proprietary format of preference to the DLS user community. The DLS offers efficient search functionality over the metadata records, based on the given format.

Real-case 2 (Archives). DLSs for management of multi-media digital objects coming with their metadata description. In this case, the digital objects are stored in the DLS back-end, can be searched through their metadata, and eventually accessed by proper protocols; e.g. streaming for video digital objects. In principle, the same digital object may be described by several metadata records, conforming to metadata formats specific to relative application scenarios. The DLS offers efficient format-based search functionality over the metadata records. Note that the same scenario is reflected by Institutional Repositories, with publications and bibliographic metadata.

Real-case 3 (Enhanced Publication Management). A special DLS for management of enhanced publication objects, intended as graphs of digital objects consisting of one publication object, with zero or one Dublin Core metadata record description and with relations to other publication objects, cited by the publication. The DLS is capable of (i) ingesting or importing publication objects, i.e. metadata records and/or payloads, from other domains in the form of “simple” compound objects (digital objects with no relationships) and (ii) allowing the user community to construct enhanced publications by specifying reference relationships over such pool of simple compound objects. It is important to observe that: the same simple objects can be part of several enhanced publications; relationships are specified in a second stage and are therefore kept apart from the objects.

Real-case 4 (Federated DLS). In a later stage, when the DLS in Real-case 3 (RC3) has been used for some time, a user community requires a DLS capable of sharing/reusing the publication objects portion of RC3 information object model so as to include its content as part of a further and separate DLS information object model. This user community is interested in growing experiment-oriented enhanced publications, consisting of publication objects in used relationship with special data source objects, that is metadata records describing an external experimental data source; e.g. a database, a Web Site, data files. The new DLS allows the user community to (*i*) ingest publication objects together with metadata records in the same collections defined by RC3 (*ii*) ingest data source descriptions and (*iii*) building experiment-oriented enhanced publications by connecting publication objects with data source objects.

DLMSS capable of supporting all the above DLS real-case scenarios should offer modeling primitivessatisfying the following requirements:

Requirement 1 (Metadata records) The DLMS should support management of metadata records of arbitrary metadata formats and also deal with them in “isolation”, since these may not necessarily come with the digital object they describe. Developers should be able to configure the DLMS to provide efficient storage and search of objects based on the formats required by the DLS model at hand.

Requirement 2 (Digital objects, i.e. payloads or files) Since payloads may obey to different media types the DLMS should provide different ways of efficiently storing and accessing such objects depending on their type. To this aim, the DLMS should be configured prior to objects ingestion by developers, who should specify the kind of digital objects (e.g. mime type) required by the DLS information object model.

Requirement 3 (Object relationships) The DLMS should enable the creation of relationships after the digital objects were, thus regard relationships as independent DLMS entities, conceptually parted from the objects they connect. Relationships should be “labeled”, i.e. suggest and include their semantic nature, and such labels should be customizable, i.e. defined by the developers. Finally, a given labeled relationship is supposed to link two objects of a given typology, not any two objects in the system.

Requirement 4 (Collections) The DLMS should be able to support management of collections of objects, i.e. groups of objects, so as to satisfy some aggregative logic specified by the DLS requirements. DLS should be able to ingest, search and access digital objects into a given collection.

Requirement 5 (Data integrity) The DLMS should be able to support management of information object models of several DLSs at the same time. Furthermore, when such information object models share part of the objects (e.g. RC4: the publication and metadata records collections), the DLMS should prevent DLSs from interfering with each other and compromise the consistency, i.e. integrity, of the information object model structure (e.g. RC4: the second DLS may, by mistake, ingest records of a different metadata format and compromise RC3).

Requirement 6 (Structure) From all requirements above, it appears that the DLMS should be aware of the structure of the metadata records, digital objects and relationships, so as to optimize their storage, ensure their efficient access and their safe manipulation (e.g. sharing between different DLSs) and safe grouping into collections.

Fedora's DLMS matches only parts of such requirements. For example, Fedora's data model does not include the notion of arbitrary metadata record and metadata format management. It is instead assumed that all objects have a mandatory Dublin Core record associated with them and efficient search and access is available only for that format. Fedora digital objects, i.e. payloads, cannot be stored or accessed depending on their media type, since the model does not include the notion of "type of objects" nor a notion of "object collection". For the same lack-of-structure reasons, data sharing and integrity cannot be supported. Different DLSs operating on top of the same Fedora DLMS find themselves managing a common graph of objects, whose consistency depends on how careful and aligned have been the developers in implementing the respective DLS applications. In recent implementations of Fedora, the notion of *Content Model* has been introduced to overcome some of these drawbacks. Content models are represented and stored as special Fedora objects, whose payload bears an XML description of the structure of other objects, i.e. the payload types they are supposed to contain. In this respect part of the benefits of type information are recovered, but others are still unsolved. For example, some form of automatic application correctness checking is possible, but type-dependent optimized and efficient data storage is still not achievable.

3 Typed Compound Object Data Model

In this section we present the foundation of a typed compound object model that meets all requirements presented in the previous Section. We can summarize such requirements in three main conclusions:

- *Object kinds*: metadata records, digital objects and relationships are three independent entities, i.e. object kinds, in the data model. Developers should be able to combine them most appropriately to match the DLS information object model at hand.
- *Object collections*: in order to organize the variety of objects present in a DLS information object model, the notion of group of objects, i.e. collection, is crucial. Objects of the same kind and structure, e.g. Dublin Core metadata records, may not necessarily belong to the same intuitive pool, but be separated in conceptually different collections; e.g. "my Dublin Core records", "your Dublin Core records".
- *Object structure*: in order to enable optimized and efficient storage and access of objects, their structure/typology should be specified prior their ingestion to the DLMS. The DLMS should be "aware" of the parts composing the DLS information object model at hand. In particular, metadata records require their format to be specified, digital objects their media type and relationships their label and the type of objects they are supposed to connect.

The typed data model proposed here adheres to these conclusions. To this aim the model supports the notions of *Object*, *Set* and *Type*. Objects belong to (and are created by) Sets, which are instantiations of Types. Accordingly, we say that Types define the *abstract* structure of objects, while Sets the concrete structure of the objects they contain. More specifically, the relation between the three entities is:

- Types have unique names and can be *instantiated* to create new Sets, whose Objects will conform to the type properties; a Set also acts as “generator” of Objects of that Set.
- Sets have unique names and are used to add, delete or update the Objects therein; in that sense, each Set defines a new unique *concrete type* for all Objects it will contain, whose structure and operators will be that of the Set’s Type; Sets are therefore (possibly empty) containers of “structurally homogeneous” Objects.
- Objects have unique identifiers and they conform/belong to the concrete type, i.e. the Set, through which they were created.

Figure 1 depicts how a Type *Cat* can be instantiated in a number of Sets, here with unique names *myCats* and *yourCats*. Such Sets will have type *Cat* (*myCat::Cat*) and will be able to generate and contain Objects *myCat* and *yourCat* with concrete types *myCats* and *yourCats* respectively (*myCat:myCats* and *yourCat:yourCats*). Objects in *myCats* and *yourCats* share the same structure and behavior of the type *Cat*, but have different concrete types.

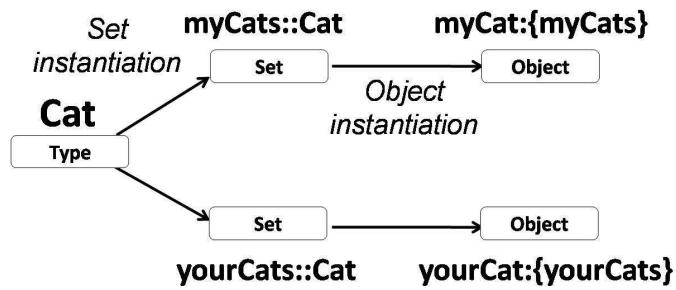


Fig. 1. Relationships between Types, Sets and Objects

In particular, Sets can be of three main Types:

- *Atom Type*: Atoms are intended as “simple digital objects”, i.e. files. A Set of atoms can be specialized to contain Objects of a specific format/media type;
- *Structure Type*: Structures are human/machine readable descriptions of digital or physical entities in the real world, thus they match the notion of metadata records as well as Description Types correspond to the metadata formats. Structure Types are defined in terms of “properties”, i.e. tuple types as sequences of (attribute,value domain) pairs, or more complex structures, such as trees of records;
- *Relation Type*: Relations represent binary relationships between Objects of two given Sets, i.e. they consist of two pointers at two existing objects in the given target sets. As all other objects, relation objects have an identity (they can be themselves target of Relations); unlike other objects, relation objects cannot exist without the

relative target Objects to exist. Relation Types depend on two Sets, relative to the objects that can be possibly associated by the Relation objects in the Relation Set of the given Type. Note that the name of the Set is the “label” to be associated to all Relation objects therein.

The model attempts to capture the essence of modern Digital Libraries, whose content may be the result of the combination of new content with existing content, in turn possibly heterogeneous and not locally available. In this scenario, Atom objects might be available from different sources and might come or not come with a human/machine readable “description”; e.g. metadata description. Equally, Description objects might exist in the system without for the objects they describe to be present; e.g. metadata catalogs. Finally, given a DLS populated with objects, relations between such objects are and should be defined independently by DLS communities, based on their needs of connecting the objects. This is why relation objects are independent from the objects they connect, i.e. they are always added in a “second stage” and can be removed with no implicit impact on the objects they were binding together.

Figure 2 illustrates a formalism through which we can describe DL information object conceptual models in terms of our data model – as well as the Entity-Relationship model is used to describe the data model of a reality of interest. According to the model, circles represent Sets of Atoms, rectangles represent Sets of Structures, whose Types are described by double-lined rectangles to which they are connected with a dashed arrow. Rhombuses represent Sets of Relations: the totality (partial or total relation) and the cardinality (one to one, one to many, many to many) of the relationship it defines are represented through different arrows. In the figure, one Atom PDF in *MyPapers* can be related to none or one structure object in *MyDublinCore*, while each structure object in *MyDublinCore* must be associated to exactly one Atom object in *MyPapers*.

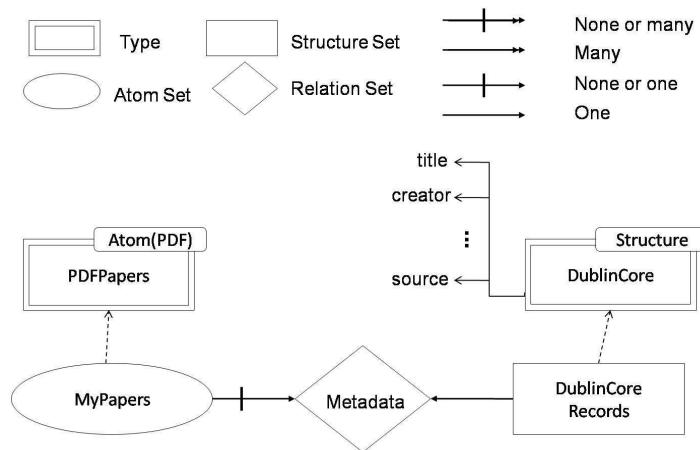


Fig. 2. Information Object Models formalism

Figure 3 shows the information object model arising from the specification of RC3 and RC4. The data model formalism clearly states the structure and content of the Objects to be managed by the two DLSs as well as their shared portion, namely the Sets *MyPapers* and *DublinCoreRecords*.

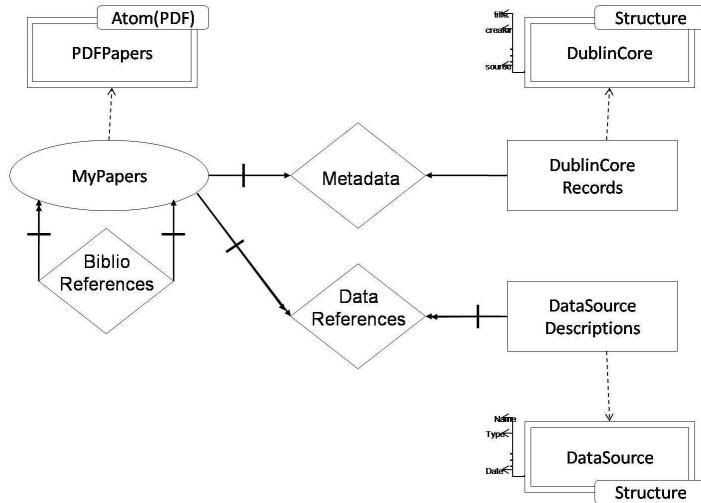


Fig. 3. Information object model for the Real-Cases 3 and 4

4 Typed flexible DLMSs: Dorothy

Such a model offers a way for designers to unambiguously specify the features of the information object model for their DL. In our research, however, this formalism also shows the minimality and completeness of the compound data model it captures. This proof of concept was crucial before starting the design and development activities of a DLMS implementing the data model. Currently, we begun the development of a prototype of a DLMS called *Dorothy* (*Digital Object Repository with Types*), offering APIs for creating Sets according to a Type and creating and querying Objects into Sets using an X-Path based query language. The DLMS exploits type information for two main reasons:

- Checking type-correctness of the DLS software operating on-top of Dorothy: programs must respect the Type constraints imposed by the DL designer and thus cannot mis-behave when manipulating the Objects of a Set;

- Optimizing the storage of objects based on Type information: Structures are stored into Sets endowed with indexes; Dorothy storage has a modular architecture, which will be incrementally enriched by modules exploiting Type information to store Atoms according to properties such as their mime type, size and access mechanisms.

The design and development of Dorothy is based on the architectural principles described in [10] and on the language and grammar presented in [11].

5 Conclusions

In this work we discussed the motivations behind the realization of DLMSs and illustrated the modeling requirements of information object models by means of four DL real-case scenarios. We have shown how existing flexible data models for DLMSs fail at satisfying part of such requirements and then presented the foundations of a typed compound object model that appears to fulfill them. Currently we are developing a prototype of a DLMS, called Dorothy, which supports such data model so as to offer a practical proof of concept of the typed data model approach to DLMS construction.

Acknowledgments This work is partially supported by the INFRA-2007-1.2.1 Research Infrastructures Program of the European Commission as part of the DRIVER-II project (Grant Agreement no. 212147).

References

1. Candela, L., Castelli, D., Ferro, N., Ioannidis, Y., Koutrika, G., Meghini, C., Pagano, P., Ross, S., Soergel, D., Agosti, M., Dobrev, M., Katifori, V., Schuldt, H.: The DELOS Digital Library Reference Model - Foundations for Digital Libraries. DELOS: a Network of Excellence on Digital Libraries (2008) ISSN 1818-8044 ISBN 2-912335-37-X.
2. Ioannidis, Y., Maier, D., Abiteboul, S., Buneman, P., Davidson, S., Fox, E., Halevy, A., Knoblock, C., Rabitti, F., Schek, H., Weikum, G.: Digital library information-technology infrastructures. International Journal on Digital Libraries **5** (2005) 266–274
3. Tansley, R., Bass, M., Stuve, D., Branschofsky, M., Chudnov, D., McClellan, G., Smith, M.: The DSpace Institutional Digital Repository System: current functionality. In: Proceedings of the third ACM/IEEE-CS joint conference on Digital libraries, IEEE Computer Society (2003) 87–97
4. Lagoze, C., Payette, S., Shin, E., Wilper, C.: Fedora: An Architecture for Complex Objects and their Relationships. Journal of Digital Libraries, Special Issue on Complex Objects (2005)
5. Ioannidis, Y.E., Milano, D., Schek, H.J., Schuldt, H.: DelosDLMS. International Journal on Digital Libraries **9** (2008) 101–114
6. Assante, M., Candela, L., Castelli, D., Frosini, L., Lelii, L., Manghi, P., Manzi, A., Pagano, P., Simi, M.: An Extensible Virtual Digital Libraries Generator. In Christensen-Dalsgaard, B., Castelli, D., Jurik, B.A., Lippincott, J., eds.: 12th European Conference on Research and Advanced Technology for Digital Libraries, ECDL 2008, Aarhus, Denmark, September 14-19. Volume 5173 of Lecture Notes in Computer Science., Springer (2008) 122–134

7. Gonçalves, M.A., Fox, E.A.: 5SL - A Language for Declaratively Specification and Generation of Digital Libraries. In: Proceedings of the 2nd ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL'02), Portland, Oregon (2002) 263–272
8. Witten, I.H., Bainbridge, D., Boddie, S.J.: Power to the People: End-user Building of Digital Library Collections. In: Proceedings of the first ACM/IEEE-CS joint conference on Digital libraries, ACM Press (2001) 94–103
9. Millington, P., Nixon, W.J.: EPrints 3 Pre-Launch Briefing. *Ariadne* **50** (2007)
10. Candela, L., Manghi, P., Pagano, P.: An Architecture for Type-based Repository Systems. In: Foundations of Digital Libraries II, Pre-proceedings of the Second Workshop on Foundations of Digital Libraries, Budapest, Hungary, September. (2007)
11. Castelli D., Candela L., M.P.M.M.P.P.: Typed Compound Objects Models for Digital Library Repository Systems. Technical report, Istituto di Scienze e Tecnologie dell'Informazione, Consiglio Nazionale delle Ricerche (2008)
12. Candela, L., Castelli, D., Manghi, P., Pagano, P.: Item-Oriented Aggregator Services. In: Third Italian Research Conference on Digital Library Systems, Padova, Italy (2007)

Design and Development of the Data Model of a Distributed DLS Architecture for Archive Metadata

Nicola Ferro and Gianmaria Silvello

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

Abstract. In this work we present the architecture of a *Digital Library System* (DLS) that enables the preservation, management and sharing of archival descriptive metadata in a distributed environment. Furthermore, we describe in detail the design and development of the data model envisioned for this DLS. The result is a flexible and scalable architecture able to deal with the complexities of archival metadata.

1 Introduction

The role of *Digital Library Systems (DLSs)* in collecting, managing, sharing and preserving our cultural heritage is increasingly prominent in several contexts. DLSs have become the fundamental tool for pursuing interoperability between different cultural organizations such as libraries, archives and museums. Collecting and managing the resources of these organizations is fundamental for providing a wide, distributed and open access to our cultural heritage. One of the most important issue to be addressed is the interoperability between different cultural organizations which may differ in their organizations, policies and data management procedures.

In this work we consider the archives which are a complex and challenging cultural organizations. When archives are considered, the most relevant resource type that must be taken into account is metadata. In the archival context metadata are called archival descriptive metadata and express the archival descriptions; these are the foremost digital resources preserved by the archives. Indeed, most archival documents are not available in digital form, but they are described and represented by metadata. In the work we have been carrying out we have underlined that DLS technologies need to be revisited to be well-suited and successfully applied to the management of archival metadata and digital objects [3].

In [4] we presented a distributed DLS architecture to address the problem of sharing archival metadata between different archives spread across a geographic region. In particular, we considered the Italian Veneto Region archives which promote a related project called *Sistema Informativo Archivistico Regionale (SIAR)*. The main goal of the SIAR project is to develop a DLS for sharing archive metadata spread across the territory. Archive metadata are geographically distributed

across the Veneto Region and they are preserved in several local archives; the SIAR objective is to develop a DLS able to provide advanced services on regional archive metadata [1].

The DLS architecture designed in the SIAR project is divided into three basic layers: the data exchange infrastructure described in [2,4], the metadata management layer and the user interfaces layer. The main goal of this work is to present and describe the data model of the SIAR system realized throughout the metadata management layer of the DLS architecture.

Section 2 reports the background of SIAR principles which are worthwhile for understanding the design and development of the data model. Section 3 presents the design of the metadata management layer of the SIAR DLS and Section 4 reports some final remarks.

2 SIAR architecture and the design choices

The SIAR architecture is based on three main layers built one upon the other: the transportation layer, the metadata management layer and the user interfaces layer. A schematic view of this three-level conceptual architecture is shown in Fig. 1.

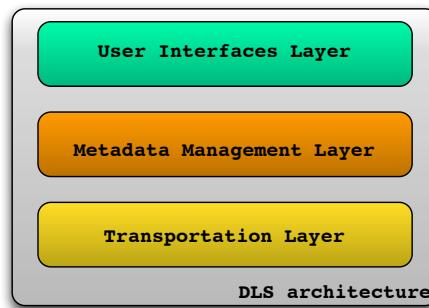


Fig. 1. The three main layers of the conceptual SIAR DLS architecture

The transportation layer is represented by the data exchange infrastructure. The main role of this infrastructure is to permit metadata exchange between the local archives spread across the territory. Basically, it is composed of the *Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH)* [11] protocol that is a standard de-facto for metadata exchange in distributed environments. OAI-PMH is based on the distinction between Data Provider and Service Provider which, respectively, offer metadata and harvest metadata to provide services. Data Providers are the components that make metadata available to the Service Providers that harvest metadata. Each Data Provider manages its own

metadata and it is independent and autonomous from the outside information systems. The Service Provider role is to harvests metadata by the different Data Providers and to performs advanced services on these harvested metadata. In our case study, the Veneto Region is the Service Provider which gives advanced services such as data and public access to the harvested metadata and the archive keepers act as Data Providers because they supply archive metadata. OAI-PMH requires *Dublin Core (DC)* metadata format as minimum requirement; DC, a tiny and lightweight metadata format, is becoming the preponderant means for the exchange of information in a wide distributed environment.

The intermediate layer is the metadata management level that has to manage archival metadata retaining their full informational power; when archival metadata are considered several issues must be taken into account. Indeed, archival metadata reflect the structure of the archive which is strongly hierarchical and related to the organization preserving them; an archive organizes its documents describing them from the general units to the specific units in a tree structure. Every document is linked with the other documents of the same hierarchy and, at the same time, with the environment in which they were created and preserved. In order to describe the archival documents correctly, the metadata have to retain the relationships between the documents themselves and with the production and conservation environments; in other words, they have to maintain the hierarchical structure and the context of archival documents.

Typically, archival metadata are described by metadata format encoded in big and complex eXtensible Markup Language (*XML*)¹ files able to retain hierarchy and context information; a relevant example is the *Encoded Archival Description (EAD)* metadata format [7] which is the standard defined by the Society of American Archivists in cooperation with The Library of Congress. EAD encourages archivists to use collective and multilevel description, and because of its flexible structure and broad applicability, it has been embraced by many repositories [6]. The use of EAD is widespread in the United States of America and also in the European Union; for instance the “Nationale Archief”² in the Netherlands preserves a big collection of EAD metadata in Dutch or the “Archives Napoléon”³ is based on EAD metadata in French. Unfortunately, EAD allows for several degrees of freedom in tagging practice, which may turn out to be problematic in the automatic processing of EAD files. Moreover, EAD files are heavy and difficult-to-move. Moreover, the EAD file cannot be accessed with a variable granularity; indeed, to access a specific archival unit it is necessary to visit all the archival tree [9], starting from the general fonds that usually is the tree root and going down the node path until the required unit is found. It has been underlined [8] that the EAD metadata standard is not well-suited for use in a distributed and dynamic environment like the SIAR is. Indeed, the SIAR system does not rely on a specific metadata format that can turn out

¹ <http://www.w3.org/XML/>

² <http://www.nationaalarchief.nl/>

³ http://www.archivesnationales.culture.gouv.fr/chan/chan/archives_napoleon-averti.htm

to be problematic in a specific domain, but it leaves the choice to the archives participating the system.

At the same time, we have to consider that not every metadata format is well-suited to deal with archival descriptive metadata. For instance, the Dublin Core metadata format is not enough informative to be used in a stand-alone way in the archival context. For these reasons we have defined a data model based on nested sets which permits to handle full informative archival metadata using small and flexible metadata formats like the Dublin Core [5]. In particular, the XML metadata files must be organized in a proper way that enables the preservation of hierarchy and context information. By means of the use of sets we can retain the full informative power of archival metadata; the sets are nested one inside another to create a hierarchy that reflects the archival organization. We can say that the archival information is brought by the documents content and by the archive structure; in the SIAR architecture the content is brought by the metadata and the structure is retained by the nested sets organization. Metadata and sets are the two basic logical entities constituting the metadata management layer of the SIAR DLS architecture.

The third level of the SIAR DLS architecture is the presentation layer constituted by the user interfaces. The system presents two main interfaces: the first is a general-purpose interface dedicated to a generic user-type such as archivists, historical researchers, public administrations or private organizations that will use the advanced services available in the SIAR DLS; the second is dedicated to specialized users who can use this interface to add, remove or update archival metadata.

3 SIAR Metadata Management Layer

The SIAR metadata management layer is the central component of the DLS architecture. Throughout this level it is possible to manage, preserve, retrieve and share full expressive archival metadata.

In Fig. 2 we can see a sketch of the 3-layers architecture with a zoom on the metadata management level composed by: the database, the data logic and the application logic. The data logic is realized by a component called *datastore*, instead the application logic is composed by the *service manager* and the *web component manager*.

We can clearly see the four main entities of the SIAR DLS: metadata, set, user and group; the main function of the database and the datastore is to create, read and update the various instances of these entities. Furthermore, they supply the data to the application logic. The service manager realizes the various services of the SIAR DLS such as the representation of metadata, the reconstruction of the set organization and the OAI-PMH data and service providers. The web component manager implements those methods that permit the interaction of the services with the web services exploiting the SIAR DLS. Currently we have designed a web service that realizes a user interface interacting with the SIAR DLS; in Fig. 2 it is represented as the presentation logic.

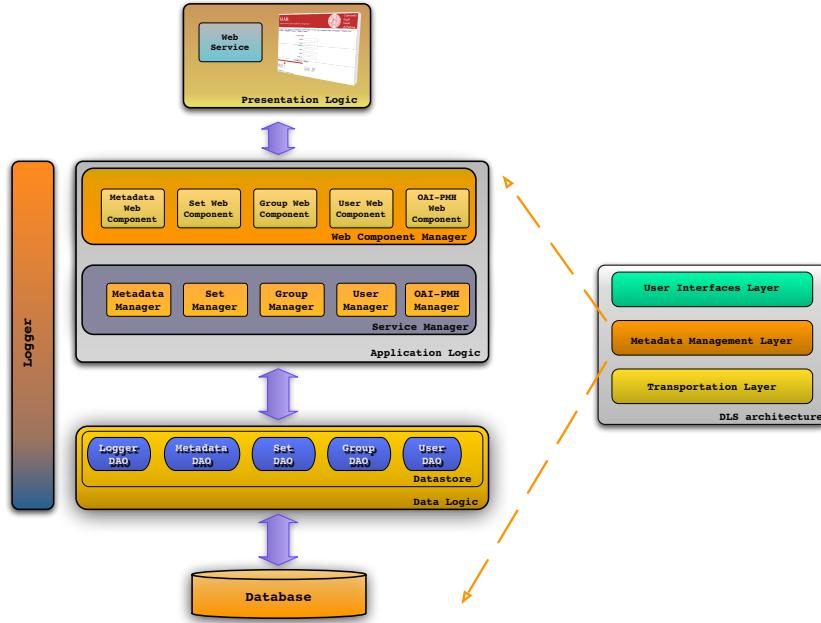


Fig. 2. Composition of the SIAR metadata management layer

The second component presented in the Fig. 2 is the logger. The logger keeps track of all the activities of the system at all the levels; indeed, it is the only component transversal to the whole metadata management layer. The logger registers all the accesses of the users and all the operations done by the system such as the creation of metadata, sets, users or groups, the access and any update of the entities or the exceptions risen and handled by the system.

The elements composing the metadata management layer are described in the following three subsections: subsection 3.1 presents the database conceptual schema in order to describe the entities treated by the SIAR DLS. Subsection 3.2 describes the structure and the components of the datastore constituting the data logic of the layer and subsection 3.3 explains the role of the service manager and the web component manager composing the application logic.

3.1 SIAR Database Design

The conceptual design of the database reflects the world that the SIAR DLS represents. In Fig. 3 we can see the conceptual schema of the database.

The conceptual schema is composed of four main entities: metadata, set, user and group. The metadata entity is defined by five attributes: *id*, this is the unique identifier of a metadata; the identifier is assigned automatically by the system, calculating an hash function of the metadata content. *body* containing

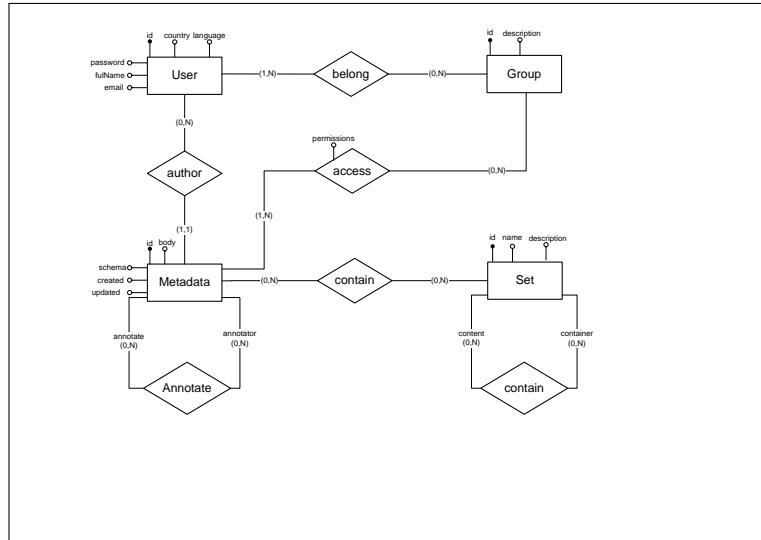


Fig. 3. SIAR DLS Conceptual Database Schema

the metadata content encoded in XML; the *body* is defined in the database as a native XML data type. Every metadata format encoded in XML can be stored in the database and the *schema* attribute reports the metadata format; throughout this attribute we are able to individuate the metadata format and to parse the XML file in a correct way. *Created* which is the attribute storing the time stamp in which the metadata was created in the database and *updated* which stores the time stamp in which a metadata could have been updated.

The recursive relationship called *annotate* indicates if a metadata annotates (*annotator*, the metadata is the annotator of other metadata) some other metadata or if it is annotated (*annotation*) by other metadata. Thanks to this relationship we can have notes on metadata expressed as metadata themselves, we can retain the metadata history preserving all the versions of metadata that have been modified and we can retain information about the original repository of the metadata if they have been harvested by means of the OAI-PMH protocol. A metadatum can annotate many other metadata, for instance in the *history case* an old metadatum is the annotator of its updated version and the *repository case* where the metadatum describing a repository is the annotator of all the metadata coming from that repository. Furthermore, *annotate* recursive relationship is useful to represents nested metadata. Metadata are the most important entities in the SIAR system and they establish relationships with all the other entities. The *author* relationship specifies that a metadatum must be created by a user also if it has been harvested via OAI-PMH protocol. The *belong* relationship indicates that a metadata can belong to a specific set or to many sets defined in the system; a metadatum can also belong to no set at all. The *access* relationship indicates that each metadatum must have at least one group

with read or write permissions on it (e.g. a metadatum has to be read at least by the group to which the creator user belongs) and can have many groups with permissions on it.

The *set* entity is defined by three attributes: *id* which is the unique identifier of a set, *name* which is a mandatory attribute indicating the name of the set and an optional attribute called *description* which can contain a free text description of the set. There is a recursive relationship called *contain* which indicates if a set is contained (it is a subset) by other sets or if it contains (it is a superset) other sets or both. Throughout this recursive relationship we express the possibility of creating a nested set organization that expresses the hierarchical structure of an archive. The structure of this entity is complaint with the structure of an OAI-set defined by the OAI-PMH protocol; indeed, it defines an OAI-set as a native feature composed by three main components: *setspec* which is the identifier of the OAI-set, the OAI-set name and the description which may contain free text or a piece of XML. This connects in a straightforward way the SIAR with OAI-PMH and permits to exploit an important feature as the selective harvesting which is the procedure that permits the harvesting only of metadata owned by a specified OAI-set [10].

The *user* entity is defined by six attributes: *id* which is the unique identifier of a user that can be seen as the username to access the system, *password* the password chosen by the user to access the system, *fullname* contains the full name of the registered user, *email* is the e-mail address of the user, *country* and *lang* indicate the origin of the user and are useful for setting up multilingual services. The *group* entity is defined by an *id* and a *description* of the group. The group and the user entities are related by the *contain* relationships that establish to which groups a user belongs. A user is not required to insert a metadatum to participate in the system but the permissions to create, read and update metadata are related to the group, thus users have to belong to at least one group to operate on metadata.

Fig. 3 does not show the entity called *log* that stores all the events handled by the logger of the SIAR system.

3.2 Data Logic

The data logic component of the metadata management layer is constituted by the SIAR datastore that defines all the methods that the application logic may call on the data logic of the SIAR system. The SIAR datastore is independent from any particular *DataBase Management System (DBMS)* and is composed of several components called *Data Access Objects* (DAOs). The DAOs are used to abstract and encapsulate all access to the database; the DAO manages the connection with the database to obtain and store data. Essentially, the DAOs act as adapters between the components and the database. In Fig. 2 we can see that the data logic is composed by five DAOs: the metadata DAO, the set DAO, the user DAO, the group DAO and the logger DAO. Every single DAO defines all of the methods that have to be provided for managing the corresponding entity in the database.

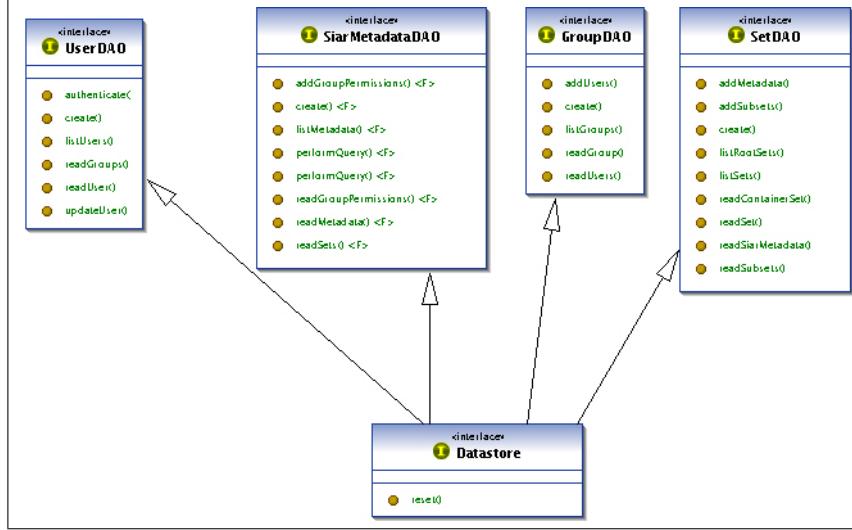


Fig. 4. DLS Data Logic: The SIAR Datastore

In Fig. 4 we can see the methods defined in the five DAOs of the SIAR datastore. For instance the Metadata DAO implements methods for creating or reading metadata, for adding a metadatum to a set, for listing all the metadata belonging to a set, for adding read and write group permission or for reading all the set to which a specific metadata belongs. In the same way Set DAO implements methods for creating a set and adding and reading metadata from the sets. Furthermore, Set DAO defines several methods for obtaining the hierarchy of supersets of a set or the list of subsets.

The other DAOs define the operations on the other SIAR entities.

3.3 Application Logic

The application logic is constituted by two components: the service manager and the web component manager. The service manager defines all the functionalities provided by the SIAR system; instead the web components manager defines all the methods that provide support for the web services implementation and elaborate external requests.

In Fig. 5 we can see how the service manager is composed: there is a service for each DAO in the datastore and an OAI-PMH service that implements the protocol functionalities. For instance the *user management service* provides the *authentication service* that permits the users authentication or the *reset password service* that allows a user to change his password.

The OAI-PMH service realizes the data and service provider features exploiting the other SIAR DLS services. The data provider component of this service answers the requests of external service providers; for instance it creates lists of

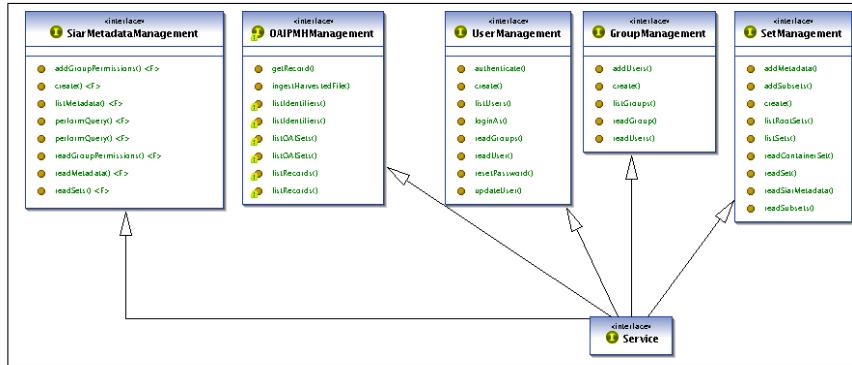


Fig. 5. DLS Application Logic: The SIAR Service Manager

metadata encoded in XML files or returns the set organization of the system formatted as an OAI-PMH response. The service provider component enables the harvesting of metadata from other repositories and their storage in the SIAR DLS. In the same way external services can also be developed and added to the SIAR.

4 Final Remarks

We presented the SIAR DLS architecture that enables the sharing of metadata between several archives in a flexible and scalable way. We explained the SIAR design choices, describing in detail the data model of the DLS represented by the metadata management layer.

Future work will concern the continuation of the development of the presentation layer composed by the user interfaces.

Acknowledgments

The study is partially supported by the TELplus Targeted Project for digital libraries, as part of the eContentplus Program of the European Commission (Contract ECP-2006-DILI-510003)⁴. The work of Gianmaria Silvello was partially supported by a grant from the Italian Veneto Region.

References

1. M. Agosti, G. Bonfiglio-Dosio, N. Ferro, and G. Silvello. *Metodologie e percorsi interdisciplinari per la ideazione di un Sistema Informativo Archivistico*. Memoria dell'Accademia Galileiana in Scienze, Lettere ed Arti in Padova, Ente di Alta Cultura (D.P.R. 27/10/49 n.1005), 2008.

⁴ <http://www.theeuropeanlibrary.org/telplus/>

2. M. Agosti, N. Ferro, and G. Silvello. An Architecture for Sharing Metadata among Geographically Distributed Archives. In C. Thanos, F. Borri, and L. Candela, editors, *DELOS Conference*, volume 4877 of *Lecture Notes in Computer Science*, pages 56–65. Springer, Heidelberg, Germany, 2007.
3. M. Agosti, N. Ferro, and G. Silvello. Proposta metodologica e architetturale per la gestione distribuita e condivisa di collezioni di documenti digitali. *Archivi*, 2(2):49–73, December 2007.
4. N. Ferro and G. Silvello. A Distributed Digital Library System Architecture for Archive Metadata. In M. Agosti, F. Esposito, and C. Thanos, editors, *Post-proceedings of the Forth Italian Research Conference on Digital Library Systems (IRCDL 2008)*, pages 99–104. ISTI-CNR at Gruppo ALI, Pisa, Italy, July 2008.
5. N. Ferro and G. Silvello. A Methodology for Sharing Archival Descriptive Metadata in a Distributed Environment. In B. Christensen-Dalsgaard et al., editor, *Proc. 12th European Conference on Research and Advanced Technology for Digital Libraries (ECDL 2008)*, pages 268–279. Lecture Notes in Computer Science (LNCS) 5173, Springer, Heidelberg, Germany, 2008.
6. K. Kiesling. Metadata, Metadata, Everywhere - But Where Is the Hook? *OCLC Systems & Services*, 17(2):84–88, 2001.
7. Society of American Archivists. Encoded Archival Description: Tag Library, ver. 2002. Society of American Archivists, 2003.
8. C. J. Prom. Reengineering Archival Access Through the OAI Protocols. *Library Hi Tech*, 21(2):199–209, 2003.
9. S. L. Shreeves, J. S. Kaczmarek, and T. W. Cole. Harvesting Cultural Heritage Metadata Using the OAI Protocol. *Library Hi Tech*, 21(2):159–169, 2003.
10. H. Van de Sompel, C. Lagoze, M. Nelson, and S. Warner. Implementation Guidelines for the Open Archive Initiative Protocol for Metadata Harvesting - Guidelines for Harvester Implementers. Technical report, Open Archive Initiative, p. 6, 2002.
11. H. Van de Sompel, C. Lagoze, M. Nelson, and S. Warner. The Open Archives Initiative Protocol for Metadata Harvesting (2nd ed.). Technical report, Open Archive Initiative, p. 24, 2003.

Combining Qualitative and Quantitative Keyword Extraction Methods with Document Layout Analysis

Stefano Ferilli¹, Marenglen Biba¹, Teresa M.A. Basile¹, Floriana Esposito¹

¹ Università di Bari, Dipartimento di Informatica,
via E. Orabona 4, 70126 Bari, Italy
{ferilli, biba, basile, esposito}@di.uniba.it

Abstract. The large availability of documents in digital format posed the problem of efficient and effective retrieval mechanisms. This involves the ability to process natural language, which is a significantly complex task. Traditional algorithms based on term matching between the document and the query, although efficient, are not able to catch the intended meaning of both, and hence cannot ensure effectiveness. To step on toward semantics, problems such as polysemy and synonymy must be tackled automatically by text processing systems. This work aims at introducing in the document processing framework of DOMINUS qualitative techniques based on the lexical taxonomy WordNet and its extension WordNet Domains for text categorization and keyword extraction, that can support the currently embedded techniques based on quantitative approaches. In particular, a density function is exploited to assign the proper importance to the involved concepts and domains. Preliminary results on texts of different subjects confirm its effectiveness.

Keywords: Lexical taxonomies, Text Categorization, Keyword Extraction.

1 Introduction

In the last years the amount of available documents in digital format has grown exponentially, which affects the retrieval of interesting and significant information at need (a problem known as “information overload”) and requires the development of proper techniques that improve the performance as regards amount and quality of the returned documents. Much of the document content is in the form of text (and hence unstructured)¹, which has been a significant motivation to the development of Natural Language Processing (NLP) techniques. The outcome of such techniques is the input to further processing aimed at indexing the documents and extracting information from them, in order to support information retrieval. In NLP two important application fields can be found: Information Retrieval (IR) and Information Extraction (IE). IR aims at selecting a relevant set of documents from a larger dataset, as an answer to a user query which expresses with a set of terms his information need. IE aims at identifying useful and relevant information that describes the content of a set of unstructured texts, and to report them in a (semi-)structured format suitable for

¹ In the following, a “document” will be intended as the set of text blocks contained in it, possibly labeled according to the role they play in the document.

filling a database or for exploitation by computers. Thus, the latter can be a significant aid to the former, in that basing document indexing (and hence IR) only on the relevant information preliminarily obtained by IE can improve the query result quality. For instance, it would be helpful to be able to search documents according to their content category, and later exploit keywords to better specify their content.

The objective of developing information extraction techniques that are able to catch the intended meaning of the text is very hard, due to peculiar ambiguities of natural language: synonymy, polysemy, phraseology (also known as *n*-grams), specific and technical terms [1]. For instance, the Italian word “Calcio” may stand for the name of a city, a chemical element, a part of a gun, or the action of kicking. In order to tackle the complexity of natural language, NLP tasks are usually divided in progressively higher-level and complex steps:

- *lexical analysis* breaks a text into tokens (usually corresponding to words);
- *syntactic analysis* organizes the tokens in a hierarchical structure according to their grammatical role;
- *semantic analysis* associates a meaning to the syntactic structure and thus, indirectly, to the text itself.

This work aims at extending the functionality of DOMINUS, a document processing framework, with text categorization and keyword extraction based on qualitative approaches applied to the lexical level only.

Text Categorization (TC) is the task of classifying words in natural language into specific categories belonging to a pre-defined set [2], or of assigning automatically a category to a corpus of documents. More formally, given a set of classes of interest and a set of documents already categorized in those classes (training set), the aim is building a decision function (classifier) that can map new documents (test set) to one or more classes according to their content. Hence, two main steps can be identified: learning and categorization. In the former, the system operates on the training set to learn information about the categories and the way to distinguish them, this way building a classifier. In the latter, using such a classifier new documents can be classified according to the given categories.

Keyword Extraction (KE) [3] is an information extraction task that aims at representing the essence of the intended message carried by the document according to the terms exploited. Two main approaches are present in the literature to tackle this problem. The quantitative approach assumes that a simple list of words included in the text can represent its subject. However, this approach, although providing a rough classification for the text, does not allow to organize it in sub-categories. To do this, semantics must be taken into account, that is the focus of the qualitative approach. This involves a semantic description of lexical objects in the text, that takes into account possible semantic domains and relationships, and can result in more specific and reliable outcomes than the quantitative approach.

In the following, after presenting the work that is at the basis of our experiment, the peculiarities of DOMINUS that make it suitable to such an integration will be introduced along with the experimental outcomes confirming the proposal viability.

2. Preliminaries

The increasing availability of linguistic resources as a support to NLP tasks, has allowed the proposed techniques approaching more closely the semantic level. One of the most famous and widely used such resources is WordNet, a lexical knowledge base designed to associate terms with a semantics based on groups of synonyms.

2.1 Ontologies

Born as a philosophical discipline, ontologies have developed in Information technology as exhaustive and rigorous conceptual schemata to formally describe a given domain. They have gained a fundamental importance with the spread of the Internet, since they represent a tool by which computer can exchange information based on its semantics rather than simple syntax. According to Tom Gruber, “*an ontology is an explicit specification of a conceptualization*” [4].

Lexical ontologies aim at characterizing (part of) a language independently of the domain, by expressing a lexical knowledge, made up of a set of words (intended as character strings), and a semantic knowledge, that encompasses word meanings and relations between words. WordNet [5, 6] is a famous lexical knowledge base aimed at overcoming the limitations of one-dimensional dictionaries. It groups terms in synsets (synonymous sets), partitioned into nouns, verbs, adjectives and adverbs. A polysemic term will belong to different synsets, and some relationships among synsets (such as hyperonymy, meronymy, etc.) are specified. Thus, WordNet is an outstanding candidate for supporting all tasks that are related to synonymy, in contrast to classical term-based representations of documents.

However, relationships expressed in WordNet are not exhaustive, so that some semantically related terms are not linked in it (e.g., 'doctor' and 'hospital'). Considering such connections in terms co-occurring in the same text would be of great help in stepping on from the syntactic/quantitative level to a semantic/qualitative one. In fact, domain labels (e.g., medicine, sports, etc.) are a powerful way to establish semantic relations between terms, and represent a fundamental semantic property on which text consistency is based, in the sense that terms in the same text tend to belong to the same domain(s). Hence, a limited number of terms (typically nouns) can determine the main domain of a text, and be of great help in text disambiguation; most terms, conversely, are not relevant because highly polysemic (typically verbs).

The *One Sense per Discourse* (OSD) hypothesis refers to the trend of terms in one discourse of having always the same sense. Conversely, the *One Domain per Discourse* (ODD) hypothesis assumes that term occurrences in a consistent portion of text tend to show the same domain. To defeat such an assumption, Krovetz [7] states that is sufficient that even one term in the same text does not fulfill it. While OSD seems not to hold, ODD does: in a text only a limited number of domains exists.

According to these principles, an extension of WordNet has been set up, called *WordNet Domains (WND)* [8, 9], that associates each synset to the corresponding domain(s), where a domain is intended as a set of words among which strong semantic relations exist. The domains are taken from a domain hierarchy made up of nearly 200 elements and inspired to the *Dewey Decimal Classification (DDE)* system

[10] used by librarians to categorize books. After manually setting some high-level synsets, the relationships already present in WordNet were exploited to automatically complete such an assignment for all synsets. The ODD hypothesis supports the use of WordNet Domain in text disambiguation, to identify the main domain in a text.

2.2 Density Function and Its Exploitation

In the following we recall the WordNet-based density computation according to the *density function* presented in [11]. Terms not included in WordNet (frequent words such as articles, pronouns, conjunctions and prepositions) are not evaluated for classification, this way implicitly performing a stop-word removal.

Given the set $W = \{t_1, \dots, t_n\}$ of terms in a sentence, each having a set of associated synsets $S(t_i)$, a generic synset s will have weights

- $p(S(t_i), s) = 1 / |S(t_i)|$ if $s \in S(t_i)$, 0 otherwise, in $S(t_i)$, and
- $p(W, s) = \sum_{i=1, \dots, n} p(S(t_i), s) / |W|$ in sentence W .

If a term t is not present in WordNet, $S(t)$ is empty and t will not contribute to computation of $|W|$. The weight of a synset associated to a single term t_i is $1 / (|W| \cdot |S(t_i)|)$. The normalized weight for a sentence is equal to 1.

Given a document D made up of m sentences W_i , each with associated weight $w_i > 0$, is $p(D, W_i) = w_i / (\sum_{k=1, \dots, h} w_k)$. The total weight for a document, given by the sum of weights of all its sentences, is equal to 1. Thus, the weight of a synset s in a document can be defined as:

$$p(D, s) = \sum_{j=1, \dots, m} p(W_j, s) \cdot p(D, W_j)$$

In order to assign a document to a category, the weights of the synsets in the document that refer to the same WordNet Domains category are summed, and the category with highest score is chosen. This Text Categorization technique, differently from traditional ones, represents a static classifier that does not need a training phase, and takes the categories from WordNet Domains.

For a successful exploitation of this technique, internal cohesion of the document is very important. Indeed, each sentence in the document conveys a portion of the information it refers to, and authors tend to limit redundancy in the whole text by means of cross-references among sentences. Thus, documents in which each sentence concerns a different topic will probably yield a uniform distribution of scores for the different categories, and prevent the identification of a unique dominant category to be assigned. However, this is not to be considered as a fault of the technique, rather as a case of problematic input in itself.

The technique proposed in [11] exploits information about the PoS of the terms to filter the relevant synsets only and thus improve weight computation. Since wrong assignment of the PoS tag to a term could negatively affect the weight computation (e.g., *marine* as a noun would denote the *military* domain, while as an adjective would indicate *biology*), we want to check the effectiveness of the technique without such knowledge.

3 Document and Text Processing in DOMINUS

DOMINUS (DOcument Management INtelligent Universal System) [12] is a framework designed to intensively exploit advanced Artificial Intelligence and Machine Learning techniques in automatic document processing. It covers all aspects and functionality involved in a digital library, from document acquisition to information retrieval, and particularly focus on the semantic aspects of the information it handles. A document submitted to the system goes through a number of steps that progressively acquire higher-level information, and specifically:

1. *acquisition*: documents in different formats are acquired and translated into a unique representation
2. *layout analysis*: the various components of the document pages are extracted and organized in a structure called layout hierarchy
3. *document image understanding*: the document is assigned to a layout class, and each component in it is associated to a label expressing its role
4. *text analysis*: text from components playing a relevant role is extracted, along with its grammatical and logical structure, and stored for future retrieval
5. *text categorization*: the document is assigned to a category expressing its domain of interest
6. *information extraction*: further information of interest for the specific domain is extracted.

Various steps include intelligent techniques that can be trained and later exploited automatically on new documents. A quality threshold is specified for each step, so that when the accuracy falls below such a threshold the system requires user confirmation before proceeding to the next step; in turn, the user intervention is exploited to improve performance on future cases and take the system above the threshold again.

Here, we focus on the last two steps, that are based on the outcome of step 4 that turns the text in a form that can be processed in a more easy and efficient way.

Specifically, after a *tokenization step* that aims at splitting the text into homogeneous components such as words, values, dates, nouns and a *language identification step*, DOMINUS also carries out additional steps that are language-dependent: PoS-tagging (each word is assigned to the grammatical role it plays in the text by means of a rule-based approach), Stopword removal (less frequent or uniformly frequent items in the text, such as articles, prepositions, conjunctions, pronouns, etc, are ignored to improve effectiveness and efficiency), Stemming (all forms of the same term are reported to a standardized form, this way reducing the amount of elements and highlighting word correspondences), Syntactic Analysis (yielding the grammatical structure of the sentences in the text) and Logical Analysis (providing the role of the various grammatical component in the sentences). For Italian, a rule-based procedure performs in a single step PoS tagging, stopword removal and stemming. For English, different modules are planned to perform these steps, of which only stemming is implemented, according to Porter's algorithm [13]. Hence, we wanted to check whether the qualitative technique described in Section 2 can work effectively even based on stemmed words only.

DOMINUS is very suitable to include a technique based on the density function presented above for various reasons. First of all, Step 3 of document processing can provide the kinds of components to be weighted differently by the density function. Second, the density function deals separately with each document, and hence fits the incremental behavior of **DOMINUS** more than statistical techniques that require statistics computed on the whole set of documents. Moreover, logical analysis of the sentence can provide phraseology to be considered as a whole instead of the single terms that make it up (although this is left to future work). Lastly, the application of the density function to text categorization and keyword extraction would cover two services that are currently needed in **DOMINUS**: indeed, the former is still missing, while the latter is currently carried out with a naïve Bayes technique [14] that could be complementary to the semantically-based approach.

The naïve Bayes technique is a quantitative method based on the concepts of frequency and position of a term and on the independence of such concepts. Indeed, a term is a possible keyword candidate if the frequency of the term is high both in the document and in the collection. Furthermore, the position of a term (both in the whole document and in a specific sentence or section) is an interesting feature to consider, since a keyword is usually positioned at the beginning/end of the text. Such features are combined according to the Bayesian Theorem in a formula to calculate the probability of a term to be a keyword in the following way:

$$P(key|T,D,PT,PS) = \frac{P(T|key) * \sum_{i=1}^{|insD|} P(D_i|key) * \sum_{j=1}^{|insT|} P(PT_j|key) * \sum_{k=1}^{|insS|} P(PS_k|key)}{P\left(\sum_{i=1}^{|insD|} D_i + \sum_{j=1}^{|insT|} PT_j + \sum_{k=1}^{|insS|} PS_k\right)}$$

where $P(key)$ represents the probability *a priori* that a term is a keyword (the same for each term), $P(T | key)$ is the standard *tf-idf* value of the term, $P(D | key)$, respectively $P(PT | key)$ and $P(PS | key)$, are computed by dividing the distance of the first occurrence of the term from the beginning of the section (D), document (PT), sentence (PS) with the number of the terms in the section, respectively document and sentence. Finally, $P(D, PT, PS)$ is computed by adding the distances of the first occurrence of the term from the beginning of the section, document and sentence. Since a term could occur in more than one document, section or sentence, the sum of the values are considered. In this way, the probability for the candidate keyword are calculated and the first k ($k=10$) with the highest probability are considered as the final keywords for the document.

On the other hand, in the qualitative method based on the density function computation, for each text block identified by **DOMINUS** in the document, the weight associated to the corresponding label and the terms extracted from it are exploited to obtain the *domain categories* to which the document content belongs and the set of *keywords* for the document. The extracted keywords can be exploited to support IR tasks, by computing query results according to keyword matching rather than complete text matching. For instance, given a document with keywords:

{jellyfish, invertebrate, marine, tentacle, cnidaria, ocean}

and that a user query contains the following terms:

{jellyfish, marine, invertebrate, medusa}

the overlapping between the query terms and the document keywords is 3 out of 4, which probably indicates the relevance of the latter to the former.

Concerning the exploitation of knowledge about the document logical structure, as provided by DOMINUS, in the density function computation, different weights will be assigned to the kinds of text blocks as a whole, depending on the role they play in the document, rather than to the single sentences. For instance, in the case of scientific papers the following labels could be considered of interest, along with sensible weights that express their relative importance: TITLE importance 4 - ABSTRACT importance 3 - BIBLIOGRAPHY importance 2 - BODY importance 1

The document weighting algorithm, after computing the density function, proceeds as follows:

1. sort the list of synsets in the document by decreasing weight;
2. assign to the document the first k ($k = 10$) terms in the document referred to the synsets with highest weight in the list, whose domain category is not “factotum”;
3. for each pair synset-weight create the pair label-weight where label is the one that WordNet Domains assignes to that synset
4. sort by decreasing weight the pairs label-weight;
5. select the first n domain labels that are above a given quality threshold.

Terms with category different from “factotum” are considered, since these have not a peculiar meaning or are frequently used (this improves the stopword removal process).

After assigning weights to all synsets expressed by the document under processing, the synsets with highest ranking can be selected, and the corresponding terms can be extracted from the document as best representatives of its content.

By exploiting WordNet to retrieve the synsets associated to words in the text, the density function identifies the most significant words that express the document subject. Each synset is associated to a weight from the density function, and the weight of the involved synsets is exploited to assign a weight to the overall categories and keywords for the document. Based on WordNet, the new functionality will extract information from documents based not only on the frequency of occurrence of the words they contain, but also on the possible concepts underlying those words.

4 Experiments

The proposed technique was implemented in Java 6.0, embedded in the document processing system DOMINUS and evaluated, for the task of Keyword Extraction and Text Categorization, according to the behavior, effectiveness and efficiency. All experiments were run on a PC endowed with an Intel Core 2 Duo T7200 2.0 Ghz processor and 1 GB RAM, working under Windows XP Professional.

For evaluation purposes, we built a small dataset made up of 5 documents, concerning very different domains and subjects, as specified below:

1. electronic computers, their birth and evolution;

2. child education from birth to adolescence;
3. jellyfish, their habitat and their main features;
4. rubber and its chemical feature;
5. rugby, with references to its rules and history.

For each document, the following figures report the extracted keywords and the corresponding weight computed through the density function. The best ranking keywords are reported, and currently only those having weight greater or equal to 0.03 are selected.

Document	1	2	3	4	5
Length	1317	1167	607	176	834
Runtime KE	13.2"	12.1"	8.4"	5.5"	10"
Runtime TC	2'17"	2'05"	1'35"	37"	1'57"
1	computer 0.04834439	teaching 0.021461325	jellyfish 0.03463425	synthesized 0.05172414	rugby 0.046938974
	electromechanical 0.011299435	education 0.021461325	tentacle 0.03429106	rubber 0.024820872	team 0.013085229
	Neumann 0.010788882	instruction 0.021461325	creatures 0.008853416	Hevea 0.023824453	goal 0.011058885
	instructions 0.006961259	pedagogy 0.021461325	animal 0.008853416	Kuhn 0.020689657	minutes 0.008610792
	transistor 0.004842615	training 0.016265217	zooplankton 0.008333334	latex 0.011050157	football 0.006467662
	store 0.00454148	instructor 0.009772334	digestive 0.0074955914	colloidal 0.0103448285	score 0.0063810167
	wartime 0.004237288	teachers 0.009772334	invertebrates 0.0069444445	entropy 0.0103448285	hit 0.0063810167

By comparing the keywords extracted to the documents content, a significant overlapping is evident, suggesting high effectiveness of the technique. Indeed, it can identify important terms even when they are specific to the domain discussed in the document (e.g., Cnidaria in Document 3, indicating a class of jellyfish). The first keywords in the ranking are sufficient to identify the domain, while the others are useful to complete and refine the idea about the document content. Runtime, as expected, increases along with the document size, and is quite fast.

As to Text Categorization, the following table reports the identified categories and the corresponding weight computed through the density function.

Doc	1	2	3	4	5
	computerscience 0,074179690	pedagogy 0,13590841	animals 0,145654480	chemistry 0,076343600	play 0,125770640

	mathematics 0,041016333	school 0,056096878	biology 0,134478210	pure_science 0,051724140	sports 0,077846320
	mechanics 0,026428163	politics 0,033727642	anatomy 0,075752420	animals 0,044569080	rugby 0,046938974
	time_period 0,026385480	sociology 0,029268516	gastronomy 0,024342252	biology 0,044454810	person 0,030388908
	person 0,025812928	administration 0,025442114	person 0,023609525	plants 0,036958255	music 0,029128496
	industry 0,024380295	number 0,024071350	chemistry 0,022065999	physics 0,034578360	music 0,021006696
	geography 0,015836516	person 0,022174576	geography 0,017274980	industry 0,033254784	animals 0,018921590
	publishing 0,015047456	university 0,020350550	food 0,014234459	fashion 0,029777769	biology 0,016300263
	buildings 0,014745613	geography 0,017418027	military 0,011330307	sexuality 0,024820872	telecommunic 0,013834923
	art 0,012498402	biology 0,015105671	economy 0,008133663	music 0,017505657	football 0,012883367

Looking at the results, it is evident that the system always succeeds in catching the proper domain category the document deals with. However, differently from the keyword extraction task, the ranking of short documents is shorter, and thus errors in assigning categories are possible due to closely weighted domains (e.g., in document 4). The threshold for assigning a domain to a document was empirically set to 0.03. A lower threshold would tend to provide a single-label classification, whereas a higher threshold would include wrong categories as well. With such a bias, Document 3 would be assigned to categories *animals* and *biology*, while Document 4 would be assigned to categories *chemistry* and *pure_science*.

Runtime is again proportional to the document length, but, in this case, as expected, is much slower. However, this task is carried out only once for each document separately, and the prototype has not been optimized. The main cause of such a behavior is the choice to scan the whole synset map in order to extract the domain categories and rank them by importance. Indeed, such a map is quite large. If the system is to be exploited for web applications/purposes, however, such runtimes are not acceptable, and some alternative way must be found to compute the ranking. A solution can be scanning only a portion of the whole map, in which case such a proportion must be properly defined in order to obtain results similar to the complete case, with a minimal quality decay. Empirical tests showed that using only 1/20th of the whole map yields imprecise and misleading results, particularly for document 5 that is very short. Thus, various experiments led us to empirically set the threshold to 1/10th, so that the results are similar to the original ones also for short documents. Indeed, the results for a short and a long document are reported in the following table, to highlight the different behavior in the two cases.

Doc	1		4	
	<i>whole synset map</i>	<i>1/10th synset map</i>	<i>whole synset map</i>	<i>1/10th synset map</i>
computer_science 0,074179690	computer_science 0,067159660		chemistry 0,076343600	chemistry 0,068965520
mathematics 0,041016333	mathematics 0,028364072		pure_science 0,051724140	pure_science 0,051724140
mechanics 0,026428163	mechanics 0,024066502		animals 0,044569080	animals 0,041827142
time_period 0,026385480	industry 0,019222680		biology 0,044454810	biology 0,029852908
person 0,025812928	time_period 0,015588739		plants 0,036958255	plants 0,029780567
industry 0,024380295	person 0,011930388		physics 0,034578360	fashion 0,024820872
geography 0,015836516	electricity 0,010278916		industry 0,033254784	industry 0,024820872
publishing 0,015047456	geography 0,010015408		fashion 0,029777769	sexuality 0,024820872
buildings 0,014745613	electronics 0,009804817		sexuality 0,024820872	physics 0,022642877
art 0,012498402	buildings 0,009618105		music 0,017505657	music 0,012056910

Document 1 took 13 sec to accomplish the task and yield the correct category *computer_science*. Document 4 took 8 sec to yield the results, assigning the document to *chemistry* and *pure_science* as in the previous experiment (additionally, category *animals* was introduced, due to references to the animal and vegetal reigns that get a higher importance in a short document).

In large documents, the dominant category becomes more neatly separate from the others, since only more important synset are considered for defining the ranking. Runtimes are neatly reduced, as desired.

A further experiment was carried out that aimed at evaluating both the qualitative and quantitative approaches embedded in DOMINUS, and in order to better test and understand the way the two approaches can be integrate. Thus, we built a dataset made up of 100 documents equally distributed on the following 10 categories: Architecture, Astronomy, Biology, Chemistry, Computers, Economics, Geography, Law, Oceanography-Meteorology, Religion. All the documents were downloaded from the web and in particular from some university libraries.

On each document (only the first page of the documents was used) both techniques were applied requiring the extraction of 10 keywords for each of the methods. Among the 1000 keywords that were extracted from the 100 documents, 518 are the same for

both the techniques on 99 documents, i.e. an average of 5.23 keywords in common were extracted (in one case no common keywords were extracted). However, we noted that in many cases, the keywords that are not in common can be used to complete the set of keywords identified by one of the two techniques. Specifically, the quantitative method, that is based on the naïve bayes technique, completes the document description with keywords that describe the topic of the document at a quite general level. On the other hand, a more detailed and specialized description of the document content can be obtained exploiting the quantitative method, that is based on the density function computation. For example, for a document on Oceanography-Meteorology topic, the quantitative method found spring, pacific, eastern, season that are more general than the keywords extracted from the quantitative method, i.e. ship, measurement, compare, maintenance.

Thus, an integration of the two methods is able to better define and identify the document content by mixing the generalized and specialized topic description that each technique is able to grasp. As regards scalability, for the qualitative method this is not an issue since in that case the keywords are computed for each single document with respect to WordNet alone, independently of all the others. For the quantitative method, based on $tf*idf$ and hence on the entire collection of documents, since DOMINUS stores in a relational database statistics concerning all tf and idf values of the terms, it is sufficient to update them incrementally, when a new document arrives, for the terms appearing in that document alone. We plan to perform extensive experiments regarding this problem by using larger collections of documents.

5 Conclusions

This work has presented an extension of the functionality of DOMINUS, a prototypical system for intelligent document processing based on Artificial Intelligence techniques, with qualitative approaches to text processing, based on the semantics of terms rather than on their number of occurrences only. Specifically, two lexical resources (WordNet and WordNet Domains) and a particular density function defined on them have been exploited to transform a document into a weighted map of synsets that describe it conceptually, thus supporting qualitative techniques for text categorization and keyword extraction. DOMINUS is particularly suited to such a technique since it can provide the role each piece of text plays in a document, can process each document separately in an incremental way and already provides quantitative techniques for keyword extraction that can be complemented by the new approach. Experimental results on both tasks are satisfactory, both for accuracy and for effectiveness.

Future works will concern defining a strategy for the selection of keywords and categories when more than one are required (also taking into account generalization relationships among them), and exploiting the extracted information to improve information retrieval and to build or refine specific domain taxonomies.

References

- [1] T. De Mauro. Il dizionario della lingua italiana. www.demaeuroparavia.it
- [2] F. Sebastiani (2002) "Machine Learning in Automated Text Categorization" ACM Computing Surveys, Vol.34 N.1, pp. 1-4. <http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.17.6513>
- [3] L. Hunyadi (2001) "Keyword extraction: aims and ways today and tomorrow". Lajos Kossuth University, Hungary, pp.1-6. www.keyword.kcl.ac.uk/redist/pdf/hunyadi.pdf
- [4] T. Gruber (1995) "Toward Principles for the Design of Ontologies Used for Knowledge Sharing". International Journal of Human-Computer Studies, Vol. 43, No. 5-6. pp. 907-928.
- [5] G. A. Miller, R. Beckwith, C. Fellbaum, D. Gross, K. Miller (1990) "Introduction to WordNet: An On-line Lexical Database". International Journal of Lexicography, Vol. 3, No. 4, pp. 235-244. wordnet.princeton.edu/5papers.pdf
- [6] C. Fellbaum (1998) "WordNet an Electronic Database", Cambridge: MIT Press. pp. 1-23.
- [7] R.Krovetz. (1998) "More than one sense per discourse". Technical report, Princeton, NEC Research Institute. Proceedings of SENSEVAL Workshop, Herstmonceux Castle, UK, pp.1-10. citeseer.ist.psu.edu/185217.html
- [8] B.Magnini, C.Strapparava, G. Pezzulo, A. Gliozzo (2002) "The role of domain Information in Word Sense Disambiguation". Natural Language Engineering, Vol. 8, No. 4 , pp. 359-373. www.istc.cnr.it/doc/1a_16p_Magnini-NLE-2002.pdf
- [9] B. Magnini, G. Cavaglià (2000) "Integrating Subject Field Codes into WordNet". ITC-irst, Proc. Second International Conf. Language Resources and Evaluation, LREC2000, pp.1-6.
- [10] (2006) "Dewey Decimal Classification" http://www.library.und.edu/research/handouts/027_Dewey_Decimal_Classification.pdf
- [11] M. Angioni, R. Demontis, F. Tuveri (2008) "A Semantic Approach for Resource Cataloguing and Query Resolution", Communications of SIWN, ISSN 1757-4439, Vol.5, pp. 62-66.
- [12] F. Esposito, S. Ferilli, T.M.A. Basile, N. Di Mauro (2008) "Machine Learning for Digital Document Processing: From Layout Analysis To Metadata Extraction" - Machine Learning in Document Analysis and Recognition 2008, pp. 105-138.
- [13] M.F. Porter (1980), "An algorithm for suffix stripping". Program, vol.14, N.3, pp. 130-137. <http://tartarus.org/~martin/PorterStemmer/def.txt>
- [14] Y. Uzun, "Keyword Extraction Using Naïve Bayes", Bilkent University, Department of Computer Science, Turkey www.cs.bilkent.edu.tr/~guvenir/courses/CS550/ Workshop/Yasin_Uzun.pdf

Handling Evolution in Digital Libraries *

Andrea Baruzzo, Paolo Casoto, Antonina Dattolo and Carlo Tasso

Department of Mathematics and Computer Science - University of Udine, Italy

{andrea.baruzzo, paolo.casoto, antonina.dattolo, carlo.tasso}@dimi.uniud.it

Abstract. Developing and maintaining a digital library requires substantial investments which are not simply a matter of technological decisions, but include also organizational aspects (which user roles are involved in content production, which workflows are needed, and so on). Moreover, starting a digital library initiative requires to handle several evolution issues (the need of new roles, workflows, and types of contents, the availability of new applications to integrate on the top of digital archives, etc.). To catch all these aspects, we outline a conceptual model based on three complementary domains: informational, technological, and social. This model tackles the typical issues affecting a digital library, especially concerning the evolution (i.e. change) of content, infrastructure (software tools), user roles, and related workflows in content creation, publication, and exploitation. These issues are addressed in the model by means of three elements: a suitable XML data metamodel, a service-oriented architecture, and a multi-agent infrastructure.

The XML metamodel abstracts the physical representation from the logical definition of data, making easier future changes. The multi-agent infrastructure helps to preserve the consistency of the stored archives when their schemata need to be changed. Finally, the service-oriented architecture simplifies the integration of new applications at the top of the digital library. As part of this architecture, we describe in particular a specific component: the PIRATES framework. This module introduces in the digital library a set of semantic services aimed at assisting final users to select from the archives the most appropriate content. Integrating semantic aspects helps to handle the evolution of both contents and user needs (i.e. interests). Techniques of user modeling, adaptive personalization, and knowledge representation are exploited to build the PIRATES services in order to fill the gap existing between traditional and semantic digital libraries.

1 Introduction

Data preservation in digital libraries is often addressed mainly by means of reliable storage mechanisms, and long-term accessibility of digital supports, both aimed at ensuring that library's contents will remain sustainable, authentic, accessible, and understandable over time. *Data evolution*, at the same time, is addressed by means of scalability of the *physical* system, concerning the modification of the stored information either in data formats or in space needed to archive them. Following this vein, many works in literature consider preservation and evolution issues mainly as technological factors

* The authors acknowledge the financial support of the Italian Ministry of Education, University and Research (MIUR) within the FIRB project number RBIN04M8S8.

(e.g. [3]). In this paper, we take a wider approach, recognizing also the role of other important aspects outside the technological domain. In particular, we explicitly address the *semantic aspects* of a (textual) content stored in a digital archive. We agree with Ross in claiming that digital preservation is more than keeping the streams of 1s and 0s used to represent information [20]. Preserving information is about maintaining the *semantic* meaning of both the digital object and its content, maintaining its provenance and authenticity, retaining its interrelatedness, and securing information concerning the context of its creation and use [21].

Moreover, a digital library is more than the sole data stored in its archives: it serves as an infrastructure to publish, retrieve, and access information fulfilling the final user needs. The user community can be huge and heterogeneous, thus *social aspects* can also play a critical role in the evolution of a digital library. For example, content production and exploitation processes involve a wide range of different users. The way users exploit their respective tasks may change over time and with experience, as well as with the upcoming of new information needs. The digital library infrastructure should then be aware of the evolution related to social aspects in order to provide users with the ability to easily update the way information is generated, accessed, aggregated, classified, and delivered. The effort needed to align this infrastructure with the evolution of user requirements is not trivial. New requirements often demand for new functional services which must be integrated into the legacy architecture in order to be effective. Hence, handling evolution aspects at the requirement level has also an important but often not recognized impact of the architectural level.

In this work we introduce a characterization of a digital library addressing specifically the evolution of both digital content and related services, taking into account the aforementioned considerations. This characterization integrates three complementary domains: *social*, *technological*, and *informational*. These domains can be incorporated into a conceptual model, as discussed in Section 4. We believe that providing such model facilitates the undertaking of the mentioned aspects. Hence, this formal model takes into account both *physical* and *semantic* evolution of the archived content, and the way such information is exploited by final users. In particular, semantic concerns are handled in the model at the architectural level by a specific component, the PIRATES framework, as described in Section 6.1. Such “semantic layer” is a first step in the direction of fully supporting the semantic digital library vision [16].

This paper is based on the results of an experimentation within the EU-India E-Dvara project¹. This project is concerned with the development of a digital platform devoted to e-content management in Indian heritage and sciences. In previous works, we presented the overall project goals [11], the technical details concerning data representation metamodels, and the general software architecture [4–6]. Here, we extend previous works, focusing especially on the characterization of a digital library according to its evolution aspects. In particular, after a brief survey of related work, we extend the ideas proposed in [6], presenting a conceptual model to handle the evolution of digital archives along multiple dimensions (Section 4 and Section 5). Then, we discuss a set of semantic, adaptive, and personalized services which can be introduced on the

¹ <http://edvara.uniud.it/india>

top of a digital library, supporting final users in retrieving, annotating, classifying and organizing the information archived in the library (Section 6).

2 Related Work

In the last few years several research projects have been proposed in order to cope with data preservation and organization [7, 18, 8]. Storage of XML-based documents has been proposed in Greenstone [2, 26], a digital library designed to provide librarians with the ability to create and publish heterogeneous collections of digital contents on the Web like text, images, videos and e-books. Each content in Greenstone can be described using *metadata* compliant with a standard schema (e.g. Dublin Core²), either imported or manually provided by librarians. However, Greenstone does not provide any role for managing the content submission process. Moreover, it does not provide functionalities concerning the evolution management of both contents and collection templates.

D-Space [24] is a digital library aimed at providing long-term preservation of heterogeneous contents, by improving some of the limitations affecting Greenstone. Authors usually submit their documents to the system, and define metadata for them. D-Space introduces also a multi-roles approach to content publishing, identifying the following actors: *authors* and *organizations*, which provide the contents, *librarians*, which perform content validation, and *users*, which are interested in content retrieval. Content-based workflows can be customized in order to cope with the needs of specific organizations, to structure content and to delegate proper activities to different stakeholders.

In order to provide a flexible and reusable solution to data preservation and organization, the Fedora Project [17] explored a service-oriented approach to data interoperability in digital libraries, by designing and developing a distributed architecture for contents publishing, aggregation, and retrieval. Composite information is obtained by aggregating physical contents, viewed as bit-streams, located worldwide into the Fedora repositories. Fedora allows content editors and archivists to define semantic connections between archived contents, treated as set of physical contents.

Other works related to content preservation in digital libraries are described in [7, 18]; the aDORe project, in particular, adopts the MPEG-21 DID content representation model to provide preservation and retrieval of heterogeneous multimedia contents.

The above mentioned systems are centered on contents, defined as *binary resources* enriched by metadata devoted to preservation, storage and retrieval purposes, but not intended for data structuring. Preservation and evolution of a data model in those approaches is implemented as a low-level mechanism, where data is processed as bit-streams instead of as instances of well-defined structures (i.e. XML Schema).

3 The E-Dvara Project

In this section we introduce the E-Dvara digital library. We examine its initial requirements and the results of an experimentation activity performed in the last three years. Building from these requirements, and overcoming some of the criticalities emerged

² See for more details: <http://dublincore.org/>

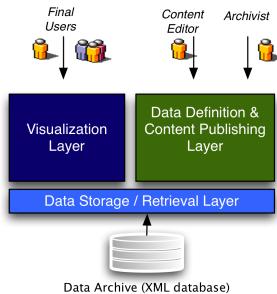


Figure 1: The high-level E-Dvara software architecture

during experimentation, we formulate the proposal for a new version of the digital library which is discussed in the rest of the article.

E-Dvara is a research project which represents our current development and experimentation in the area of digital libraries. E-Dvara is focused on the development of a new platform for the access and storage of digital contents [11]. Since its inception, it has been explicitly designed to overcome several limitations that characterize the process of building digital content. In particular, E-Dvara was initially meant to:

1. reduce the effort required by the archivist to define the data structure used to represent data into the platform;
2. provide to archivists with no expertise in data management a set of wizards devoted to data schemata creation in a completely automatic and transparent way (with respect to the physical database);
3. allow content providers to easily share their archives on the Web by means of built-in Web interfaces or with several dedicated applications, allowing archivists and system administrators to define the way data should be displayed to final users.
4. allow archivists to provide for each archive of digital contents a specific visualization template and a set of search forms.

In order to cope with these requirements, the E-Dvara platform has been designed adopting a three level modular architecture, illustrated in Figure 1. This architecture is constituted by a core layer devoted to data storage and persistence, a set of tools for data definition and content publishing, and a visualization layer devoted to content delivery and rendering.

Digital contents are represented and stored as XML documents, in order to fulfill both the requirements related to data interoperability and separation of concerns between data representation and data visualization. Data visualization is achieved by means of an XSLT engine, which can transform each XML document into HTML, PDF and WAP files, according to a specific style sheet defined by the archivist.

Data definition and content publishing are provided by a set of Web applications devoted to archivists and content publishers; archivists can easily create new data schemata, defining the required simple or complex fields, their multiplicity and data types. Con-

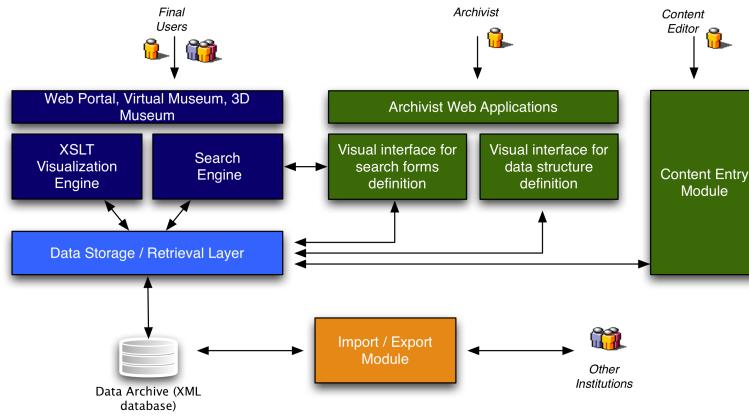


Figure 2: The architecture of the first prototype of E-Dvara platform

tent providers, according with data schemata created by archivists, can publish their data using Web forms automatically generated from the selected data structure.

The third layer concerns the visualization and delivery of the available contents to end users which can access data using either a Web browser or a WAP browser on a mobile phone. Data visualization is exploited automatically according to the device employed, the user preferences, and the style sheets provided by the archivists. Visualization layer is also responsible for providing to final users the search forms defined, for each archive, by the archivists. Search forms are represented as XML documents too. In Figure 2 an overall representation of the original architecture of E-Dvara is provided; the XML archive is the core component of the proposed architecture, on which the three described layers are based. A prototype of the E-Dvara platform was developed in 2005.

In the last three years, E-Dvara has been largely tested by expert users involved into professional content publishing for cultural heritage. From this experimentation, we have identified several problems with the first prototype. In particular, the evolution issues, the technical weaknesses, and the mistakes done led us to formulate the conceptual model and a new software architecture, both of which are discussed in the next sections of this paper.

4 Modeling evolution

To better handle the evolution aspects of a digital library, we propose an explicit conceptual model. We discuss here its main characteristics and the evolution issues which have inspired it.

4.1 The conceptual model

Our conceptual model is inspired by that provided by Yates [27], and incorporates the vocabulary suggested by Rowlands-Bawden [22], which is more suited to describe mod-

ern digital environments. Moreover, we introduce the concept of *evolution dimension* [6], which best addresses the evolutionary nature of a digital library. This new model highlights three domains:

- The *Informational domain*, which describes the knowledge organization and description (e.g. metadata) of digital archives.
- The *Technological domain*, which describes the knowledge organization and discovery provided by an appropriate technological infrastructure (e.g. software agents), the technical impacts on information transfer chains, and technology factors such as human-computer interactions.
- The *Social domain*, which describes human and organizational factors, information laws and policies, social impacts on the information transfer chain, and library management concerns.

On the basis of these domains, we define three specific *evolution dimensions* by means of a pairwise combination of domains. In fact, we recognize a mutual influence among domains when we consider evolving aspects in a digital library. More specifically, the open issues faced during our experimentation with E-Dvara may be classified along three dimensions:

1. *Informational-Technological dimension*, which identifies all *data evolution* problems due to changes in the underlying data model (e.g. the invalidation of entire archives of documents that conform to the old schema version).
2. *Technological-Social dimension*, which identifies problems concerning the need to adapt the technical infrastructure of a digital library in order to fulfill new user requirements (e.g. the integration of heterogeneous services to support the interaction with new user roles).
3. *Social-Informational dimension*, which concerns the diverse workflows needed to support the activities of such different community of users, and their impact on documents (e.g. a virtual museum curator has to describe the items of a document taking into account constraints imposed by user interfaces in order to effectively show tool-tips when a visitor moves the mouse over a particular exhibit in the scene). New roles can have a different view of documents, so the digital library should provide them the information required with formats suitable to their needs.

For each evolution dimension, we propose a conceptual model element which links together two model domain, as illustrated in Figure 3. We discuss in depth each element in Section 5.

4.2 Three classes of evolution problems in digital libraries

In this section we briefly summarize three classes of open issues concerning the evolution of digital libraries that we have identified during our experimentation with the first prototype of E-Dvara. In this context we also introduce the solution to each issue we are proposing in the new prototype which is under construction. Such solutions are detailed in Section 5. A more exhaustive presentation of the problems with representative examples is provided in [6].

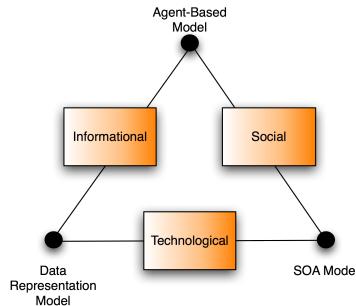


Figure 3: The evolution conceptual model

The data evolution problem The first prototype of E-Dvara provides users a flexible way to define and update the metadata associated to each digital archive. In particular, users can define a set of *schemata* which, associated with a specific archive, supplies the structure adopted for storing documents. However, metadata definition can take place during the whole life-cycle of the digital collection, leading to the problem of correctly handle the evolution of data. Such an approach requires the introduction of methodologies devoted to perform *data validation* accordingly to the schema evolution. In fact, each schema update should be properly spread to the previously validated archives, in order to automatically adapt the existing content to the new schema without introducing potential inconsistencies between data. Clearly, in order to both handle these issues and grant the preservation of yet existing contents, a set of dedicated tools should be provided. Hence, in the new prototype we are integrating the concept of *mutable templates*. To understand what is a mutable template, consider Figure 4 which describes the reference data model included in our conceptual model (a complete description of such model follows in Section 5.1). The main idea is to introduce flexibility by separating the data definition (in level M1) with the description of the corresponding types (in level M2). Furthermore, we would allow more than a single possible map between data definitions and types in order to cope with changes in both data formats and structure (e.g. consider the same set of bibliographical items which can be showed to different users using different formats). This is exactly the purpose of mutable templates: to allow type variations in data mappings. The evaluation we performed in E-Dvara has suggested us that the flexibility of *mutable templates* (i.e. evolving schemata) must be considered as an essential feature for our platform.

Technical infrastructure adaptation Other issues we have faced is the need for integrating new heterogeneous modules at the top of the digital library (e.g. virtual museums, meta-search engines, or applications for mobile devices). These requests posed unforeseen challenges on the software infrastructure. For each new application, we needed to rewrite a lot of ad-hoc business logic, without mentioning the fact that we had to duplicate some system services in order to adapt them to a new programming interface.

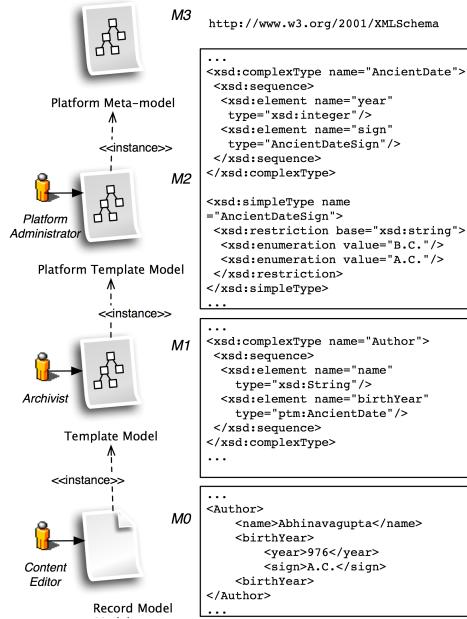


Figure 4: XML Representation of Data Model

Another tricky aspect was sharing similar behaviors and data located in different components which were based on incompatible communication protocols. These issues clearly demanded for a *reusable integration layer* that we lacked as part of our initial software architecture. Moreover, we have learned also that integrating several applications in a common environment requires a substantial investment in understanding and implementing their *orchestration*, in order to handle incompatibilities between different business logics in a standard and transparent way. In conclusion, in our first prototype we failed to recognize the importance of these issues, which are the building blocks to describe the interaction between different applications. If we had realized that since the beginning, composing together heterogeneous functionalities would be simpler to achieve. Furthermore, also the resulting deployment would be expedited, especially by moving the orchestration logic from the inside of a component to an external configuration file (i.e. by means of a XML description file associated to each component), enabling a flexible and dynamic setup.

Coordination issues of different user roles and workflows The continuous integration of different applications on the top of the existing software infrastructure were typically a manifestation of new requirements involving user roles and information access policies. An example of this situation is an external service which, based on its own data management policy, defines *when* a particular workflow is required to organize the archived contents. In E-Dvara, a *workflow* expresses a set of roles, related activities, and constraints that define together the structure of the information manage-

ment process. As a typical workflow, consider the curator of a digital museum which has to arrange a new gallery, composed by paintings, ancient books, and movies hosted in three archives, respectively, and owned also by three different archivists. When the curator wants to incorporate in this gallery a set of features to search, organize, and enrich the existing records, he may want to add new fields describing the position that each item should have in the virtual museum scene. Moreover, final users may also improve the exhibition quality, by creating new relations between existing contents (e.g. opinions and links to a specific related content in a typical Web 2.0 style). All these scenarios pose many issues that must be faced to provide flexibility in data management. Such issues concern *intellectual property* (Who is allowed to use/modify a content?), *coupling between archives* (How do the schemata of archive B evolve according to the evolution of both schemata and data in archive A, if any dependency between A and B exists?) and *coupling between workflows* (Do the activities in the workflow A overlap those in workflow B?).

5 Handling evolution of content production workflows

This section describes three conceptual model elements (the XML metamodel, the service-oriented architecture, and the multi-agent infrastructure), each one associated to a specific evolution dimension introduced in Section 4.1. Moreover, we describe also how these model elements are characterized by mutable templates, integration layers, and workflows management which are discussed in previous section.

5.1 The Informational - Technological dimension

In order to handle the evolution problems concerning the changes in data format and schemata described in Section 4.2, our conceptual model is based on a four-layer data representation model (Figure 4). At the bottom of the hierarchy, we place the *records* (level M0, Record Model), aimed at representing the archived data (documents). A record is an instance of a document stored in the digital platform. Every document must also conform to a *document template* (level M1, Template Model), which provides structural definitions (e.g. the document contains the Title, Author, and Date fields) and constraints (e.g. the Data field must conform to the mm/dd/yy format or the Title field is mandatory). Document templates are themselves conformed to *platform template* (level M2, Platform Template Model) devoted to define both business rules and data types the archivists can use to build document templates (e.g. each record in every archive must contain the Creation Date and Owner fields). Finally, platform templates are instances of a more general layer, the *platform meta-model* (level M3, Platform Meta-Model), which defines a set of common low-level structures (e.g. primitive data types as xsd:String) and operations (e.g. data sequencing) available in order to define more complex data structures. This level corresponds to that of the OMG XML Schema specifications³.

The overall data model involves the interaction with three different actors:

³ <http://www.w3.org/XML/Schema>

- *Content editor*, devoted to data entry, with respect to a specific document template; however, he is not allowed to perform any template change.
- *Archivist*, devoted to document templates definition.
- *Platform administrator*, devoted to the management of platform templates (e.g. the templates provided by archivists should be validated against the platform template model each time they are created/modified or when the platform template model itself is updated).

This hierarchical data model provides *automatic data validation policies* which play a central role in our vision. Indeed, validation is applied both to the templates and (recursively) to all the records stored in the platform archives. Templates which do not respect the business rules defined in the platform template model should be manually updated by either archivists or content providers in order to become consistent. This type of validation is propagated then to the platform meta-model (level M3) which acts as a template for the platform template model (level M2). In order to develop the proposed model, we present an implementation approach based on the XML technology and standards, focusing our attention on the features provided by XML Schema.

5.2 The Technological - Social dimension

Sometimes content production workflows change because new actors emerge, playing roles that was not foreseen in advance. Developers of digital libraries are then forced to consider new requirements when it is more expensive to integrate them in an existent technical infrastructure. Hence, to handle evolution issues concerning the adaptation to such new requirements (e.g. the integration of heterogeneous services described in Section 4.2), we conceive the second prototype of E-Dvara according to a Service-Oriented Architecture (SOA) model (Figure 5), characterized by:

- The introduction of an explicit *Integration layer*, which forms the “architectural glue” that brings the digital library beyond the scope of a single application, unifying the interfaces of different subsystems into the same interoperable environment.
- The migration toward *services*.
- The adoption of a *peer-to-peer; message-based communication protocol* supported by the *Enterprise Service Bus* (ESB)

The standard set of Web service technologies (XML, SOAP⁴, WSDL⁵) provides the means to describe, locate, and invoke a Web Service. However, it is often necessary to compose different services with a specific business logic in order to complete a task. This is where orchestration plays a crucial role, deploying sophisticated and complex Web services as a single, whole functionality. Thus, the orchestration engine (the ESB component) acts as a centralized authority to coordinate interaction between services and applications. At the top of our SOA architecture we have applications such as administration interfaces to manage users and archives, publication interfaces to produce new content in the digital library, or virtual museums to exhibit a document

⁴ <http://www.w3.org/TR/soap>

⁵ <http://www.w3.org/TR/wsdl>

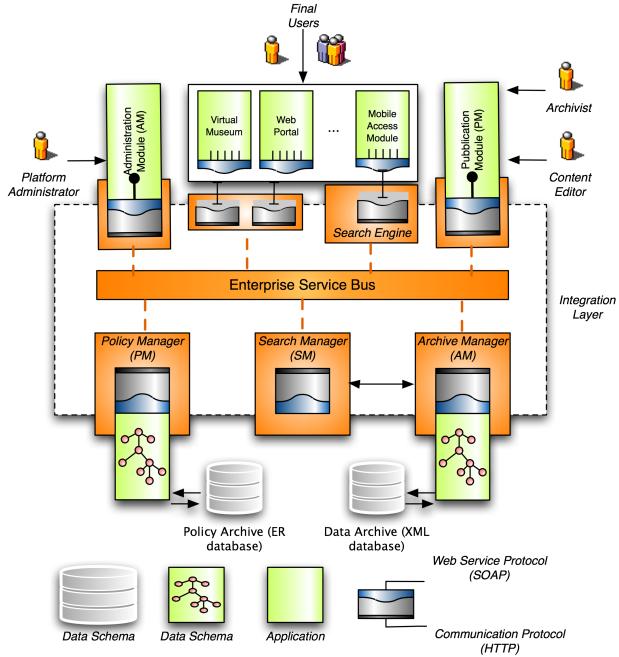


Figure 5: The architecture of the second prototype of E-Dvara

archive in a “museum-like” setting. All these heterogeneous modules can exploit any reusable service available in the Integration layer, (e.g. to perform searches in the platform archives). Finally, the archives are placed at the bottom of the architecture, which are managed by two custom applications: the *Archive Manager* which stores and retrieves documents, and the *Policy Manager* which manages users, accessing policies, and archives. The Archive Manager isolates the business logic needed to realize the data-model described in Section 5.1, whereas the Policy Manager implements the data validation rules, decoupling them from other architectural components.

5.3 The Social - Informational dimension

The introduction of mutable templates in content representation provides the ability to update a schema during the whole life-cycle of a data collection, but leads also to several challenges such as the evolution and re-validation of existing archives. In this section, we introduce a *multi-agent approach* to tackle the problem, aimed (when possible) to automatically resolve evolution issues.

The levels from M1 to M3, proposed in Figure 4, can be affected by updates during the digital library life-cycle. In particular, such updates can involve XML Schema definitions (level M3, with a low frequency), Platform Template Models (level M2, with a low-medium frequency) and Template Models (level M1, with a rather high frequency). Each schema is connected by a dependency link with the schemata on its top for val-

idation purposes. However, in a collection one level can be related to another also by means of relations between different data types (e.g. an instance of the template `Book` in M1 can be related with one or more instances of the template `Author` in M0). At the same time, we can also have a relation connecting templates in different collections (e.g. instances of the template `GalleryRoom` in a virtual museum application can be related with instances of `Book` and `Painting` templates taken from different collections). Hence, such dependencies require evolution mechanisms that must be propagated both in a specific level and across multiple levels. This propagation mechanism is achieved by means of a multi-agent system. Each agent is assigned to a specific schema, monitoring its evolution; an agent can interact with other agents assigned to depending schemata, send them messages and apply evolution to the instances of its schema.

A *coordinator agent* is assigned to each instance of the platform, in order to monitor the updates of the Platform Template Model and to activate the agents connected to each schema when required. The coordinator agent is also devoted to the creation of a new agent every time a new schema is defined (even if it does not act directly on data because such task is delegated to agents located at level M1).

A *schema agent* is devoted to the evolution of contents related to a specific template at level M1. They can perform a set of actions on the existing data, accordingly to the updates affecting related schemata. Agents perform several evolutionary operations on data, in order to preserve data validity and, at the same time, to prevent archivists and content editors to spend a lot of time re-entering the whole set of existing contents. In [14, 15] a complete taxonomy of updates, which can affect a generic XML schema, is described; actually only a subset of the listed operations has been implemented in E-Dvara, covering the set of updates which can be performed by archivists. For example, we provide the utilities to rename or add elements and attributes of the Template Model (level M1). In order to cope with the complexity of the evolution tasks and the amount of data yet available in E-Dvara, our attention is focused on simple updates which commonly occur during the life-cycle of a collection. A typical evolution task is represented by the extraction of a vocabulary (a closed list of predefined strings) from the set of values assigned to a free-text `String` element. In our experience such an update is rather frequent, specifically when we are not able to know *a priori all* the values assignable to a specific element. In this case, when an archivist decides to change the type of the element `Name` from `String` to `Vocabulary`, the agent assigned to that schema should access each instance of the template and perform a `change_item_type`, verifying if the old values assigned to `Name` are validated with respect to the values admitted by the new element type. When this task is completed, the agent should notify the schema updates to the related agents (according to the dependency chain between schemata), in order to grant the consistency of any inter-dependent data.

6 Toward semantic digital libraries

In Section 5 our attention was focused on handling evolution issues from the content providers point of view, more focused on data schemata and publication services. Here we take into account aspects more related to the semantic nature of the information stored in a digital archive. Our perspective, now, is that of a typical final user which uses

the digital library services to fulfill a specific information need (e.g. submitting textual queries to the digital library search engines). In this context, we recognize that such need has a dynamic nature. The interactions between final users and the library may evolve, reflecting changes in information needs. (e.g. which is the semantics of a user query? Which content archived in the library best fits the query?) Moreover, according to the growth of Web 2.0 philosophy, new ways to access such information should be provided to users: they could add their own contributions to documents (e.g. tags, comments, etc.), share them with other users improving in such a way the effectiveness of information access.

In conclusion, a digital platform should provide to its users an environment capable of dealing with information retrieval tasks where it is not important the presence of the “exact word” (string matching approach), but of the *intended meaning* underlying the information need. Our proposal involves the creation of a system that will be able to provide accurate search results exploiting several tools coming from automatic categorization algorithms, information filtering and retrieval techniques, personalization, adaptation, and Web 2.0/Semantic Web features. To achieve this goal, we have designed the PIRATES framework. Its integration on the top of the E-Dvara platform is aimed at providing a “semantic layer” into the digital library.

6.1 The Pirates framework

PIRATES (Personalized Intelligent Recommender and Annotator TEStbed) is a general framework for text-based content retrieval and categorization and exploits social tagging, user modeling, information filtering, and information extraction techniques. The main feature of PIRATES concerns a novel approach that automates in a personalized way some typical manual tasks (e.g. content annotation and tagging). The framework operates on the contents archived in the Information Base (IB) repository (e.g. the digital library archives) and suggests for these some personalized tags recommendations and other forms of textual annotations (e.g. key-phrases) in order to classify them. The original contents are then annotated with these tags, forming enriched archives that we store in a Knowledge Base (KB) repository. These two different types of archives denote our particular approach to access and manage information provided by digital libraries. In our view, annotating a specific content with semantic information such as that potentially conveyed by tags or key-phrases, we shift from the perspective of *data* to the perspective of *knowledge*.

Personalization is achieved exploiting user profiles (which represent the user interests), personal ontologies, personal tags, etc. Furthermore, PIRATES provides several mechanisms of user feedback that help to adapt the content retrieved and showed by the digital library services to the needs of the final user, even where these needs change.

6.2 The PIRATES modules

The PIRATES architecture is illustrated in Figure 6. On the left-hand side, all the possible input sources are shown: single textual documents, specific IB repositories which can be contained within an e-learning or knowledge management environment, and the Web, with specific (but not exclusive) focus on Web 2.0 portals, social networks, etc.

The right-hand side shows the annotations automatically suggested to the user and the resulting KB repository. The main modules of our PIRATES framework can be classified in two categories: modules that discover new content (*Web agents and filters*), and modules that extract information from these new contents, in the form of textual annotations (*content annotators*). Typically, the modules devoted to retrieve new contents are started first, initializing repositories and the “semantic environment”.

Web agents and filters

- IFT Web Agents, which continuously monitor the Web (and the blogosphere) looking for new information, cooperates with IFT to filter contents according to the user model, loads and updates the IB repository as soon as new relevant information is available.
- IFT (*Information Filtering Tool*), which evaluates the relevance (in the sense of topicality) of a document according to a specific model of user interests represented with semantic (co-occurrence) networks [1]. IFT and its Web agents form together the Cognitive Filtering module discussed in [9].

As soon as new content is available, one or more annotators can be executed. The number of annotators and the exact order of execution is user defined.

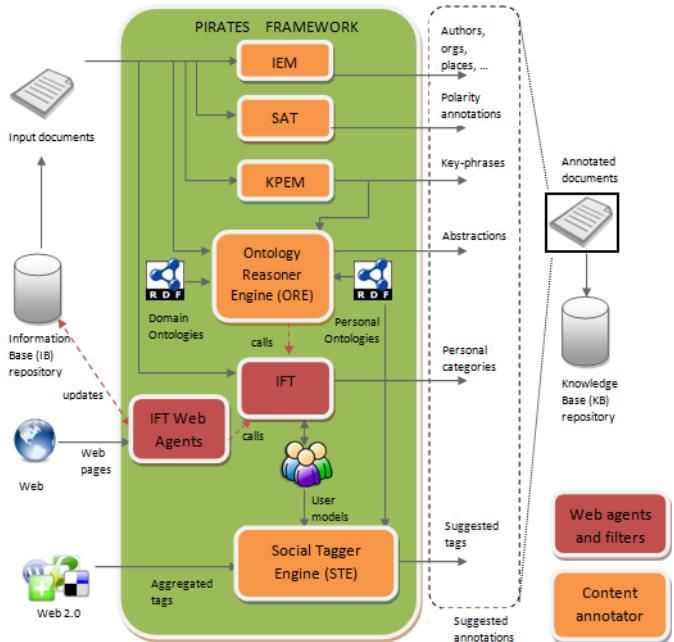


Figure 6: Overall architecture of PIRATES.

Content annotators

- IEM (*Information Extraction Module*), which is currently based on the GATE platform [12] to extract named entities, adjectives, proper names, etc. from input documents, contained in the IB.
- SAT (*Sentiment Analysis Tool*), which is a specific plug-in for personalized sentiment analysis (typically to be activated for marketing intelligence applications), that is capable of mining consumer opinions in the blogosphere and classify them according to their polarity (positive, negative, or neutral)[10].
- KPEM (*Key-Phrases Extraction Module*), which implements a variation of the KEA algorithm [13] for key-phrase extraction. KPEM identifies n-gram key-phrases (typically with n between 1 and 4) that summarize each input document. This information is provided to the user, but is also given in input to possibly subsequent modules.
- ORE (*Ontology Reasoner Engine*), which suggests new *abstract concepts* by navigating through ontologies, classification scheme, thesauri, lexicon (such as WordNet), etc. An abstract concept is identified by looking for a match between the annotations found by the other modules (IEM, KPEM, IFT, and STE) and the concepts stored in ontologies. When a match is found, ORE navigates through the ontology, looking for the common parent node which represents the more abstract term suggested as annotation. ORE also assists users in creating personal ontologies with techniques similar to those described in [23].
- STE (*Social Tagger Engine*), which suggests new annotations for a document relying on *aggregated tags*, i.e. the user's personal tags (tags previously exploited) and the more popular tags used by the community of people that classify the same document in social bookmarking sites such as Del.icio.us⁶, Faviki⁷ or Bibsonomy⁸. This social information is integrated with content-based analysis techniques as discussed in [25].

Our main goal in building the PIRATES framework is to empower the services provided by digital libraries, allowing users to exploit social bookmarking tools in order to easily add new contents in the library archives and categorizing such content by means of keywords (tags) in a personalized and adaptive way. This work is a first step toward the generation and sharing of personal information spaces described in [9]. We have designed PIRATES keeping in mind several applications where it can provide innovative adaptive tools enhancing user capabilities:

- in e-learning for supporting the tutor and teacher activities for monitoring (in a personalized fashion) student performance, behavior, and participation;
- in knowledge management contexts (including for example scholarly publication repositories [19]) for supporting document filtering and classification and for alerting users in a personalized way about new posts relevant to individual interests;
- in online marketing for monitoring and analyzing the blogosphere where word-of-mouth and viral marketing are nowadays more and more expanding.

⁶ <http://delicious.com>

⁷ <http://www.faviki.com/pages/welcome/>

⁸ <http://bibsonomy.org>

7 Conclusions

In this paper we have introduced three specific evolution dimensions which characterize our conceptual model for handling evolution in content production within a digital library initiative. Furthermore, we have proposed the introduction of a semantic layer on the top of the E-Dvara digital library aimed at better addressing the changes in the final users information needs and improving the effectiveness of the information access. To support this new semantic layer, we have designed a framework based on adaptive and personalized services that can empower the three dimensions of our conceptual model (especially in the social domain), distinguishing the digital library from a old-fashioned DBMS/structured archive system. Give access to the semantics of contents helps to realize the vision of semantic digital library which is possibly one of the most innovative evolutions of current digital libraries. These proposals come from the lessons learned during the experimentation with the first prototype of the E-Dvara platform. We are now working to complete a second version of E-Dvara which will embody the improvements discussed in this paper. Our future plans include a validation of the new prototype in different areas, concerning the exploitation of both information and services by means of mobile applications, virtual museums, and Web 2.0 environments.

References

1. F. A. Asnicar, M. Di Fant, and C. Tasso. User model-based information filtering. In A. Evans, S. Kent, and B. Selic, editors, *AI*IA 97: Advances in Artificial Intelligence - Proceeding of the 5th Congress of the Italian Association for Artificial Intelligence*, volume 1321 of *Lecture Notes in Artificial Intelligence*, pages 242–253, Berlin, 1997. Springer-Verlag.
2. D. Bainbridge, G. Buchanan, J. Mcpherson, S. Jones, A. Mahoui, and I.H. Witten. Greenstone: A platform for distributed digital library applications. In *ECDL '01: European Digital Library Conference*, pages 137–148, Berlin, 2001. Springer-Verlag.
3. B.R. Barkstrom, M. Finch, M. Ferebee, and C. Mackey. Adapting digital libraries to continual evolution. In *JCDL '02: Proceedings of the 2nd ACM/IEEE-CS Joint Conference on Digital Libraries*, pages 242–243. ACM, 2002.
4. A. Baruzzo and P. Casoto. A flexible service-oriented digital platform for e-content management in cultural heritage. In *IABC '08: Intelligenza Artificiale nei Beni Culturali Workshop*, pages 38–45, 2008.
5. A. Baruzzo, P. Casoto, P. Challapalli, and A. Dattolo. An intelligent service oriented approach for improving information access in cultural heritage. In *IACH '08: Information Access in Cultural Heritage (IACH) Workshop, European Conference on Digital Libraries*, Berlin, 2008. Springer-Verlag.
6. A. Baruzzo, P. Casoto, A. Dattolo, and C. Tasso. A conceptual model for digital libraries evolution. In *WEBIST '09: Proceedings of 5th Informational Conference on Web Information Systems and Technologies*, pages 299–304, Berlin, 2009. Springer-Verlag.
7. J. Bekaert, X. Liu, and H. Van de Sompel. aDORe: A modular and standards-based digital object repository at the Los Alamos National Laboratory. In *JCDL '05: Joint Conference on Digital Library*, pages 367–367. ACM, 2005.
8. D. Candela L., Castelli and P. Pagano. A reference architecture for digital library systems: Principles and applications. In *Digital Libraries: Research and Development, 1st International DELOS Conference*, pages 22–35, 2007.

9. P. Casoto, A. Dattolo, F. Ferrara, N. Pudota, P. Omero, and C. Tasso. Generating and sharing personal information spaces. In *Proc. of the Workshop on Adaptation for the Social Web, 5th ACM Int. Conf. on Adaptive Hypermedia and Adaptive Web-Based Systems*, pages 14–23, 2008.
10. P. Casoto, A. Dattolo, and C. Tasso. Sentiment classification for the Italian language: A case study on movie reviews. *Journal of Internet Technology*, 9(4):365–373, 2008.
11. S.R.C.P. Challapalli, M. Cignini, P. Coppola, and P. Omero. E-dvara: an xml based e-content platform. In *AICA: Associazione Italiana per l'Informatica e il Calcolo Distribuito*, 2006.
12. H. Cunningham. Gate, a general architecture for language engineering. *Computers and the Humanities*, 36:223–254, 2002.
13. E. Frank, G.W. Paynter, I.H. Witten, C. Gutwin, and C.G. Nevill-Manning. Domain-specific keyphrase extraction. In *IJCAI '99: Proc. of the Sixteenth International Joint Conference on Artificial Intelligence*, pages 668–673. Morgan Kaufmann, 1999.
14. G. Guerrini, M. Mesiti, and R. Rossi. Impact of xml schema evolution on valid documents. In *WIDM '05: Proceedings of the 7th Annual ACM International Workshop on Web Information and Data Management*, pages 39–44. ACM, 2005.
15. G. Guerrini, M. Mesiti, and M. A. Sorrenti. Xml schema evolution: Incremental validation and efficient document adaptation. In *Database and XML Technologies, 5th International XML Database Symposium*, pages 92–106, 2007.
16. S.R. Kruk and B. McDaniel. *Semantic Digital Libraries*. Springer Verlag, 2008.
17. C. Lagoze, S. Payette, E. Shin, and C. Wilper. Fedora: An architecture for complex objects and their relationships, 2005.
18. F. Lutzenkirchen. MyCoRe - ein open-source-system zum aufbau digitaler bibliotheken. *Datenbank-Spektrum*, 4:23–27, 2002.
19. P. Omero, N. Polesello, and C. Tasso. Personalized intelligent information services within an online digital library for medicine: the BIBLIOMED system. In *IRCDL '07: Proc. of the Third Italian Research Conference on Digital Library Systems*, pages 46–51, 2007.
20. S. Ross. Approaching digital preservation holistically. In A. Tough and M. Moss, editors, *Information Management and Preservation*, pages 115–153, Oxford, 2006. Chandos Press.
21. S. Ross. Digital preservation, archival science and methodological foundations for digital libraries. In *ECDL '07: European Digital Library Conference*, Berlin, 2007. Springer-Verlag.
22. I. Rowlands and D. Bawden. Digital libraries: A conceptual framework. *Libri*, 49:192–202, 1999.
23. M. Speretta and S. Gauch. Using text mining to enrich the vocabulary of domain ontologies. *Web Intelligence and Intelligent Agent Technology, IEEE/WIC/ACM International Conference on*, 1:549–552, 2008.
24. R. Tansley, M. Bass, D. Stuve, M. Branschofsky, D. Chudnov, G. McClellan, and M. Smith. The DSpace institutional digital repository system: Current functionality. In *JCDL '03: Joint Conference on Digital Libraries*, pages 87–97. IEEE, 2003.
25. C. Tasso, P.G. Rossi, C. Virgili, and A. Morandini. Exploiting personalization techniques in e-learning tools. In *SW-EL '04: Proc. of the Workshop on Applications of Semantic Web Technologies for Adaptive Educational Hypermedia*, 2004.
26. I.H. Witten, R.J. McNab, S.J. Boddie, and D. Bainbridge. Greenstone: A comprehensive open-source digital library software system. In *ICDL '00: International Conference on Digital Libraries*. ACM, 2000.
27. J. Yates. *Control Through Communication*. The Johns Hopkins University Press, 1989.

WibNED: Wikipedia Based Named Entity Disambiguation

Anna Lisa Gentile, Pierpaolo Basile, and Giovanni Semeraro

Dipartimento di Informatica, Università di Bari

Via E. Orabona, 4 - 70125 Bari - Italia

{al.gentile, basilepp, semeraro}@di.uniba.it

Abstract. Natural Language is a mean to express and discuss concepts, which are taken to be abstractions from perceptions of the experienced real world: what texts describe consist of objects and events. Objects of the real world are identified by proper names, which are words, thus raising the problem of proper linkage between the textual reference and the real object. This work addresses the problem of automatically association of meanings to words within an unstructured text and focuses the attention on words representing Named Entities. The proposed solution consists of a Knowledge based algorithm for Named Entity Disambiguation: we used an *ad hoc* builded corpus, extracted form Wikipedia's articles to prove the soundness of the algorithm.

1 Introduction

A proper name is a word or a list of words that refers to a real world object. According to Frege, a proper name has a reference (Bedeutung) and a sense (Sinn) [6]. The reference is the object that the expression refers to (different linguistic expressions can have the same reference). The sense is the *cognitive significance*, the way by which the referent is presented. Linguistic Expressions with the same reference may have different senses, so it is necessary to disambiguate between them. In Natural Language Processing field Named Entity Disambiguation is the task that aims to solve this issue. NLP operations include text normalization, tokenization, stop words elimination, stemming, Part Of Speech tagging, lemmatization. Further steps, as Word Sense Disambiguation (WSD) or Named Entity Recognition (NER), are aimed at enriching texts with semantic information. Named Entity Disambiguation (NED) is the procedure that solves the correspondence between real-world entities and mentions within text. The proposed approach automatically associates each entity in a text with a unique identifier, a URI from Wikipedia¹, which is used as an "entity-provider". The contribution of this work is twofold: firstly a novel knowledge based approach for NED is proposed; secondly the work shows a method to build a testbed dataset from Wikipedia. The suggested solution is completely knowledge-based, with the advantage that no training data is needed: indeed, manually annotated data for this task is not easily available, so acquiring such data can be expensive.

¹ <http://www.wikipedia.org>

The work is structured as follows: Section 2 proposes an overview of Named Entity Disambiguation task, together with a description of available solutions to exploit Wikipedia for the issue of Named Entities. Section 3 presents the proposed Wikipedia based Named Entity Disambiguation algorithm, named WibNED. Section 4 presents the dataset used for experiments, which are then described in section 5. Conclusions close the paper.

2 Related Work

Named Entity Disambiguation is the problem of mapping mentions of entities in a text with the object they are referencing. It is a step further from Named Entity Recognition (NER), which involves the identification and classification of so-called named entities: expressions that refer to people, places, organizations, products, companies, and even dates, times, or monetary amounts, as stated in the Message Understanding Conferences (MUC) [8]. The NED process aims to create a mapping between the *surface form* of an entity and its unique dictionary meaning. It can be assumed to have a dictionary of all possible entity entries. In this work we use Wikipedia as such a dictionary. Many studies that exploit Wikipedia as a knowledge source have recently emerged [12, 13, 16]. In particular, Wikipedia turned to be very useful for the problem of Named Entities due to its greater coverage than other popular resources, such as WordNet [5] that, resembling more to a dictionary, has little coverage on named entities [13]. Lots of previous works exploited Wikipedia for the task of NER, e.g. to extract gazetteers [14] or as an external knowledge of features to use in a Conditional Random Field NER-tagger [9], to improve entity ranking in the field of Information Retrieval [15]. On the other hand, little has been carried out on the field of NED. The most related works on NED based on Wikipedia are those by Bunescu and Pasca [3] and Cucerzan [4]. Bunescu and Pasca consider the problem of NED as a ranking problem. The authors define a scoring function that takes into account the standard cosine similarity between words in the context of the query and words in the page content of Wikipedia entries, together with correlations between pages learned from the structure of the knowledge source (mostly using Wikipedia Categories assigned to the pages). Cucerzan proposes a very similar approach: the vectorial representation of the document is compared with the vectorial representation of the Wikipedia entities. In more details the proposed system represents each entity of Wikipedia as an *extended vector* with two principal components, corresponding to context and category information; then it builds the same kind of vector for each document. The disambiguation process consists of maximizing the *Context Agreement*, that is the overlap between the document vector for the entity to disambiguate and each possible entity vector. Both described works are based on the Vector Space Model, which means that a pre-computation on the Wikipedia knowledge resource is needed to build the vector representation. The proposed solution, differently to previous methods, exploits words in the context of an entity in a simple way, calculating the gloss overlapping between context and dictionary entries.

For the task of NED little resources and benchmark data are publicly available. On the other hand lots of data is available for the task of Named Entity Recognition: multilingual benchmarking and evaluations have been performed within several events, such

as the Message Understanding Conferences (MUC) series organized by DARPA, the International Conference on Language Resources and Evaluation (LREC), the Computational Natural Language Learning (CoNLL) workshops, the Automatic Content Extraction (ACE) series organized by NIST, the Multilingual Entity Task Conference (MET), the Information Retrieval and Extraction Exercise (IREX). The problem with data shared within these events is that entities are not labelled with a URI, but only classified within a set of predefined entity classes, which means that is not directly reusable for the task of NED. Some useful data for NED has been provided by Cucerzan, but the dataset only contains information about Named Entities and not all the text, so it is not useful for the purpose of this work. Moreover we wanted to evaluate our algorithm for the italian language and the available dataset is in english. For these reasons a specific dataset has been built to validate the proposal, automatically extracted from italian Wikipedia articles containing ambiguous entities, maintaining both entities and other words within the text.

3 WibNED: Wikipedia based Named Entity Disambiguation

The goal of a WSD algorithm consists in assigning a word w_i occurring in a document d with its appropriate meaning or sense s , by exploiting the *context* C in which w_i is found. The context C for w_i is defined as a set of words that precede and follow w_i . The sense s is selected from a predefined set of possibilities, usually known as *sense inventory*.

The Lesk algorithm is a classical algorithm for Word Sense Disambiguation for all words in unrestricted text. It was introduced by Mike Lesk in 1986 [11]. The basic assumption is that words in a given neighbourhood will probably share a common topic. Apart from knowledge about the context (the immediate surrounding words), the algorithm requires a machine readable dictionary, with an entry for each possible sense for a word. The original algorithm takes into account words pairwise and computes overlap among sense definitions: the sense pair with the highest overlap score is chosen.

The proposed algorithm, named WibNED, is an adaptation of Lesk dictionary-based WSD algorithm [1]. In the WibNED algorithm the words to disambiguate are only those representing an Entity.

WibNED takes as input a document $d = \{w_1, \dots, w_j, e_{j+1}, w_{j+2}, \dots, w_h, e_{h+1}, w_{h+2}, \dots\}$ and returns a list of Wikipedia URIs $X = \{s_1, s_2, \dots, s_k\}$ in which each element s_i is obtained by disambiguating the *target entity* e_i on the ground of the information obtained from Wikipedia for each candidate URI (Wikipedia page content of the URI) and words in the *context* C of e_i . We define the *context* C of the target entity e_i to be a window of n words to the left and another n words to the right, for a total of $2n$ surrounding words. In the current version of the algorithm if other entities occur in the context of the target entity, they are considered as words and not as entities.

Algorithm 1 describes the complete procedure for the disambiguation of entities. The input for the algorithm is an ordered list W

$$W = (w_1, \dots, w_j, e_{j+1}, w_{j+2}, \dots, w_h, e_{h+1}, w_{h+2}, \dots)$$

of words in a document, processed with a NLP tool: w_i are common words while e_i are tagged as Named Entities.

The NLP tool used to process the document collection is META [2], that performs the following operations:

- tokenization;
- part of speech (POS) tagging;
- stop words elimination;
- word sense disambiguation by using WordNet;
- Entity Recognition performed with Yamcha [10], a NER annotator which uses a Support Vector Machine techniques.

Each document of the collection is then transformed in an ordered list of common words w_i and Named Entities e_i :

$$W = (w_1, \dots, w_j, e_{j+1}, w_{j+2}, \dots, w_h, e_{h+1}, w_{h+2}, \dots)$$

The list W is the input for the main procedure, named WibNED, that finds the proper Wikipedia URI for each polysemous entity e_i in W . WibNED uses several sub-procedures. The subprocedure QueryWiki finds all possible Wikipedia pages answering the query e_i . Each element of the returning set contains the page URI and a short textual description of the page. The subprocedure pickPage computes the overlap between the context of the target entity and the description obtained from the Wikipedia page for each candidate sense. It returns the candidate entity which maximizes the overlap.

4 Dataset

The dataset used for Experiments consists of 752 documents extracted from italian Wikipedia. The evaluation of a NED algorithm which gives as output Wikipedia URIs needs a dataset containing ambiguous entities, already tagged with the correct URI belonging to Wikipedia.

We implemented a procedure to build an automatic annotated corpus, starting from a list of ambiguous entities (i.e. entities whose surface form has an associated *Disambiguation Page*² in Wikipedia). We used a list of 100 ambiguous surface form, taken from italian Wikipedia, $A = (a_1, \dots, a_{100})$. For each a_i we accessed the related *Disambiguation Page* on Wikipedia and we picked the most significative senses for a_i , $s_{a_i} = (s_1, \dots, s_j)$, with $j \leq 4$, considering only senses referring to Named Entities and using heuristics to reject poor senses. For example, considering the *disambiguation page* for the italian word "mosca", the sense referring to *Mosca (Moskva), the capital and the largest city of Russia*, has been stored whereas the sense referring to *Muscomorpha, a group of flies*, has been ignored because it is a common noun word. Starting from $S = (s_{a_1}, \dots, s_{a_t}, \dots, s_{a_{100}})$ for each s_j in s_{a_t} we choosed up to 5 generic Wikipedia articles that contain at least a link to the sense s_j . Each article has been processed using META [2] and has been stored as a single file, using a IOB like format, similar to CoNLL 2003 Named Entity Recognition corpus³: each row contains a token, its Part Of Speech tag, its lemma and a final tag which is valued as O if the token has not been

² http://en.wikipedia.org/wiki/Category_Disambiguation_pages

³ <http://www.cnts.ua.ac.be/conll2003/ner/>

Algorithm 1 *WibNED*, the algorithm for the disambiguation of entities

```

procedure WibNED(W)
    W = ( $w_1, \dots, w_j, e_{j+1}, w_{j+2}, \dots, w_h, e_{h+1}, w_{h+2}, \dots$ )
    W is the list of words in a document, processed with a NLP tool.
     $w_i$  are common words while  $e_i$  are tagged as Named Entities.

    S  $\leftarrow$  W
    for all  $e_i \in S$  do
         $Context_{e_i} \leftarrow \{w_{i-n}, \dots, w_{i+n}\}$ 
         $Candidate_{e_i} \leftarrow QueryWiki(e_i)$ 
         $s_{e_i} \leftarrow pickPage(Context_{e_i}, Candidate_{e_i})$ 
        S.replace( $e_i, s_{e_i}$ )
    end for
    return S
end procedure

procedure QueryWiki( $e_i$ )
    Candidate  $\leftarrow \{\}$ 
    WikiResults  $\leftarrow$  all pages from Wikipedia answering the query  $e_i$ .
    Each element of the returning set contains the page URI
    and a short textual description of the page.

    Candidate.add( $c_i$ )
    for all  $wikis_i \in WikiResults$  do
         $c_i.uri \leftarrow wikis_i$ 
         $c_i.description \leftarrow describe(wikis_i)$ 
        describe(wikis_i) builds a short textual description of the page.
    end for
    return Candidate
end procedure

procedure pickPage(Context, Candidate)
    maxOverlap  $\leftarrow 0$ 
    bestPage  $\leftarrow \text{null}$ 
    for all  $c_i \in Candidate$  do
         $currentOverlap \leftarrow computeOverlap(c_i.description, Context)$ 
        if  $currentOverlap > maxOverlap$  then
            computeOverlap( $c_i.description, Context$ )
            calculate the number of overlapping words
            between Context and  $c_i.description$ 
             $bestPage \leftarrow c_i.uri$ 
            maxOverlap  $\leftarrow currentOverlap$ 
        end if
    end for
    return bestPage
end procedure

```

recognized as an entity, B-<Wikipedia URI for the entity> if the token is the beginning of an entity and I-<Wikipedia URI for the entity> if the token continues an entity.

Each text contains on the average 89 entities. In table 1 we report a piece of a document to show an example of text. It is a document included in the corpus, specifically it is an article taken from italian Wikipedia about *The Beastie Boys*. In this piece of text there is only one entity, represented by the two words John Berry. Together with entity annotations, the corpus also contains the Part Of Speech and the stem for each word, thus allowing to refine and improve computation over the corpus.

Table 1. Corpus example

Il	RS	Il	O
nome	SS	nome	O
"	XPO	"	O
Beastie	SP	beastie	O
"	XPO	"	O
,	XPW	,	O
inventato	VSP	inventare	O
dall	SN	dall	O
,	XP	,	O
ex	SN	ex	O
componente	SS	componente	O
John	SPN	John	B- http://it.wikipedia.org/wiki/John_Berry
Berry	SPN	Berry	I- http://it.wikipedia.org/wiki/John_Berry
,	XPW	,	O
é	VI	essere	O
l	SN	l	O
,	XP	,	O
acronimo	AS	acronimo	O
della	ES	della	O
frase	SS	frase	O
"	XPO	"	O
Boys	YF	Boys	O
Entering	YF	Entering	O
Anarchistic	YF	Anarchistic	O
States	YF	States	O
Towards	YF	Towards	O
Inner	SPN	Inner	O
Excellence	SPN	Excellence	O
"	XPO	"	O

5 Experiments

We performed the experiment following the methods generally used to evaluate Word Sense Disambiguation (WSD) algorithms. The entity WSD is not an end in itself but

rather an intermediate task which contributes to an overall task such as information retrieval, ontology building, etc. This opens the possibility of two types of evaluation for WSD work (using terminology borrowed from biology): *in vitro* evaluation, where WSD systems are tested independently of any application, using specially constructed benchmarks; and *evaluation in vivo*, where, rather than being evaluated in isolation, results are evaluated in terms of their contribution to the overall performance of a system designed for a particular application (e.g., information retrieval). In this instance we adopt *in vitro* evaluation in order to evaluate the accuracy and the potentialities of the algorithm in an independent way. *In vitro* evaluation, despite its artificiality, enables close examination of the problems plaguing a given task. In its most basic form this type of evaluation involves comparison of the output of a system for a given input, using measures such as precision and recall. Alternatively, *in vitro* evaluation can focus on study of the behavior and performance of systems on a series of test suites representing the range of linguistic problems likely to arise in attempting WSD. Considerably deeper understanding of the factors involved in the disambiguation task is required before appropriate test suites for typological evaluation of WSD results can be devised. The *in vitro* evaluation demands the creation of a manually sense-tagged reference corpus containing an agreed-upon set of sense distinctions. The idea is to build a corpus, find the entity into the corpus and annotate the entity with relative sense. In our experiments we use Wikipedia as Sense Inventory for the entities because it's the same used by the WibNED algorithm.

WibNED is implemented in JAVA, by using Lexical Collector web service [7] as accessing point to the Wikipedia Sense Inventory. We run the WibNED algorithm on the dataset described in Section 4 in order to evaluate its effectiveness. We used the accuracy metric that describes the ratio between entities correctly labelled by WibNED and total number of entities within the document.

Experimental results are showed in figure 1: on the x-axis are reported single documents while on the y-axis is reported the accuracy of WiBNED for each document. Documents on x-axis are ordered according to ascending accuracy.

Total Number of Entities	67029
Total Number of correctly labelled Entities	19131
Average number of Entities per Document	89
Average number of Correctly labelled Entities per Document	25
Total Accuracy	0,285
Average Accuracy on single Document	0,282
Minimal Accuracy	0,000
Maximal Accuracy	0,833

Table 2. WibNED Results

Some statistics are reported in table 2. The table provides the total number of entities within the corpus, the total number of correctly labelled entities by WiBNED

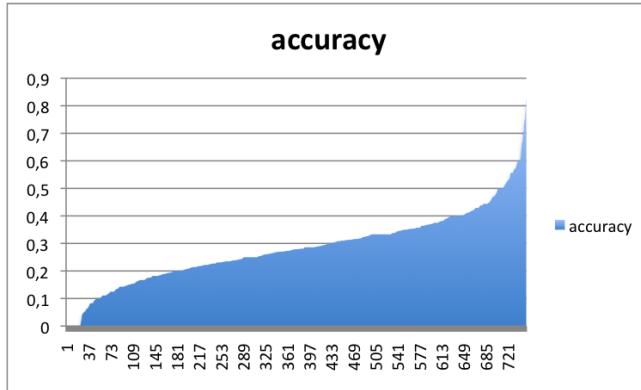


Fig. 1. Accuracy of WiBNED on 752 documents

algorithm, the average number of entities on single documents, the average number of correctly labelled entities per document. Results about accuracy are showed: total accuracy which is calculated considering the total number of entities and total number of correctly labelled entities, regardless of distribution between documents, average accuracy considering the mean of accuracy of all documents. The minimal and maximal values reported for accuracy are listed in the table.

The results are quite encouraging if we consider that WibNED is a very simple algorithm based only on the string-matching between the words in Wikipedia definition and the words within the context of the target entity. The algorithm achieves on average 28,2% of accuracy. Taking into account more informative features borrowed from Wikipedia, such as category labels associated to each article or internal links, could improve results of the algorithm.

6 Conclusions

In this paper we presented the WibNED algorithm for Named Entity Disambiguation and we evaluated it on Italian language using an *ad hoc* builded corpus, automatically annotated with Wikipedia URIs.

The task of disambiguating Named Entities within a text and the problem of identity and reference are important issues for many research field, most of all for NLP: the WibNED algorithm associates unique references to words and uses popular URLs (Wikipedia's) as "canonical" URIs.

An ongoing work is focused on improving accuracy of WiBNED algorithm, relying on more Wikipedia features, such as links and categories instead of using only words within the disambiguation process.

References

1. S. Banerjee and T. Pedersen. An adapted lesk algorithm for word sense disambiguation using wordnet. In *CICLING'02: Proceedings of the Third International Conference on Computational Linguistics and Intelligent Text Processing*, pages 136–145. Springer-Verlag, 2002.
2. P. Basile, M. de Gemmis, A.L. Gentile, L. Iaquinta, P. Lops, and G. Semeraro. Meta multi-language text analyzer. In *Proceedings of the Language and Speech Technology Conference LangTech 2008*, pages 137–140, 2008.
3. R. C. Bunescu and M. Pasca. Using encyclopedic knowledge for named entity disambiguation. In *Proceedings of EACL 2006, 11st Conference of the European Chapter of the Association for Computational Linguistics*, pages 9–16. ACL, 2006.
4. S. Cucerzan. Large-scale named entity disambiguation based on Wikipedia data. In *Proceedings of EMNLP 2007: Empirical Methods in Natural Language Processing*, pages 708–716, 2007.
5. C. Fellbaum. *WordNet: An Electronic Lexical Database*. MIT Press, 1998.
6. G. Frege. Über sinn und bedeutung. In Mark Textor, editor, *Funktion - Begriff - Bedeutung*, volume 4 of *Sammlung Philosophie*. Vandenhoeck & Ruprecht, Göttingen, 1892.
7. A. L. Gentile, P. Basile, L. Iaquinta, and G. Semeraro. Lexical and semantic resources for nlp: From words to meanings. In *Knowledge-Based Intelligent Information and Engineering Systems*, volume 5179/2008 of *Lecture Notes in Computer Science*, pages 277–284. Springer Berlin / Heidelberg, 2008.
8. R. Grishman and B. Sundheim. Message understanding conference-6: A brief history. In *COLING*, pages 466–471, 1996.
9. J. Kazama and K. Torisawa. Exploiting wikipedia as external knowledge for named entity recognition. In *Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 698–707, 2007.
10. T. Kudo and Y. Matsumoto. Fast methods for kernel-based text analysis. In Erhard Hinrichs and Dan Roth, editors, *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 24–31, 2003.
11. M. E. Lesk. Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from an ice cream cone. In *Proceedings of the Fifth International Conference on Systems Documentation*, pages 24–26, Toronto, CA, 1986. ACM.
12. S. P. Ponzetto and M. Strube. Exploiting semantic role labeling, wordnet and wikipedia for coreference resolution. In Robert C. Moore, Jeff A. Bilmes, Jennifer Chu-Carroll, and Mark Sanderson, editors, *HLT-NAACL*. ACL, 2006.
13. M. Strube and S.P. Ponzetto. Wikirelate! computing semantic relatedness using wikipedia. In *AAAI*, pages 1419–1424. AAAI Press, 2006.
14. A. Toral and R. Munoz. A proposal to automatically build and maintain gazetteers for named entity recognition by using wikipedia. In *Workshop on New Text, EACL 2006, 11th Conference of the European Chapter of the Association for Computational Linguistics*. ACL, 2006.
15. A. Vercoustre, J. A. Thom, and J. Pehcevski. Entity ranking in wikipedia. In R. L. Wainwright and H. Haddad, editors, *SAC*, pages 1101–1106. ACM, 2008.
16. T. Zesch, I. Gurevych, and M. Mühlhäuser. Analyzing and accessing wikipedia as a lexical semantic resource. In *Biannual Conference of the Society for Computational Linguistics and Language Technology*, 2007.

A Continuous Language Modelling Approach for Assessing Real-valued Attributes of Documents

Richard Bache¹ and Fabio Crestani²

¹Department of Computer and Information Science,
University of Strathclyde, Scotland
richard.bache@cis.strath.ac.uk

²Faculty of Informatics
University of Lugano, Lugano, Switzerland
fabio.crestani@unisi.ch

Abstract: Probabilistic Language Modelling has been widely used for document classification and document ranking. This paper focuses on the classification task. Since language models are generative models, each language model is assumed to emit terms with (fixed) probabilities that can notionally generate the document to be classified. The fixed probabilities are derived from a training set of documents. Such an approach assumes a finite number of competing models representing each category or alternatively two models representing the presence or absence of some Boolean attribute. However, it has been widely recognised that a set of fixed models cannot cope adequately with either fuzzy classification or real-valued attributes. We therefore propose a continuous language model where the probability that a term is emitted is determined by a real-valued parameter. This model uses logistic regression to estimate a function of this parameter for each term in the vocabulary. Using just one continuous model and Bayesian approach is it possible to estimate the parameter value, representing the attribute in question, for a given unclassified document. Although other diverse applications are also mentioned, the research was motivated by the analysis of crime data. Therefore, in the experimentation presented in this paper, documents are descriptions of crimes and the attributes of interest are age of offender and distance travelled from the offender's home, as examples of two continuous attributes. Estimates from the model shows to correlate significantly with actual values establishing a relationship between the behaviour described in the crime report and both age and distance travelled.

1. Introduction

Police routinely collect large quantities of crime data and store this electronically. Such data consists typically of descriptions of how a crime was committed, comprising free text and structured data. Where a crime has been solved, further data is then available about the offender. Such data archives lend themselves to data mining techniques to identify relationships between crimes and the characteristics of offenders such as age, sex, and ethnic group or where they lived.

Our motivation for conducting this investigation is to develop automatable quantitative techniques for relating features of the crimes to the kinds of people who may have committed them; this is often referred to as *offender profiling*. Traditionally, criminal profiling has been conducted subjectively with little empirical basis. More recent work has sought to apply a quantitative approach such as Canter and Fritzon's analysis of arson [5] where one of the characteristics of interest was age. Given that this type of analysis of crime data is still in its infancy, finding even weak statistical relationships would be considered a useful result and could spur further investigation. Furthermore, if we used transparent techniques where inferences can be linked back to the underlying variables then this enables us to explore which features differentiate kinds of offenders. It should be noted that where such techniques were used by law enforcement agencies such as the police, transparency is essential in order to justify the conclusions. The fact that some opaque mathematical model has made some inference would not convince senior officers or courts of law.

The work presented here was carried out in the context of the project, iMOV (Interactive Modus Operandi Visualisation), funded by EPSRC (the UK Engineering and Physical Sciences Research Council). The project is a joint work of the Department of Computer and Information Sciences of the University of Strathclyde, Glasgow, and the Department of Investigative Psychology of the University of Liverpool. This particular study wished to investigate the relationship between the behaviour of the offender when committing the crime and both the age and distance travelled from the home base. If we consider the description of the offender's behaviour (free text and other variables) to comprise a (virtual) document describing the crime, then we can see that relating the behaviour to a characteristic such as gender or racial group of the offender is a document classification problem. However, since age and distance travelled are not naturally categorical data, attempts to 'classify' on these bases would require defining arbitrary categories e.g. *young* and *old* or *near* and *far* with an arbitrary age or distance boundary.

Bayesian approaches have been used previously to analyse crime data [11]. Language Modelling is a Bayesian approach and has been applied widely to the problem of document classification [4, 12, 13]. Specifically in the crime domain, this approach has been used to link the behaviour of known offenders [1] and also for identifying Boolean characteristics of offenders [2]. The fact that probabilities are associated with each term means that it is possible to peek inside the model and relate terms in the document (e.g. words or behavioural features) to characteristics of the offender. Language models in this respect offer advantages over other classification techniques such as vector support machines or neural nets which are essentially opaque. The drawback of the traditional Language Modelling approach is that it assumes that a document is placed in exactly one of a finite set of collectively exhaustive and mutually exclusive categories. Since distance and age are defined on a numeric scale, existing language models cannot deal adequately with a continuous attribute.

We will henceforth refer to traditional language models as *constant* models. Although they represent a stochastic process, the probabilities associated with each term in the vocabulary are assumed to be fixed for each model, having been first calibrated on some training set. Although it would be possible to apply two or more constant models over intervals of a numeric scale, such an approach has three drawbacks:

- The output can only be a range of values corresponding to a discrete category rather than a point estimate or a confidence interval.
- The model assumes catastrophic changes at (often arbitrary) boundaries and so training points at these boundaries will contribute 100% to either one language model or the other.
- Any attempt to make the intervals smaller necessarily requires increasing the number of models and spreading the training data more thinly.

Here we propose an alternative approach. We assume a single language model that has continuous parameters. These parameters alter the probabilities of each term in the vocabulary since each term has associated with it a probability function rather than a fixed probability. Using a Bayesian approach we then seek to estimate the value of these parameters for a document for which the attribute is unknown. We called this class of models: *continuous language models*. We will consider here the specific case where there is a single parameter representing the attribute of interest although such a method could, in principle, be extended to multiple parameters to deal with situations where many attributes are being estimated simultaneously.

The models were presented by Bache and Crestani in [3]. This paper extends [3] by explaining the models in more detail and by presenting an extensive evaluation of their effectiveness. The potential of these models extends well beyond the analysis of crime data. There will be situations where some attribute of a text is not dichotomous (e.g. reading age or degree of sexual or violent content) or where classification may be fuzzy (e.g. news stories that are in varying degrees about two topics). However, these applications are not considered here.

The rest of the paper is organised as follows. Section 2 describes the continuous language model by analogy with classification by discrete language models. In section 3 we describe the crime data used for the analysis. Section 4 presents the empirical results by comparing the continuous models with the alternative of using a dichotomous set-up with an arbitrary cut-off point. In Section 5, we show how the extraction of tokens can show differences in behaviour between older and younger offenders as well as those who travel near or far. Section 6 offers some conclusions.

2. Deriving Continuous Language Model

We can identify three distinct steps when applying traditional Language Modelling to a classification problem and this provides a useful template for explaining continuous language models.

- The training data is used to calibrate one language model per category, that is to assign probabilities for each term in each model. Smoothing is usually applied at this stage to account for the terms that do not appear in a specific document (see later).
- For a document D , we calculate the probability $P(D|M_i)$ that each of the models M_i could have generated that document. This is achieved by multiplying the probabilities for each term in the vocabulary. Here *term* refers to words or phrases from free text or the name of a category for categorical data.

- Bayes' Theorem is then used to invert the probabilities by assuming some prior, so that we can calculate a probability that a particular model generated a given document, that is $P(M_i | D)$.

Continuous language models seek to estimate some attribute relating to the document and thus we assume instead a single language model with a parameter v representing age, distance or some other quantity. Without loss of generality, we will henceforth consider age as the parameter of interest. Using the above template, firstly, we use logistic regression to assign a function to each term in the vocabulary; this function yields a probability for a given age. The second step produces a probability of the document being generated as a function of age by multiplying the probabilities of each term. Finally we use the continuous version of Bayes' Theorem by assuming a prior distribution of ages over solved offences and calculating the posterior distribution of the age of the offender of the crime under analysis. A point estimate can be made by, for example, taking the mean of this distribution.

2.1 Multinomial versus Multiple Bernoulli

It is worth noting here that, whereas it is usual to speak of a language model generating a document, strictly speaking the model actually generates the document's index. Stopwords are removed, words may be stemmed or lemmatised and word order is lost to yield a bag of terms; categorical variables are added as extra words or phrases to the text. It is because we are using a bag of terms that the language models used are unigram and thus take no account of context or word order.

We argue that applications of Language Modelling to document classification and to document retrieval are closely related. We can thus divide the traditional language models used in both into two groups: multinomial and multiple Bernoulli. The former takes into account the number of incidences of a term in the index and the latter considers only whether a term is present or absent. The seminal paper by Ponte and Croft [14] used Language Modelling for document retrieval and proposed the Bernoulli approach although many recent developments tended to favour multinomial models [5]. Losada [9] argues that in sentence retrieval, Bernoulli models can offer an advantage. McCallum and Nigam [10] show that Bernoulli models work better where the vocabulary is smaller. It is worth noting that for the data considered here the vocabulary is small and, in common with sentence retrieval, the documents are short (typically 20 words). It is also true that words (other than stopwords) rarely appear more than once in a document. All of these points would gravitate towards using a Bernoulli approach. However, Bernoulli models also offer a further advantage in that the probability of each term being emitted is independent of all probability of other terms. This is not true of the multinomial models and would make adaptation to a continuous model more problematic. We now explain this point in more detail.

For all unigram language models, the emission of any term is assumed to be independent of any other. However multinomial models assume that the document comprises n terms emitted randomly from the vocabulary set with replacement. Thus a probability $P(t_i)$ is assigned to each term in the vocabulary V . It follows that:

$$\sum_{t_i \in V} P(t_i) = 1 \quad (1)$$

Therefore, for any one term to become more frequent, at least one term other must become less frequent. Multiple Bernoulli models assume independent trials in which each specific term in the vocabulary is either emitted or not. Thus one term can become more frequent while not affecting the frequency of any other terms. In other words, it is possible to estimate the frequency of one term independently of all the others. It is this property of the Bernoulli approach that allows us to construct a continuous model.

2.2 Calibrating the Model

Since each term can be treated separately, we can express the probability of each term in the vocabulary as a function of the parameter v representing the age of the offender:

$$P(t_i) = f_i(v) \quad (2)$$

This function, which yields a value between 0 and 1, must be estimated from some set of training data. For each document in that set either a term will be present or absent. So, since we wish to estimate a probability function from Boolean data, we use logistic regression with an affine function:

$$\log \text{it}(P(t_i)) = \log \left(\frac{P(t_i)}{1 - P(t_i)} \right) = a_i v + b_i \quad (3)$$

It is then possible to estimate the values of a_i and b_i so that for a given v we can calculate the probability of term t_i appearing in the document. However, there are two problems that arise from the use of logistic regression because it necessarily requires a numerical approximation algorithm.

Firstly, there may be terms which are either always present in or always absent from every training document. Any logistic regression algorithm will fail to estimate a_i and b_i in this situation. The maximum likelihood estimate of the probabilities of these terms will be 1 and 0 respectively for any value of v . However, this will lead to a problem known to affect traditional language models, the *Zero-Probability Problem* (ZPP). If a language model gives a probability of zero to a term then any document which contains that term cannot, by definition, be generated by the model. Thus the probability of generation will be zero. Similarly if the probability is 1 then any document that fails to contain it also cannot be generated. The usual solution to this problem is smoothing but here we shall show that smoothing is unnecessary. Clearly, if a term is always present or absent in the training documents then we have to assume it is independent of v . We can then assign some constant probability to that term, such as:

$$P(t_i) = k_i \quad (4)$$

However, it is easy to show that any constant term it would cancel out in the numerator and denominator of Bayes' theorem. Since the value of the constant assigned is irrelevant, we can simply disregard the term altogether.

The second problem occurs for a small minority of sparse terms where the logistic regression algorithm fails to converge. This will typically be no more than two or three terms in a vocabulary of thousands. Inspection indicated that these are terms that occurred in only one document, were unlikely to occur in the test set and had little behavioural importance anyhow. Thus the pragmatic solution was to remove these terms from the analysis too. An investigation into various logistic regression algorithms may reduce this, although there is no guarantee that the issue can be eliminated with certainty. Thus the elimination of a small minority of problematic terms may always be necessary.

For terms, for which a_i and b_i could be estimated, $f_i(v)$ will always produce a value such that $0 < f_i(v) < 1$ for any value of v . Thus ZPP cannot occur with this model. Nevertheless, it could be argued that smoothing could improve the performance of the model. However, when applied empirically, the smoothing did not show any improvement in performance.

2.3 Calculating the Probability of the Attribute

The probability of the language model generating the document (or, strictly speaking, its index) for a given value of v may be written as:

$$P(D|v) = \prod_{t_i \in D} P(t_i) \times \prod_{t_i \notin D} (1 - P(t_i)) \quad (5)$$

Since v is modelled as continuous, the continuous form of Bayes' theorem gives us:

$$pdf(v|D) = \frac{pdf(v) \cdot P(D|v)}{P(D)} \quad (6)$$

Here $P(D)$ can be calculated by assuming that the area under the distribution $pdf(v|D)$ will be 1. The prior distribution $pdf(v)$ can be estimated from the training data. Since here we wish only to calculate the mean of the posterior distribution, the fact that this consists of a number of discrete points and is thus rather lumpy does not matter. However, if we wished to produce a graphical representation of the likely ages (or distances), a smoothed curve could be fitted.

3. Data Analysed

The data was drawn from a digital police archive from a large city for crimes reported over a period of years. Nine datasets of various crime types were available with age

data and hailed from one district of the city over a four-year period. One set of burglary data contained coordinates of both crime scene and the offender's home base and thus Euclidean distance could be calculated. It covered the whole city but this was restricted to prolific burglars over more than 10 years. These constraints were imposed by the availability of the data and were outside the control of the researchers. Only solved crimes were considered for the analysis. For the age data, crimes with more than one offender were removed since there would be more than one age associated with that crime. For the distance data there was always one offender per crime. Table 1 summarises the data sets and their respective sizes.

Table 1: Summary of Datasets.

Set No.	Crime Type	Continuous Variable	Median Value	Num Points	No. Offenders
1	Theft from Vehicles	age	16	248	159
2	Other Theft	age	25	326	284
3	Shoplifting	age	29	2060	1618
4	Assault	age	31	2073	1881
5	Criminal Damage	age	26	849	724
6	Damage to Vehicles	age	27	220	207
7	Burglary	age	27	1126	556
8	Street Robbery	age	18	263	210
9	Sexual offences against women	age	37	123	117
10	Burglary	distance (meters)	2689	1376	83

Words from free text were lemmatised using a lemmatiser based on WordNet [7]. Stopwords from a standard list were removed. Certain addition words which may identify age or sex (e.g. *young* or *female*) were removed. Others, which revealed either age or sex, were mapped to neutral terms (e.g. *child*, *man*, *woman* to *person*). Codes were of two types. Allegation codes indicated the type or subtype of offence and one was always present. Features codes related to some observed behaviour and any number could be present although the typical number was one or two. The codes were mapped to phrases which were hyphenated and marked with a \$-sign to remain separated from free text words.

4. Empirical Validation

The purpose of the experiments presented here was to demonstrate firstly that the continuous language models have predictive power and secondly that they perform comparably with a dichotomous model comprising two constant language models splitting the continuous variable at the median. It can be argued that a model that yields a numeric value or a distribution of numeric values is intrinsically preferable than one that assigns categories so it would be sufficient to show that it performs as well as a dichotomous model. Our third purpose was to compare two varieties of dichotomous models, multinomial and multiple Bernoulli. If the multinomial model outperformed the Bernoulli model then it would call into question the Bernoulli assumption for the

continuous model. Both constant types of constant language model were applied with Jelinek-Mercer smoothing [8] taking $\lambda = 0.5$.

4.1 Experimental Design

A common strategy when testing models such as the ones presented here is to assign the data randomly into a training set and a test set, typically with 50% of the points in each. However such an approach was shown to give very different results for each random allocation. Given the relatively small size of the datasets, it was practical to use the *leave-one-out* (LOO) or jack knife strategy of selecting training and test data. Usually, this technique consists of removing one point from the data set and using all the remaining data to train the model. The model is then tested on the removed data point. This procedure is repeated for each data point until each point has been excluded precisely once. However, this potentially leads to the problem of serial crimes where several crimes are committed by the same offender and thus the training set contains crimes by the same offender as the test set. Although we expect to find commonalities in behaviour between people of the same age, this will never be as strong as the consistency of behaviour of one individual. So if a 25-year-old burglar has a data point in both the training set and the test set then the model may well identify a connection peculiar to that individual rather than to offenders of that age in general.

One possible solution to the serial crime problem was to remove the serial crimes. For many of the data sets this would have been a viable option although it would reduce the size of the datasets. For data set 7 and particularly data set 10 of Table 1 this would create a problem since it lost more than half the data points. Thus the solution was to use a *leave-one-offender-out* strategy. Here the test set comprises all the offences by one offender and the training set comprises all offences committed by other offenders. This procedure is then run once for each offender.

The two types of model yield different results. The dichotomous models predict a category whereas the continuous model predicts a distribution of values of which we may take some measure of central tendency such as the mean. Nevertheless, there should be a correlation between the predicted and actual values although different statistical tests would be applied. For the dichotomous model we use the Chi-squared test. For the continuous model we use Pearson's correlation coefficient between the mean estimated age and the actual age.

4.2 Results

Table 2 shows the significance of correlation for both the dichotomous models and the continuous model using one-sided tests. The fact that the models detected significances also demonstrates that age is a factor in the behaviour of the offender in all but one dataset. Dataset 9 is particularly small and this may explain why no significant correlation was found. The continuous model shows significance in all other cases and therefore outperforms either of the two other models. These models fail to find a significant relationship in burglary either based on age or distance travelled. From these

data we can conclude that the continuous model is capable of finding relationships that a dichotomous model cannot.

Table 2: Significance of correlation (better than 5% shown in bold).

Set No.	Crime Type	Variable	Dichotomous Models – Chi squared		Continuous Model --Pearson
			Multivariate	Multiple Bernoulli	
1	Theft from Vehicles	age	0.015	0.002	0
2	Other Theft	age	0.046	0.022	0.013
3	Shoplifting	age	0	0	0
4	Assault	age	0	0	0
5	Criminal Damage	age	0	0	0.026
6	Damage to Vehicles	age	0.006	0.002	0.001
7	Burglary	age	0.155	0.081	0.007
8	Street Robbery	age	0	0	0
9	Sexual offences against women	age	0.266	0.399	0.353
10	Burglary	distance	0.4	0.537	0

5 Exploring the Contribution of Individual Terms

As mentioned in Section 1, the model proposed here has a degree of transparency so that it is possible to determine what impact each term in the vocabulary has on implying either an older or younger offender. This has two possible uses:

- For any unsolved crime, we can determine whether each term in the document (excluding stopwords) implies younger or older offenders thus shedding some light as to why the model has estimated a given age. It would also indicate whether the terms were concordant in this implication or were giving mixed messages.
- By looking at the offences together, we can indicate which terms tend to imply older or younger behaviour in general. This information has a qualitative application for law enforcers to apply results of the analysis in the field and well as being interest to investigative psychologists.

The first use could be achieved by displaying a crime report with the various terms colour coded depending on the extent that term relates to younger or older behaviour, as shown in Figure 1.



Figure 1: Colour coding of police report to indicate evidence of the sex of the offender.

Here we consider a second use in more detail although both uses require a measure of age sensitivity defined as follows. Given that each term has a probability that is a function of age, we calculate the derivative of the probability with respect to age. A strongly negative value indicates it is more common amongst younger offenders and a strongly positive value will indicate it relates older behaviour. Values around zero suggest that such a feature is not influenced much by age. Where we have terms from free text, then the relationship between identifiable features and words used is a complex one because of synonymy, polysemy and words that do not relate to behaviour at all such as proper nouns. Nevertheless, ranking the vocabulary by this derivative does yield interesting and intuitive results. Rearranging equation 3 gives:

$$P(t_i) = \frac{e^{av+b}}{1 + e^{av+b}} \quad (7)$$

and differentiating it with respect to v yields:

$$\frac{dP(t_i)}{dv} = \frac{ae^{av+b}}{(1 + e^{av+b})^2} \quad (8)$$

There is the problem of serial offences when performing this analysis. A particular term may relate to the behaviour of a single offender. A prolific offender with a unique recurring feature may lead us erroneously to infer that it was common to all offenders of that age. Thus serial crimes were removed so that there was exactly one crime per offender.

Table 3 shows the top and bottom ranked terms for assault. Note that terms starting with a \$-sign are derived from codes and also that *parent*, *spouse* and *offspring* are gender-neutral terms for *mother*, *father* etc. Inspection of the terms shows that whereas young offenders are involved in more serious assaults (e.g. Section 18 assault which can carry life imprisonment) with punching and kicking and attacks involving the head being common features, older offenders are more likely to be involved in domestic incidents resulting in less serious assaults on family members.

Analysis of other data sets reveals interesting patterns too. Younger offenders are more likely to burgle non-residential premises whereas older burglars target homes. Burglars who operate close to their home base are more likely to climb into properties or enter through doors; those who have travelled further tend to enter through windows and then conduct an untidy search. For older offenders, damage to vehicles is more likely to occur as a result of a road rage incident whereas younger offenders are more likely to vandalise an unattended vehicle. A similar pattern emerges for criminal damage where older offenders damage property of an individual with whom they have a dispute. Younger offenders are more likely to engage in ‘victimless’ acts of vandalism (i.e. not directed at an individual) such as spray painting public property. In shoplifting, theft of bottles of spirits (whisky and vodka) is associated with older offenders.

Table 3 – Terms Ranked by Derivative of Probability with regards to Age.

Rank	Lowest Ranked (young)		Highest Ranked (old)	
	Word	Value	Word	Value
1	\$victim-kicked	-0.00323	\$common-assault	0.00203
2	kick	-0.00236	argument	0.00203
3	victim	-0.00186	parent	0.00144
4	\$victim-punched	-0.00171	spouse	0.00130
5	punch	-0.00166	slap	0.00113
6	suspect	-0.00138	offspring	0.00106
7	head	-0.00114	grab	0.00105
8	attack	-0.00110	arm	0.00092
9	\$assault-section-eighteen	-0.00103	domestic	0.00077
10	\$sharp-instrument	-0.00085	\$victim-threatened	0.00075

These results were validated by police officers and considered extremely interesting by the investigative psychologists that were our partners in the iMOV project. They are currently subject to further analysis and validation by field studies.

6 Conclusions

The experimental results show that the continuous language model is able to produce estimates of age and distance for a notionally unsolved crime that correlate significantly with the actual quantities. This model outperforms the alternative dichotomous model. The use of multiple Bernoulli models is appropriate for the nature of the data analysed since the dichotomous Bernoulli model performs as well as the multinomial one. The fact that the models exhibit a degree of transparency is shown to be useful in both explaining an inference of the model but also to identify the different styles of behaviour related to different ages or distances travelled.

Acknowledgements

We would like to thank Prof. David Canter and Dr. Donna Youngs of the Department of Investigative Psychology of the University of Liverpool for many interesting discussions, which inspired the development of the models presented in this paper, and also for providing the data on which the models were tested.

References

1. Bache, R., Crestani F., Canter D., Youngs D., Application of Language Models to Suspect Prioritisation and Suspect Likelihood in Serial Crimes, *International Workshop on Computer Forensics*, pages 399-404, Manchester, UK, 2007.
2. Bache, R., Crestani F., Canter D., Youngs D., Mining Police Digital Archives to Link Criminal Styles with Offender Characteristics, *International Conference on Asian Data Libraries (ICADL)*, pages 493-494, Hanoi, Vietnam, 2007.
3. Bache, R., Crestani F., Estimating Real-valued Characteristics of Criminal from their Recorded Crimes. *The Seventeen ACM Conference on Information and Knowledge Management (CIKM)*, pages 1385-1386, Napa Valley, USA, 2008.
4. Bai J., Nie J., Paradis F., Text Classification Using Language Models. *Asian Information Retrieval Symposium (AIRS)*, Poster Session, Beijing, 2004.
5. Canter, D., Fritzson, K., Differentiating arsonists: A model of firesetting actions and characteristics, *Legal and Criminal Psychology*, vol. 3, pp 73-96, 1998
6. Croft, W.B., Lafferty J., *Language Modeling for Information Retrieval*, Kluwer Academic, 2003.
7. Fellbaum C. (Ed.): WordNet – An Electronic Lexical Database, MIT Press, 1998.
8. Jelinek, F., Mercer, R., Interpolation estimation of Markov source parameters from sparse data. *Workshop on Pattern Recognition in Practice*, Amsterdam, The Netherlands, 1980.
9. Losada D., Language Modeling for Sentence Retrieval: A comparison between Multiple-Bernoulli Models and Multinomial Models, *Information Retrieval Workshop*, Glasgow, Scotland, 2005.
10. McCallum A., Nigam, K., A Comparison of Event Models for naïve Bayes Text Classification, *Proc. AAAI/ICML-98 Workshop on Learning for Text Categorisation*, pages 41-48, Madison, Wisconsin, 1998.
11. Oatley G.G., Ewart B., Crimes Analysis Software: Pins in Maps, Clustering and Bayes Net Prediction. *Expert Systems with Applications*, 25(4):569-588, 2003
12. Peng, F., Schuurmans, D., Combining naive Bayes and n-gram language models for text classification, in *Twenty-Fifth European Conference on Information Retrieval Research (ECIR)*, pages 335-350, Pisa, Italy, 2003.
13. Peng, F., Schuurmans, D., Wang, S. Augmenting Naive Bayes classifiers with statistical language models. In *Information Retrieval*, 7(3):317-345, 2003.
14. Ponte J.M., Croft W.B., A Language Modeling Approach to Information Retrieval, in *Proceedings of the Twenty First Conference on Information Retrieval Research (SIGIR)*, Melbourne, Australia, page 275-281, 1988.

CLEF Ad-hoc: A Perspective on the Evolution of the Cross-Language Evaluation Forum

Nicola Ferro¹ and Carol Peters²

¹ Department of Information Engineering, University of Padua, Italy
ferro@dei.unipd.it

² ISTI-CNR, Area di Ricerca – 56124 Pisa – Italy
carol.peters@isti.cnr.it

Abstract. *MultiLingual Information Access (MLIA)* is a key topic in Digital Libraries. In the last decade the *Cross-Language Evaluation Forum (CLEF)* has stimulated system research and development in this field through the organization of large-scale evaluation campaigns. We discuss the achievements of CLEF in the light of the evolution of one of its main tasks, the Ad Hoc track, which studies multilingual document retrieval, and include details of experimentation with collections provided by The European Library.

1 Introduction

MultiLingual Information Access (MLIA) is a key topic in Digital Libraries. This is the reason why the European Commission, and more specifically the unit for Digital Libraries has sponsored the activity of the *Cross-Language Evaluation Forum (CLEF)*³ over the last decade. CLEF is the major evaluation initiative for the experimentation and testing of MLIA systems operating on European languages.

This paper shows how CLEF has stimulated system research and development in this field by focusing on the evolution of one of its core tasks, the Ad Hoc track, which studies techniques for multilingual document retrieval on document collections in multiple languages and on different genres: news data and library catalog cards from the archives of The European Library. Our efforts over the years have resulted in a wide coverage of the building blocks for multilingual system development (e.g. tools, components, resources and lexicons).. We feel that it is now time to shift the focus of our activity to acquiring a deeper understanding of the underlying issues.

This change in direction is helped by the launching in 2008 by the European Commission of the TrebleCLEF Coordination Action⁴ which intends to promote research, development, implementation and industrial take-up of multilingual, multimodal information access functionality [1]:

³ <http://www.clef-campaign.org/>

⁴ See <http://www.trebleclef.eu/>

- by continuing to support the annual CLEF system evaluation campaigns with tracks and tasks designed to meet the requirements of the user and application communities;
- by constituting a scientific forum for the MLIA community of researchers enabling them to meet and discuss results and new directions;
- by providing a central reference point for anyone interested in studying or implementing MLIA functionality.

In the following sections we briefly introduce the CLEF evaluation campaigns, describe the evolution of the Ad-hoc track, and provide an outlook for the future.

2 The CLEF Evaluation Campaigns

CLEF actually began life in 1997 as a track for *Cross Language Information Retrieval (CLIR)* within the *Text REtrieval Conference (TREC)* organized in the US by NIST and DARPA⁵. The aim was to provide researchers with an infrastructure for evaluation that would enable them to test their systems and compare the results achieved using different cross-language strategies [2]. However, after three years within TREC, it was decided that Europe was better suited for the coordination of an activity that focused on multilingual aspects of information retrieval. A major motivation for this decision was that it was far easier in Europe to find the people and groups with the necessary linguistic competence to handle the language-dependent issues involved in creating test collections in different languages.

While the first efforts within TREC concentrated on assessing the performance of cross-language systems in which queries in one language were matched against target collections in another, CLEF and *NII-NACSIS Test Collection for IR Systems (NTCIR)*⁶ have taken the concept of “cross-language system evaluation” much further by also including monolingual retrieval in multiple languages and truly multilingual retrieval, i.e. retrieval against target collections containing documents in several languages, in their evaluation exercises.

When we launched CLEF in 2000, our focus was on text and document retrieval. However, over the years our scope has gradually expanded to include different kinds of text retrieval across languages (ie not just document retrieval but question answering and geographic IR as well) and different kinds of media (i.e. not just plain text but collections also containing images and speech). The goal has been not only to meet but also to anticipate the emerging needs of the R&D community and to encourage the development of next generation multilingual IR systems. This has meant that the number of tracks offered by CLEF has increased over the years, from just two in 2000 to nine separate tracks in 2008. Most tracks offer several different tasks and these tasks normally vary each year,

⁵ See <http://trec.nist.gov/>

⁶ See <http://research.nii.ac.jp/ntcir/>

according to the interests of the track coordinators⁷ and participants. Figure 1 shows when tracks have been introduced and when they have been terminated.

3 Multilingual Textual Document Retrieval (Ad Hoc)

The Ad Hoc track is considered as our core track. It is the one track that has been offered each year, from 2000 through 2008, and will be offered again in 2009. The aim of this track is to promote the development of monolingual and cross-language textual document retrieval systems through the exploration of a comprehensive set of CLIR-related topics:

- **experimental collections:** test collections are built for as many European languages as possible, attempting to cover diverse language typologies;
- **tasks:** groups are stimulated to experiment with retrieval over unusual pairs of languages and retrieval from collections of multiple languages with diverse characteristics, such as long documents, sparse information, and so on;
- **linguistic resources:** the development and/or use of language resources, such as stop lists, dictionaries, lexicons, aligned and parallel corpora, etc., is supported;
- **linguistic components:** the development and/or application of linguistic tools, such as stemmers, lemmatizers, decompounders, part of speech taggers, and so on, is fostered;
- **translation approaches:** groups are encouraged to experiment with different approaches for crossing language barriers, such as *Machine Translation (MT)*, and dictionary-based, parallel corpora-based, or conceptual network-based translation mechanisms;
- **IR models:** different models are studied and applied – boolean, vector space, probabilistic, language models, and so on – to improve retrieval performances across languages;
- **advanced IR techniques:** advanced techniques, such as data fusion and merging or relevance feedback, are adopted to address issues such as the need for query expansion to improve translation or the fusion of multilingual results;
- **metrics and evaluation techniques:** metrics to analyse system behaviour in a multilingual setting and compare performances across languages and tasks are developed and employed.

From 2000 - 2007, the track exclusively used collections of European newspaper and news agency documents and worked hard at offering increasingly complex and diverse tasks, adding new languages each year. As can be seen from Table 1, the different ad-hoc tasks present varying degrees of difficulty: there are more basic tasks, such as the monolingual tasks or the bilingual English tasks,

⁷ It is impossible to acknowledge all the research organisations that are involved in the coordination of CLEF. A complete list can be found on the homepage of the CLEF website at <http://www.clef-campaign.org/>.

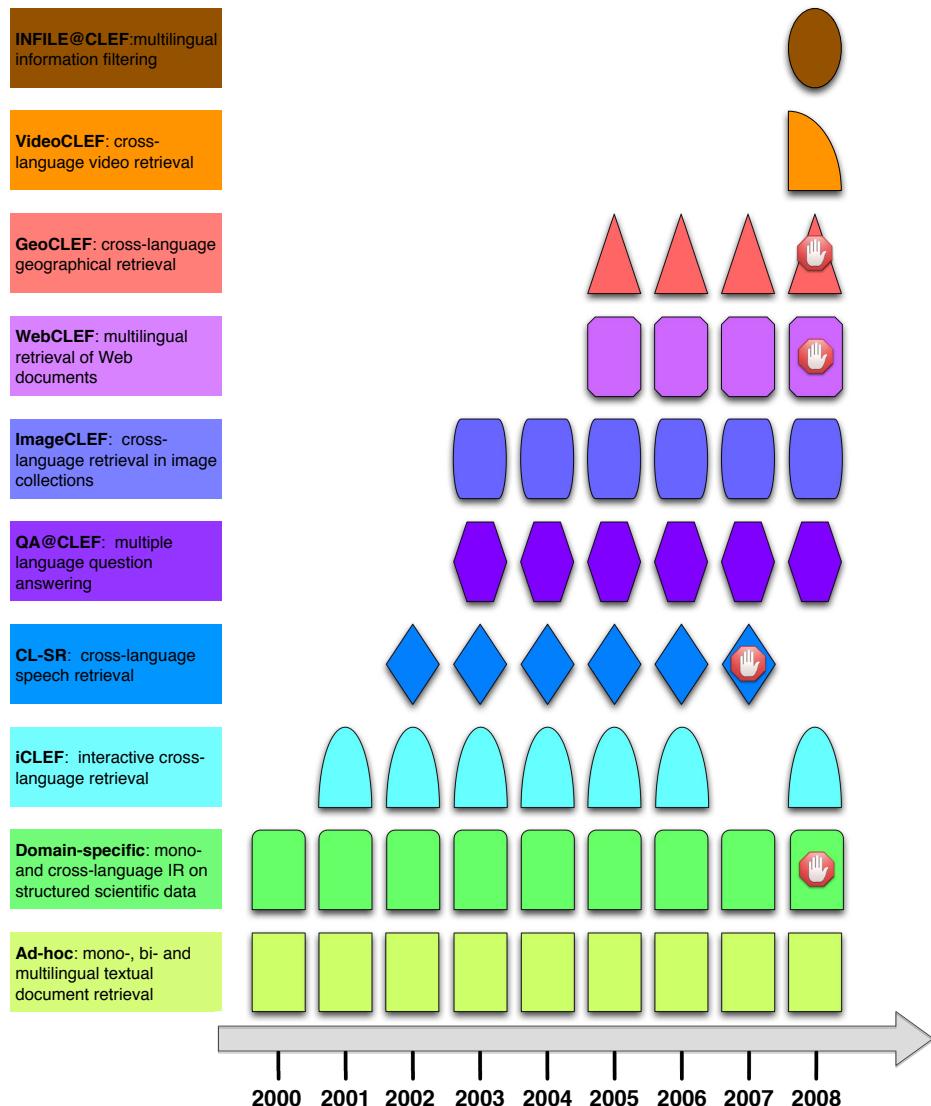


Fig. 1. CLEF 2000 – 2008 tracks. Full details of the activities and results of each track can be found on the CLEF website at <http://www.clef-campaign.org/>.

Table 1. CLEF 2000–2008 Ad Hoc Tasks. The following ISO 639-1 language codes have been used: **am**=Amharic; **bg**=Bulgarian; **bn**=Bengali; **de**=German; **en**=English; **es**=Spanish; **fa**=Farsi; **fi**=Finnish; **fr**=French; **hi**=Hindi; **hu**=Hungarian; **id**=Indonesian; **it**=Italian; **mr**=Marathi; **nl**=Dutch; **or**=Oromo; **pt**=Portuguese; **ru**=Russian; **sv**=Swedish; **ta**=Tamil; **te**=Telugu.

	Monolingual	Bilingual	Multilingual
CLEF 2000	de;fr;it	x-en	x-de;en;fr;it
CLEF 2001	de;es;fr;it;nl	x-en x-nl	x-de;en;es;fr;it
CLEF 2002	de;es;fi;fr;it;nl;sv	x-de;es;fi;fr;it;nl;sv x-en (newcomers only)	x-de;en;es;fr;it
CLEF 2003	de;es;fi;fr;it;nl;ru;sv	it-es de-it fr-nl fi-de x-ru x-en (newcomers only)	x-de;en;es;fr x-de;en;es;fi;fr;it;nl;sv
CLEF 2004	fi;fr;ru;pt	es;fr;it;ru-fi de;fi;nl;sv-fr x-ru x-en (newcomers only)	x-fi;fr;ru;pt
CLEF 2005	bg;fr;hu;pt	x-bg;fr;hu;pt	Multi8 Zyrsion (as in CLEF 2003) Multi8 Merge (as in CLEF 2003)
CLEF 2006	Robust de;en;es;fr;it;nl	x-bg;fr;hu;pt am;hi;id;te;or-en Robust it-es fr-nl en-de	Robust x-de;en;es;fr;it;nl
CLEF 2007	bg, cz, hu Robust en;fr;pt	x-bg;cz;hu am;id;or;zh-en bn;hi;mr;ta;te-en Robust x-en;fr;pt	
CLEF 2008	fa TEL de;en;fr Robust WSD en	en-fa TEL x-de;en;fr Robust WSD es-en	

designed to encourage inexperienced groups to experiment and increase their knowhow; there are intermediate tasks, such as the bilingual task with unusual pair of languages, where groups can try to apply more advanced techniques or experiment their own consolidated techniques in a more challenging scenario; finally, there are advanced tasks, such as the multilingual and robust tasks, where groups have to address difficult issues and discover innovative solutions. In this way, over the years, we have offered different entry points to the fields of CLIR and MLIA in order to support the creation and growth of a research community with diversified expertise.

The results have been considerable; it is probably true to say that this track has done much to foster the creation of a strong European research community in the CLIR area. It has provided the resources, the test collections and also the forum for discussion and comparison of ideas and results. Groups submitting experiments over several years have shown flexibility in advancing to more complex tasks, from monolingual to bilingual and multilingual experiments. Much work has been done on fine-tuning for individual languages while other efforts have concentrated on developing language-independent strategies.

There is also substantial proof of significant increase in retrieval effectiveness in multilingual settings by the systems of CLEF participants. [3] provides a comparison between effectiveness scores from the 1997 TREC-6 campaign and the

CLEF 2003 campaign in which retrieval tasks were offered for eight European languages. While in 1997 systems were performing at about 50%–60% of monolingual effectiveness for multilingual settings, that figure had risen to 80%–85% by 2003 for languages that had been part of multiple evaluation campaigns. In the recent campaigns, we commonly see a figure of about 85%–90% for most languages.

In 2008 there was a big change in focus in this track and we started to move from a breadth-wise exploration of the CLIR field to a deeper investigation of each specific area with the objective of acquiring a more profound understanding of the basic mechanisms. To this end, we introduced very different document collections, a non-European target language, and an *Information Retrieval (IR)* task designed to attract participation from groups interested in *Natural Language Processing (NLP)*. The track was thus structured in three distinct streams.

The first task offered monolingual and cross-language search on library catalog records. It was organized in collaboration with The European Library (TEL)⁸ and used three collections from the catalogs of the British Library, the Bibliothèque Nationale de France and the Austrian National Library. The underlying aim was to identify the most effective retrieval technologies for searching this type of very sparse multilingual data. In fact, the collections contained catalog records in many languages in addition to English, French or German. The task presumed a user with a working knowledge of these three languages who wants to find documents that can be useful for them in one of the three target catalogs. Records in other languages were counted irrelevant.

The TEL task is an example of how the ad-hoc track has started to explore finer-grained questions in the CLIR scenario. Indeed the sparsity of the data and their intrinsic multilinguality are particularly challenging from a retrieval point of view, since the catalog records have to be suitably processed and expanded and the intrinsic multilinguality of the collection has to be catered for with techniques that go beyond the traditional fusion strategies adopted in previous multilingual tasks. It is not expected that new language resources or linguistic tools will be produced in this task but rather that already existing ones will be exploited. The TEL task represents an example of a task that focuses more on retrieval issues than on language or linguistic aspects and is further evidence that cross-language information retrieval is much more than simple machine translation plus information retrieval. Exploiting the CLIR acronym, we could say that it is a **CLIR** task, meaning that it stresses the importance of the retrieval techniques in a multilingual setting.

The Persian@CLEF activity was coordinated in collaboration with the Database Research Group (DBRG) of Tehran University. It was the first time that CLEF offered a non-European language target collection. We chose Persian for several reasons: its challenging script (a modified version of the Arabic alphabet with elision of short vowels) written from right to left; its complex morphology (extensive use of suffixes and compounding); its political and cultural impor-

⁸ See <http://www.theeuropeanlibrary.org/>

tance. This task focuses on the creation of new experimental collections to develop both new linguistic resources (lexicons, dictionaries, and so on) and new linguistic components (stop lists, stemmers, part of speech taggers, and so on). From a retrieval point of view, the necessary techniques are well-known and it is not expected that participants produce new IR components. In this case, we could say that this is a **CLIR** task.

The robust task ran for the third time at CLEF 2008. This year it used English test data from previous campaigns but, in addition to the original documents and topics, the organizers provided *Word Sense Disambiguated (WSD)* documents and topics. Both monolingual and bilingual experiments (topics in Spanish) were activated. This task focuses on the benefits that a deeper and more sophisticated linguistic analysis can produce in a multilingual setting, especially when hard topics are being handled and the aim is to achieve robust performances across the set of topics. As in the previous case, it is not expected that participants produce new IR components. On the other hand, the development and adoption of word sense disambiguation algorithms and their introduction into a consolidated retrieval pipeline puts attention on the linguistic part of the process. Again, In this case, we define this as a **CLIR** task.

This deeper investigation will be taken a step further in the Grid@CLEF Pilot task⁹ which has been proposed for CLEF 2009 with the following goals in mind [4]:

- to look at differences across a wide set of languages;
- to identify best practices for each language;
- to help other countries to develop their expertise in the IR field and create IR groups.

Indeed, individual researchers or small groups do not usually have the possibility of running large-scale and systematic experiments over a large set of experimental collections and resources. It is our hypothesis that a series of systematic grid experiments can re-use and exploit the valuable resources and experimental collections made available by CLEF in order to gain more insights about the effectiveness of, for example, the various weighting schemes and retrieval techniques with respect to the languages. This knowledge could then be disseminated to both the research and the application communities.

In order to run these grid experiments, we need to set up a framework in which participants can exchange the intermediate output of the components of their systems and create a run by using the output of the components of other participants. For example, if the expertise of participant A is in building stemmers and decompounds while participant B's expertise is in developing probabilistic IR models, we would like to make it possible for participant A to apply his stemmer to a document collection, pass the output to participant B, who tests his probabilistic IR model, thus obtaining a final run which represents the result of testing participant A stemmer + participant B probabilistic IR model.

⁹ <http://ims.dei.unipd.it/gridclef/>

The Pilot Grid task in CLEF 2009 will provide a framework for a first set of experiments which will allow us to start to explore the interaction among IR components and languages. This initial knowledge will allow us to tune the overall protocol and framework, to understand what directions are more promising, and to scale the experiments up to a finer-grain comprehension of the behaviour of IR components across languages.

4 Conclusions

In this paper, we have discussed the evolution of the CLEF by focussing on the activities carried out in the Ad Hoc track. Much of the effort of CLEF over the years has been devoted to the investigation of key questions such as “What is CLIR?”, “What areas should it cover?” and “What resources, tools and technologies are needed?” since CLEF began when CLIR was just starting to be recognized as an independent sub-discipline and thus promoted much pioneering work in the field.

Now, we are in the position of conducting a much deeper investigation of the core issues of the field and we will focus on:

- in-depth analyses on how the various components of MLIA systems (stemmers, IR models, relevance feedback, translation techniques) behave with respect to languages;
- the organization of evaluation exercises modeled on the results of MLIA user profiling studies;
- transfer of the research results to the relevant applications.

Acknowledgements

The work reported has been partially supported by the TrebleCLEF Coordination Action, FP7 ICT programme for Digital Libraries and Technology-enhanced Learning. Grant agreement: 215231.

References

1. Braschler, M., Di Nunzio, G.M., Ferro, N., Gonzalo, J., Peters, C., Sanderson, M.: From CLEF to TrebleCLEF: promoting Technology Transfer for Multilingual Information Retrieval. In Second DELOS Conference - Working Notes (2007)
2. Harman, D.K., Braschler, M., Hess, M., Kluck, M., Peters, C., Schäuble, P., Sheridan, P.: CLIR Evaluation at TREC. In Cross-Language Information Retrieval and Evaluation: Workshop of Cross-Language Evaluation Forum (CLEF 2000), LNCS 2069, Springer (2001) 7–23
3. Braschler, M.: Combination Approaches for Multilingual Text Retrieval. Information Retrieval **7**(1/2) (2004) 183–204
4. Ferro, N., Harman, D.: Dealing with MultiLingual Information Access: Grid Experiments at TrebleCLEF. In Post-proceedings of the Fourth Italian Research Conference on Digital Library Systems (IRCDL 2008), ISTI-CNR at Gruppo ALI (2008) 29–32

Towards an Integrated Approach to Music Retrieval

Emanuele Di Buccio¹, Ivano Masiero¹, Yosi Mass², Massimo Melucci¹, Riccardo Miotto¹, Nicola Orio¹, and Benjamin Sznajder²

¹ Department of Information Engineering – University of Padova
Padova, Italy

{dibuccio,masieroi,melo,miottori,orio}@dei.unipd.it

² IBM Research Lab
Haifa, Israel
{yosimass,benjams}@il.ibm.com

Abstract. This paper describes a research work on peer-to-peer music search based on the combination of content based descriptors and textual metadata. The envisaged scenario is the one of a user who searches for documents according either to their melodic and rhythmic content, and to additional information about the title, the instrumentation, or tempo in the form of textual metadata. Two different overlay networks are used to deal with music content and text. A user interface has been developed, which allows the user to perform a query, to merge the results using alternative approaches, and to listen to the retrieved music documents.

1 Introduction

The main goal of SAPIR³ (Search in Audio-visual content using Peer-to-peer Information Retrieval), a project funded by the EU, is the development of a large-scale, distributed peer-to-peer (P2P) infrastructure to allow users searching in audio-visual content using a *Query By Example* paradigm. According to this approach, the user queries the system using an example of what he is looking for, possibly with additional metadata in order to better describe his information need. SAPIR aims at providing large scale search capabilities in P2P network for different types of media. The SAPIR vision is illustrated in Fig. 1 and it includes a media analysis frameworks for different media – images, video, speech, music, and text – together with scalable and distributed P2P index structures supporting similarity search and support for multiple devices embedding social networking in a trusted environment. In this paper we focus on a particular component of the SAPIR architecture, that is a music search based on content descriptors and textual metadata.

Music search is increasingly gaining interest because of the wide diffusion of music files in P2P networks. Yet, few P2P architectures allow for a content based search of music files. One major problem in music searching over a P2P network

³ <http://www.sapir.eu>

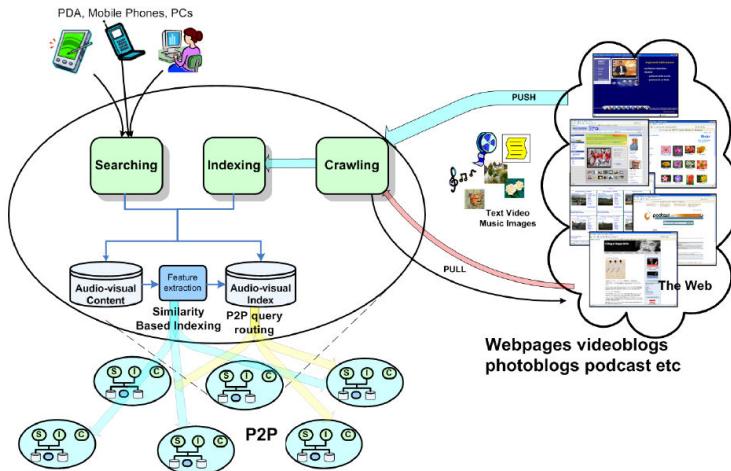


Fig. 1. SAPIR Components and Functions

is, as in the general case, the lack of knowledge of the content location. To overcome this problem, the work reported in [1] proposes a DHT-based system applied to music retrieval. The system exploits both manually specified tag-like information – e.g. artist, album, title – and automatic feature extraction techniques to search by content.

Even though structured networks allow for efficient query routing, they require an high degree of collaboration among peers. The latter may not be a suitable solution for networks that are highly dynamic, heterogeneous, or protective of intellectual property. This kind of networks is well-matched by unstructured overlays. To this end, two P2P music retrieval systems for unstructured networks have been proposed, in [2] and in [3] respectively. Yet, both approaches lack in terms of P2P searching algorithms: the former routes the query to all the peers, whereas the latter exploits a breadth-first search algorithm. The approach proposed in [4] utilizes a centralized coordinator that stores the identifiers of all the peers in the network, and for each peer also its PC-feature, that is a feature which describes the music content of the peer and that is used to select the most promising peers to answer to a given query. A more efficient solution in terms of network traffic can be obtained by decreasing decentralization.

In this paper we present an approach to P2P music search based on the combined use of descriptive metadata and content features. The approach has been developed as part of the SAPIR project, according to one of the envisaged scenarios – called *music and text* – where a user interacts with the system using a query by example paradigm but is also able to provide additional information about his information need in the form of textual metadata.

2 Description of Music Documents

We propose to retrieve music documents through the integration of two alternative descriptions: textual metadata and content features. Both descriptions are automatically extracted from the documents themselves, without using external sources. The approach has been developed and tested on a collection of about 60,000 music files in MIDI format [5].

MIDI files have been downloaded through a focused crawling on a number of Web sites that granted free access for non commercial usages. Files are typically provided by end users and thus there is no control on their quality. Moreover, there could be a replication of the same songs, sequenced and provided by different users, often with slightly different names. Exact copies may also be present in different Web sites, although this happened for less than 5% of the files. Given all these characteristics, the music collection can be considered a good approximation of the content that can be found in typical P2P network.

2.1 Textual Metadata

Two kinds of metadata can be found in MIDI files: textual metadata and coded information useful for the representation of the music score. Textual metadata is directly embedded in the file format by the user who sequenced the file, in order to provide additional information about the file content. It is important to notice that this information is not displayed on screen by most of the available music players, and thus is used freely by the users and sometimes is automatically added by the sequencing software. Textual information is structured in a number of fields that are defined by the MIDI standard, which includes: the name of the different music tracks that may be present in the file, the name of the instruments associated to each track, copyright information, and the markers that are added to a given position in the music score – which are often used to insert the lyrics.

Unfortunately, most of the MIDI files available on the Web are provided by end users on a voluntary basis, and thus there is no guarantee that the different fields for structuring metadata are used consistently. For instance, an analysis on some samples of our music collection showed that the copyright field may contain either the name of the artists, or the name of the software used to generate the file, or even the URL of the Web site where the file has been made available. Analogously, the information about the tracks can be used either to represent the name of the performer in the original recording, or the name of the played instrument, or other information unrelated to the track.

The information about artist and title, which is usually not included in the MIDI format – there is actually no field to represent information about title and artist in the MIDI standard – has been extracted by analyzing the structure of the Web pages containing the links to the MIDI files and the text surrounding the links to the music files. In about 90% of the cases we were able to identify either the title alone or the title and artist. Being provided by the Web sites, the information about title and artist is likely to be more reliable than textual metadata added by the user.

The second kind of metadata that can be extracted from MIDI files regards information strictly related to the music language. It can be computed through the automatic analysis of the music content of the files and thus is somehow more reliable than textual metadata. We choose to extract some general information about the main features of a music piece, including: tonality (e.g., C major or E flat minor), time division (e.g., 4/4 or 3/4) and tempo in beats per minutes. The latter information has been represented using a simple perceptual scale from “very slow” to “very fast”. Moreover, the information about the sound samples to be used during playback – it is worth noting that all MIDI players synthesize the music by using a bank of internal sound samples – has been used as an alternative description of the music instruments that are associated to each track (in general, this information can be different from the one manually provided by the user as textual metadata).

The use of music terms may not be particularly easy for a user without a music education. The fact that a song is in a given tonality or has a particular time division may not give additional information about the relevance to a given information need. Nevertheless, there are applications where such information can be relevant. For instance, in case that professional user is interested to create a playlist of songs or to mix different songs to create a new music product, or when a musician is interested in retrieving songs that are more suitable for the tonality of his instrument.

2.2 Content-based Descriptors

The approach to content description of music documents is based on the use of high level features, which should be perceptually relevant for the final users. According to the approaches presented in the literature on music information retrieval, the most relevant content descriptors are the rhythm and the melody of the leading voice. This information is readily available in the case of MIDI files, while it can be computed from audio documents using signal processing techniques achieving an accuracy of more than 85% in onset detection and note description. In this paper, we focus on the use of MIDI format.

The first step regards the automatic identification of the track containing the main melody, which has to be analyzed in order to extract the relevant melodic information. It has to be noted that a MIDI file contains a number of tracks, related to the different instruments that are employed in the overall music score. The approach to identify the main track, which achieves an accuracy of 97.7%, is described in [6]. In case the selected track was polyphonic, it has been transformed to monophonic using the approach described in [7].

The melodic information has been quantized, in order to take into account local variations of pitch and tempo. To this end, the fundamental frequency values has been quantized to the 12 semitones of the chromatic scale used in Western music, while rhythm has been normalized using state of the art techniques for music transcription. After quantization and normalization, melodic information was described by two main parameters:

- Pitch intervals (PIT): that is the distance between two subsequent tones.
- Interonset interval log-ratio (IOI): that is the ratio, in logarithmic scale between the duration of two subsequent tones.

The second step regards the segmentation of the extracted melody in musical lexical units, which are used as content descriptors. To this end, pattern analysis techniques have been applied to the sequence of symbols forming either the melody or the rhythm, highlighting all different patterns with a length from 3 to 6 notes, as described in [8]. These thresholds have been experimentally evaluated with the test collection available for the Music Information REtrieval Evaluation eXchange campaign [9].

The final step in the representation regards a suitable coding of the patterns, which differs according to the features to be represented. Given that all features undergo quantization, a textual representation is used to describe the patterns, by assigning a different symbol, taken from a small alphabet, to all the elements in melodic or rhythmic music patterns. The size of the alphabet depends on the choice of the quantization step. Commonly used values are 15 symbols for PIT and 9 symbols for IOI. The relative frequency with which a pattern occurs in a documents has been computed as well for each feature.

3 Music Retrieval in SAPIR

The information extracted from music documents is indexed using two different SAPIR overlays: a text overlay developed by IBM, and a content based music overlay, called SPINA, developed at the University of Padova [10]. The two overlays index the same collection of music documents in MIDI format. The file URI is used as unique identifier for the music documents.

Document content is represented using a common framework. The SAPIR approach aims at making use of standards concerning metadata representation as well as content analysis methods: metadata are expressed in XML derived from MPEG-7 (ISO/IEC 15938), while the UIMA framework [11] has been chosen to analyze content and extract its relevant features. Additional information about content analysis and representation can be found in [12]. Music information is represented as a set of melodic and rhythmic patterns, as described in Sect. 2.2. The textual metadata extracted from the music files is indexed using the structured fields as described in Sect. 2.1 and can be retrieved either by using the structure or as free text. Given the considerations made in Sect. 2.1 the UI is more oriented towards queries with unstructured text.

In order to use the textual information as real metadata, in the present implementation the user has to format the query using a XML-like representation, which is a simplified version of MusicXML [13]. For instance, if a user is interested in searching music documents played with a fast tempo and where the electric guitar is one of the instruments, he can either explicitly express his information need with a structured query like `<tempo>fast</tempo> <instrument>electric guitar</instrument>` or simply using a free text query such as *fast electric guitar*.

A stand-alone Java application implementing a SAPIR peer has been developed, providing a user interface (UI) for creating complex queries that combine music content and textual metadata. The UI allows the user:

- To upload a music file to be used as an example of his information need.
- To insert a textual query, possibly using the musical terms that describe the music content (e.g., tonality and tempo)
- To combine the content based description with the textual metadata.

The peer is connected to a SPINA super peer that acts as the entry point to the SPINA overlay and to a text overlay. Once a user creates a query using the interface of the peer, the query is parsed by a component called *query analyzer*, which parses the XML representation of the combined query – music content plus textual metadata (in the terminology proposed in SAPIR these are two *FeatureGroups*) and sends the music content features to the SPINA overlay and the textual part to the text overlay. It is important to note that SAPIR architecture allows for more complex combinations of features. This characteristics could be exploited in case the music overlay were combined with other media, such as e.g. video.

3.1 Merging of the Results

Each overlay processes independently the textual and music components of the query. The individual results can be merged using different strategies. For the particular application of music retrieval, one possible scenario is of a user who is mainly interested in retrieving music documents that are similar to the one provided as a query, while the textual information can be used only to refine the results. Alternatively, the user could be interested in obtaining two different lists of potentially relevant documents, one for the music features and one for the textual part, and a third list with the results merged together. A screenshot of the music and text peer is shown in Fig. 2.

Given these considerations, at the moment the SPINA peer presents the retrieval results through four different lists of documents:

1. A list of music documents retrieved by the SPINA music overlay alone, with additional information about the peer that actually contains the document; the user can download and listen to the music file by simply clicking on an icon, because the SPINA peer can directly contact the peer containing the file.
2. A list of music documents retrieved by the text overlay; in this case it is not possible to listen to the music files, which are stored as XML documents containing the metadata; a possible extension of the functionalities may regard the possibility to access to the textual metadata, in order to have a better description of the file content.
3. A merged list of music documents, retrieved either by the SPINA and the text overlays. The list is obtained by applying data fusion techniques. Given that in this list there could be documents retrieved only by the text overlay, the user can listen only to a limited number of documents in this list.

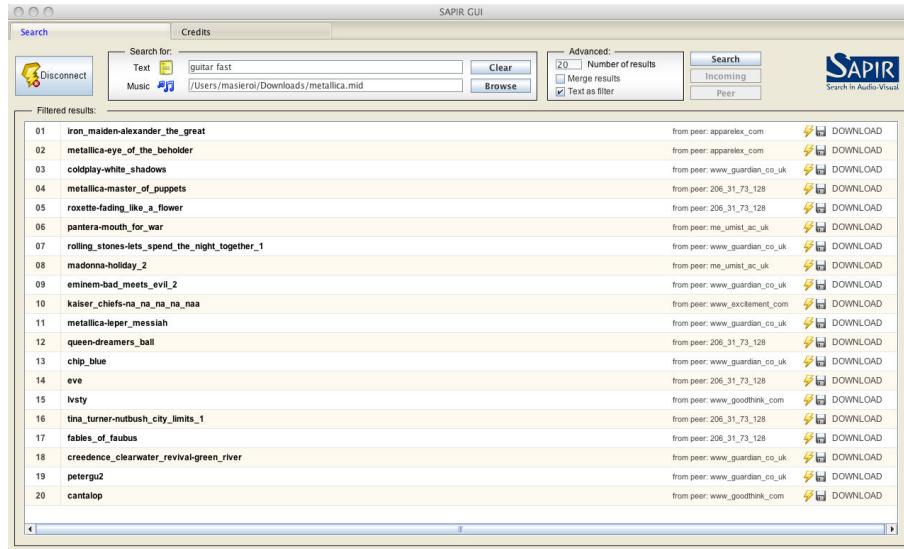


Fig. 2. Screenshot of the user interface for the music and text peer

4. An alternative merged list, where the documents retrieved by the SPINA overlay are reranked according to the results of the text overlay. In this case, which is probably the most useful for a user who needs to evaluate the results by listening to the retrieved documents, the user can listen to all the documents in the list.

The fact that a user is presented with a single rank list, including the query results on the two overlays as for case 3 and case 4, is an important step towards an integrated approach to retrieve music documents. For case 3, it can be noted that, being based on different retrieval schemes, the two overlays may retrieve different music documents and the data fusion approach allows for giving a high rank to music documents that are both similar to the music query and relevant for the textual query. At the same time, also music documents that are particularly relevant either for their content or for their metadata can have a high final rank. For case 4, the reranking approach of documents retrieved by the SPINA overlay allows the user for a content based search where additional textual information can be used to give an higher rank to documents with the required metadata.

4 Conclusions

The approach has been tested using a collection of about 60,000 MIDI files. Some files were replicas of existing ones, in order to simulate a P2P network. The approach showed to be scalable in the number of documents. Further analysis

need to be carried out on the effectiveness of the content based description of music documents. Preliminary results using a centralized system and a small test collection of about 600 music documents in MIDI format showed that the method for music retrieval can achieve a mean average precision of 0.54. Future work will address the effects of a distributed collection on retrieval effectiveness.

Acknowledgments

This work was partially supported by the SAPIR project, funded by the European Commission under IST FP6 (Sixth Framework Programme, Contract no. 45128). The authors thank Maristella Agosti, Nicola Ferro and Giorgio Di Nunzio for their valuable support in the development of the proposed approach.

References

1. Tzanetakis, G., Gao, J., Steenkiste, P.: A scalable peer-to-peer system for music content and information retrieval. In: Proceedings of the International Conference on Music Information Retrieval. (2003)
2. Yang, C.: Peer-to-peer architecture for content-based music retrieval on acoustic data. In: Proceedings of the International Conference on World Wide Web. (2003) 376–383
3. Karydis, I., Nanopoulos, A., Papadopoulos, A., Manolopoulos, Y.: Musical retrieval in p2p networks under the warping distance. (2006)
4. Wang, C., Li, J., Shi, S.: A kind of content-based music information retrieval method in peer-to-peer environment. In: Proceedings of the 3rd International Conference on Music Information Retrieval. (2002)
5. Rothstein, J.: MIDI: A comprehensive introduction. A-R Editions, Madison, WI (1991)
6. Orio, N., Zen, C.: Song identification through hmm-based modeling of the main melody. In: Proceedings of International Computer Music Conference. (2007) 248–251
7. Uitdenbogerd, A., Zobel, J.: Manipulation of music for melody matching. In: Proceedings of the ACM Conference on Multimedia. (1998) 235–240
8. Neve, G., Orio, N.: Indexing and retrieval of music documents through pattern analysis and data fusion techniques. In: Proceedings of the International Conference on Music Information Retrieval. (2004) 216–223
9. Mirex-2006: Second annual Music Information Retrieval Evaluation eXchange (July 2006) <http://www.music-ir.org/mirex2006/>.
10. Di Buccio, E., Ferro, N., Melucci, M.: Content-based information retrieval in SPINA. In: Proceedings of the Italian Research Conference on Digital Library Systems. (2008)
11. Ferrucci, D., Lally, A.: UIMA: an architectural approach to unstructured information processing in the corporate research environment. Natural Language Engineering **10**(3-4) (2004) 327–348
12. Allasia, W., Falchi, F., Gallo, F., Kacimi, M., Kaplan, A., Mamou, J., Mass, Y., Orio, N.: Audio-visual content analysis in p2p: the sapir approach. In: Proceedings of the AEIMPro08 Workshop. (2009)
13. MusicXML: Recordare: Internet music publishing and software (December 2008) <http://www.musicxml.org/>.

Searching 100M Images by Content Similarity

Paolo Bolettieri¹, Fabrizio Falchi¹, Claudio Lucchese¹, Yosi Mass²,
 Raffaele Perego¹, Fausto Rabitti¹, Michal Shmueli-Scheuer²

¹ ISTI-CNR, Pisa, Italy

² IBM Haifa Research Lab, Israel

Abstract. In this paper we present the web user interface of a scalable and distributed system for image retrieval based on visual features and annotated text, developed in the context of the SAPIR project. Its architecture makes use of Peer-to-Peer networks to achieve scalability and efficiency allowing the management of huge amount of data and simultaneous access by a large number of users. Describing the SAPIR web user interface we want to encourage final users to use SAPIR to search by content similarity, together with the usual text search, on a large image collection (100 million images crawled from Flickr) with realistic response time. On the ground of the statistics collected, it will be possible, for the first time, to study the user behavior (e.g., the way they combine text and image content search) in this new realistic environment.

1 Introduction: the SAPIR Project

Non-text data, such as images, music, animations, and videos is nowadays a large component of the Web. However, web tools for performing image searching, as the ones provided by Google, Yahoo! and MSN Live Search, simply index the text associated with the image.

Image indexing methods based on content analysis or pattern matching (i.e. features, such as colors and shapes) are usually not exploited at all. In fact, for this kind of data the appropriate search methods are based on similarity paradigms (e.g. range queries and nearest neighbor queries) that are computationally more intensive than text search. The reason is that conventional inverted indexes used for text are not applicable for such data.

The European project SAPIR (Search on Audio-visual content using Peer-to-peer Information Retrieval)¹ aims at breaking this technological barrier by developing a large-scale, distributed Peer-to-Peer infrastructure that will make it possible to search for audio-visual content by querying the specific characteristics (i.e., features) of the content. SAPIR's goal is to establish a giant Peer-to-Peer network, where users are peers that produce audiovisual content using multiple devices (e.g., cell phones) and service providers will use more powerful peers that maintain indexes and provide search capabilities

¹ <http://www.sapir.eu/>

“A picture is worth a thousand words” so using an image taken by a cell phone to find information about e.g. a monument we bump into or singing a melody as a search hint for a full song, combined with optional metadata annotations and user and social networking context will provide the next level of search capabilities and precision of retrieved results.

2 SAPIR Architecture

Although many similarity search approaches have been proposed, the most generic one considers the mathematical metric space as a suitable abstraction of similarity [14]. The metric space approach has been proved to be very important for building efficient indexes for content based similarity searching. A survey of existing approaches for centralized structures (e.g. M-tree), can be found in [14]. However, searching on the level of features exploiting similarity paradigms, typically exploiting range queries and nearest neighbor queries, exhibits linear scalability with respect to the data search size.

Recently scalable and distributed index structures based on Peer-to-Peer networks have also been proposed for similarity searching in metric spaces and are used in the context of the SAPIR project - i.e. GHT*, VPT*, MCAN, M-Chord. These index structures have been proved to provide scalability for similarity search adding resources as the dataset grows (see [2] for a comparison of their performances). Peer-to-Peer architectures are convenient approach and a common characteristic is the autonomy of the peers with no need of central coordination or flooding strategies. Since there are no bottlenecks, the structures are scalable and high performance is achieved through parallel query execution on individual peers.

In SAPIR also text is indexed using a Peer-to-Peer architecture called MINERVA [3]. In MINERVA each peer is considered autonomous and has its own local search engine with a crawler and a local index. Posting meta-information into the Peer-to-Peer network the peers share their local indexes. This meta-information contains compact statistics and quality-of-service information, and effectively forms a global directory. The Peer-to-Peer engine uses the global directory to identify candidate peers that are most likely to provide good query results. More information about MINERVA can be found in [3].

An IR-style query language for multimedia content based retrieval has been developed for SAPIR. It exploits the XML representation of MPEG-7 and it is an extension of the “ML Fragments” query language that was originally designed as a Query-By-Example for text-only XML collections. Detailed information can be found in [10].

In SAPIR it is also possible to perform complex similarity search combining result lists obtained using distinct features, GPS information and text. To this aim, state of the art algorithms for combining results are used (e.g., [6]). In Section 4 combined search algorithms and functions are described.

In SAPIR the possibility of retrieving the results of content-based queries from a *cache* located in front of the system has also been investigated [8]. The

aim is to reduce the average cost of query resolution, thus boosting the overall performance. The used cache is very different from a traditional cache for WSEs. In fact, our cache is able to return an answer without querying the underlying content-based index in two very different cases: (a) an *exact* answer when exactly the same query was submitted in the past, and its results were not evicted from the cache; (b) an *approximate* answer composed of the closest objects currently cached when the quality of such approximated answer is acceptable according to a given measure. For further information see [8].

For the scope of improving throughput and response time, during the SAPIR project a metric cache was developed [7]. Unlike traditional caching systems, the proposed a caching system might return a result set also when the submitted query object was never seen in the past. In fact, the metric distance between the current and the cached objects is used to drive cache lookup, and to return a set of approximate results when some guarantee on their quality can be given.

3 Dataset: CoPhIR

The collection of images we used consists of a set of 100 million objects randomly selected from the CoPhIR collection². CoPhIR is the largest publicly available collection of high-quality images metadata. Each contains five MPEG-7 visual descriptors (*Scalable Color*, *Color Structure*, *Color Layout*, *Edge Histogram*, *Homogeneous Texture*), and other textual information (title, tags, comments, etc.) of about 60 million photos (still increasing) that have been crawled from the Flickr photo-sharing site³.

Since no collection of this scale was available for research purpose, we had to tackle the non-trivial process of image crawling and descriptive feature extraction using the European EGEE computer GRID. In particular, we had the possibility to access the EGEE (Enabling Grids for ESciencE) European GRID infrastructure⁴ provided to us by the DILIGENT IST project⁵.

4 Combined Search: Algorithms and Functions

Queries in SAPIR can combine both image and text. Top-k queries are used to find the best results that match both a given image and a given text. Given a query it is possible to get from the image index and from the text image a list of objects sorted by descending order of relevance to the appropriate query. Top-k queries are usually done by merging those lists into a single ranked result list using some aggregate function over the objects' scores from the different lists.

² <http://cophir.isti.cnr.it> - CoPhIR stands for COntent-based Photo Image Retrieval

³ <http://www.flickr.com>

⁴ <http://www.eu-egee.org/>

⁵ <http://www.diligentproject.org/>

4.1 Merge algorithms

The state-of-the-art solution for merging several lists (also known as the top-k problem) is the family of Fagin's TA (Threshold Algorithm) and NRA (No Random Access) algorithms [5]. Although these algorithms have been proved to be instance optimal, their running time can degrade into complete scans of the input lists. Moreover, we show that their basic form is not appropriate for a P2P setting since they may consume high network bandwidth. In this section we describe briefly those algorithms and then describe various optimizations and extensions we developed in SAPIR, in particular:

- P2P Optimizations to TA
- P2P Optimizations to NRA
- Filtered algorithm

P2P Optimizations to TA Inspired by the state-of-the-art algorithms, we implemented Fagin's TA [5] algorithm with several extensions and optimizations. The TA algorithm defines the notion of sorted and random accesses. In sorted access, the next object in the descending order of scores is retrieved from the list associated, whereas, random access retrieves the score of a random given object from the list. A TA algorithm performs a mixture of sorted and random accesses to the lists. At any time during the execution of such an algorithm, there is complete knowledge of the already seen objects. Given m lists, the algorithm starts with sorted access to list i , "sees" object o , and then performs random accesses to the remaining lists to fetch o 's score, thus having the complete score for o . In addition, the TA maintains the score of the object at the current cursor position for every list i (denoted as $high_i$). An object whose aggregated score is within the best k already seen objects becomes part of the top-k set. The TA terminates when the object with the lowest score in the top-k set is higher than the threshold value defined as the aggregated score of the $high_i$'s.

We now discuss different optimizations and improvements that we applied on top of the TA algorithm.

Sorted access in Batches The TA as described above considers only costs for sorted and random accesses. However, in a peer-to-peer (P2P) environment, one should not ignore the network and communication overhead. Specifically, the overhead comprises the network latency incurred by message rounds and the network bandwidth consumption incurred by the data exchange among the peers. The abovementioned TA execution in a P2P environment will generate communication message to get the next object as well as performing the random accesses, which can result in high overheads. Thus, the first optimization we applied is to reduce the network overhead by a "fetch in batches" execution. As suggested in [12], to reduce network communication, successive Result Objects can be batched into one message; instead of getting only one object every time that the peer contacts a list, it will receive B sorted objects. To support the batched execution in the SAPIR implementation, one of the parameters for the query execution is the batchSize, the size of the result list that a peer wants to fetch.

Random Access in batches In the original TA, the random accesses are done immediately when a new object is seen, means that a communication message is send to the list after each new object. As discussed above, these communication overheads are expensive. Thus, in our implementation, for each list, the random accesses requests are batched into one array and only one communication message is sent.

P2P Optimizations to NRA We now discuss the NRA algorithm [5]. The main assumption in this algorithm is that no random accesses to the lists are allowed; thus, with sorted only access it needs to determine the k best results. The NRA starts with sorted access to the different lists, in each step it sees the next object. Thus, at any time during the execution some objects may have been only partially seen in a subset of lists, so there is some uncertainty about the final score of the object. The algorithm therefore keeps, for each seen object d , two values to bound its final score: $worstScore(d)$ and $bestScore(d)$. $worstScore(d)$ is computed as the sum of the seen scores of d , assuming a score of 0 for the remaining dimensions, and serves as a lower bound for d 's final score. $bestScore(d)$ is computed as the sum of $worstScore(d)$ and the $high_i$ values of lists where d has not yet been seen, where $high_i$ is the value at the current scan position of list i , $bestScore(d)$ is therefore an upper bound for d 's final score. Objects are then kept in two sets: The k objects with the currently highest worstScores form the current top-k answers, and the remaining objects reside in the candidates set. The algorithm can safely stop when the object with the highest bestScore of the candidates set has a bestScore that is smaller than the worstScore of the object with the min worstScore from the top-k set. Similar to the TA case, we applied the "Sorted access in Batch" optimization to the NRA. In addition we applied two more optimizations: Bounded Candidate List and Update Upper Bound Once which are described in the following subsections.

Bounded Candidates List As described above, every object o that does not qualify for the top-k set ($worstScore(o) < worstScore(d)$ where d is the object with the min score from the top-k set) and could not be eliminated ($bestScore(o) > worstScore(d)$), is inserted into the candidates set. However, many of these objects have a very low probability to be qualified for the top-k. Keeping all the objects in the candidates set means maintaining a very large set. The cost of maintaining such a set is $O(n)$ which is not suitable for an online algorithm [13]. Thus, as suggested in [13] we can limit the size of the set and keep only the r (typical r could be 200) best candidates.

Update Upper Bound Once The $bestScore(d)$ value is based on the scores for the unseen lists at the current position ($high_i$). When the NRA algorithm scans the next row, the bestScore of all the relevant objects need to be updated. Again, such updates could impose very high overheads on an online algorithm. It is worth noting, that when the query processor gets the results in batches, it can exploit this situation as follows - whenever an object is seen in one of the lists, it is then immediately probed in the other lists with negligible cost. To update the bestScore efficiently, if the object appears in the remaining lists, the

worstScore and bestScore are updated. However, if not, then the worstScore is set to 0, and the bestScore is set to the lowest score of the list.

Filtered Algorithm The main purpose of this merge algorithm is to improve the efficiency by considering only the results that were returned by one of the indices and then re-rank or filter out the results by the other index. For example the query can be first sent to the image index and then the returned results are sent to the text index to check if the queried text appears in each of the results. This algorithm does not allow the text to introduce results that did not already appear in the image list.

4.2 Aggregate functions

The majority of top-k techniques assume monotonic aggregation functions. Using monotone aggregation functions is common in many practical applications, especially in web settings [11]. Thus, many top-k processing scenarios involve linear combinations of multiple scoring predicates. Specifically, in SAPIR we have implemented the following functions: Sum, Weighted Sum, Fuzzy AND and Fuzzy OR.

The following aggregation functions were implemented in SAPIR.

- Sum: $\sum_{i=0}^{n-1} x_i$
- Weighted Sum: $\sum_{i=0}^{n-1} w_i \cdot x_i$
- Fuzzy AND: $\min_{i=0}^{n-1} (w_i \cdot x_i)$
- Fuzzy OR: $\max_{i=0}^{n-1} (w_i \cdot x_i)$
- Weighted AND: $\begin{cases} \sum_{i=0}^{n-1} w_i \cdot x_i, & \text{if } \forall x_i, x_i \neq 0 \\ 0, & \text{else} \end{cases}$

where n is the number of lists, x_i and w_i are the score and the weight of object x in list i correspondingly. It is worth noting that for the Fuzzy AND we only considered the image score.

The main purpose of supporting different aggregation functions is to give the user high flexibility and sometimes improve the effectiveness as follows.

The AND operations namely, fuzzy and weighted AND, are stricter in the sense that they require that the object will appear in all lists. Objects that appear in both lists basically have more "evidence" so that the probability that it is a good object increases. This is very important in the presence of merging content-based and metadata. Previous works [9, 4] suggested that only content-based image search is not effective enough because of the gap between visual feature representations and metadata such as user tagging and extracted semantic concepts. Thus a combination of content based search with associated



Fig. 1. SAPIR demo homepage

metadata is expected to yield the best results. Nevertheless, when the user has only a broad idea about the results that she wants and if she can tolerate more fuzziness, then aggregation function such as Weighted Sum and Sum might be more adequate.

5 Guided tour of the tool

For both testing and demonstration, we developed a web user interface to search between indexed images. In the following we briefly describe the web user interface which is public available at <http://sapir.isti.cnr.it/>.

In Figure 1 we report a snapshot of the dynamic web page that is used as starting point for searching. From that page it is possible to perform a fulltext search, a similarity search starting from one of the random selected images, a similarity search starting from an image uploaded by the user or a combined search.

In Figure 2 we report a typical results page from which it is possible to: go back to the home page, access the advanced options described before open a

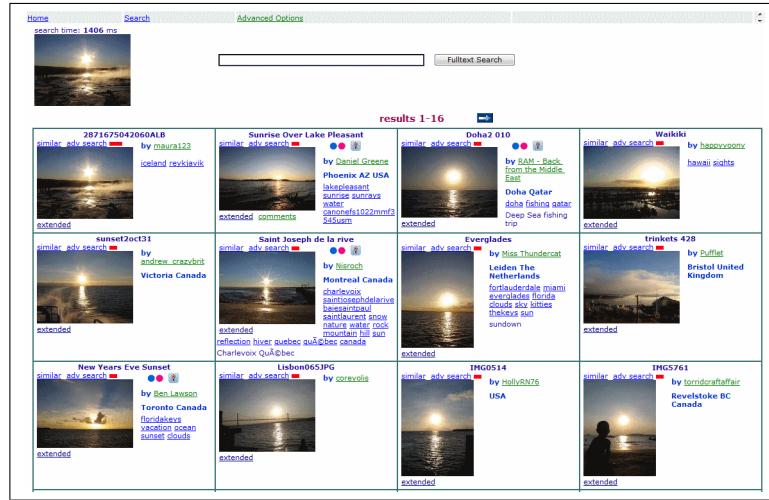
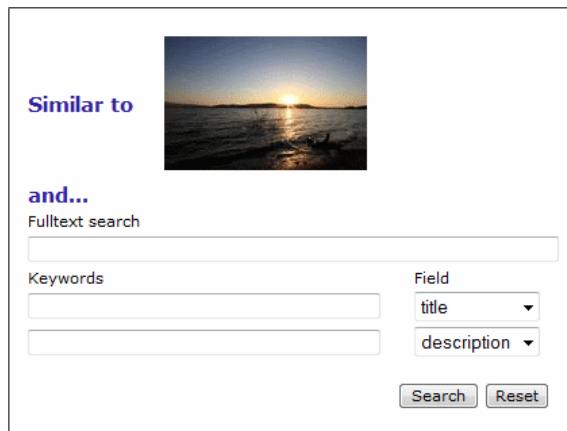


Fig. 2. Results page

window from which it is possible to start with a new query, lunch a new text query. For each result two text links are reported just over the image:

- *similar*: can be used to perform a similarity search with the given result as query. The similarity is evaluated comparing the five MPEG-7 visual descriptor used in CoPhIR. The weight of each descriptor has been fixed following the work reported in [1].
- *adv search*: can be used to access a pop-up window from which it is possible to perform a combined search using both the result as query for similarity and any given text combination as shown below:



For each result displayed the following information is reported:

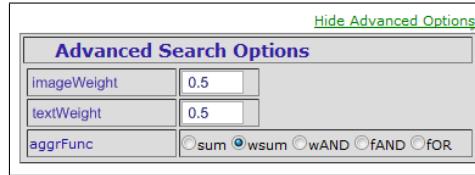
- the image title

- *score*: a red bar visually reports the score assigned to each result
-  and  buttons are used to link respectively to Flickr maps and Googlemaps whenever the geographic position during the take is present
- clicking on the result image itself it is possible to access the related Flickr page
- below these buttons we report the author's name
- the location name is reported
- the image tag
- comments can be found following the comments link
- the image description

Finally, at the bottom of the page there is button that can be used to see the next results in order of relevance to the query

The setting of the combined image and text search can be configured in the Advanced option screen. In particular it is possible to set:

1. imageWeight: the weight to give to the image
2. textWeight: the weight to give to the text
3. aggFunc: the aggregate function to be used (for details see Section 4.2)



Advanced Search Options	
imageWeight	0.5
textWeight	0.5
aggFunc	<input type="radio"/> sum <input checked="" type="radio"/> wsum <input type="radio"/> wAND <input type="radio"/> fAND <input type="radio"/> fOR

In the SAPIR demo homepage, a link is reported that can be used as a bookmarklet. Adding the bookmarklet to the browser bookmarks, it is possible to use any given image found on any web page as query. In Figure 3a we show the results obtained for a text search using Google Images. Clicking on the bookmarklet the images that are on the displayed webpage are reported in a separate page (see Figure 3b). Clicking on one of them, the selected one is used as query and then the results are displayed in Figure 3c.

6 Main Research Results and Future Work

Making this tool available to a large community of user will be important for two main reasons

- From the point of view of search engine technology, it will be the first time that a prototype system based on similarity search for multimedia data is actually used by many users concurrently on such a large image and text collection. We will collect information on the weakness and strength of the system under realistic load.

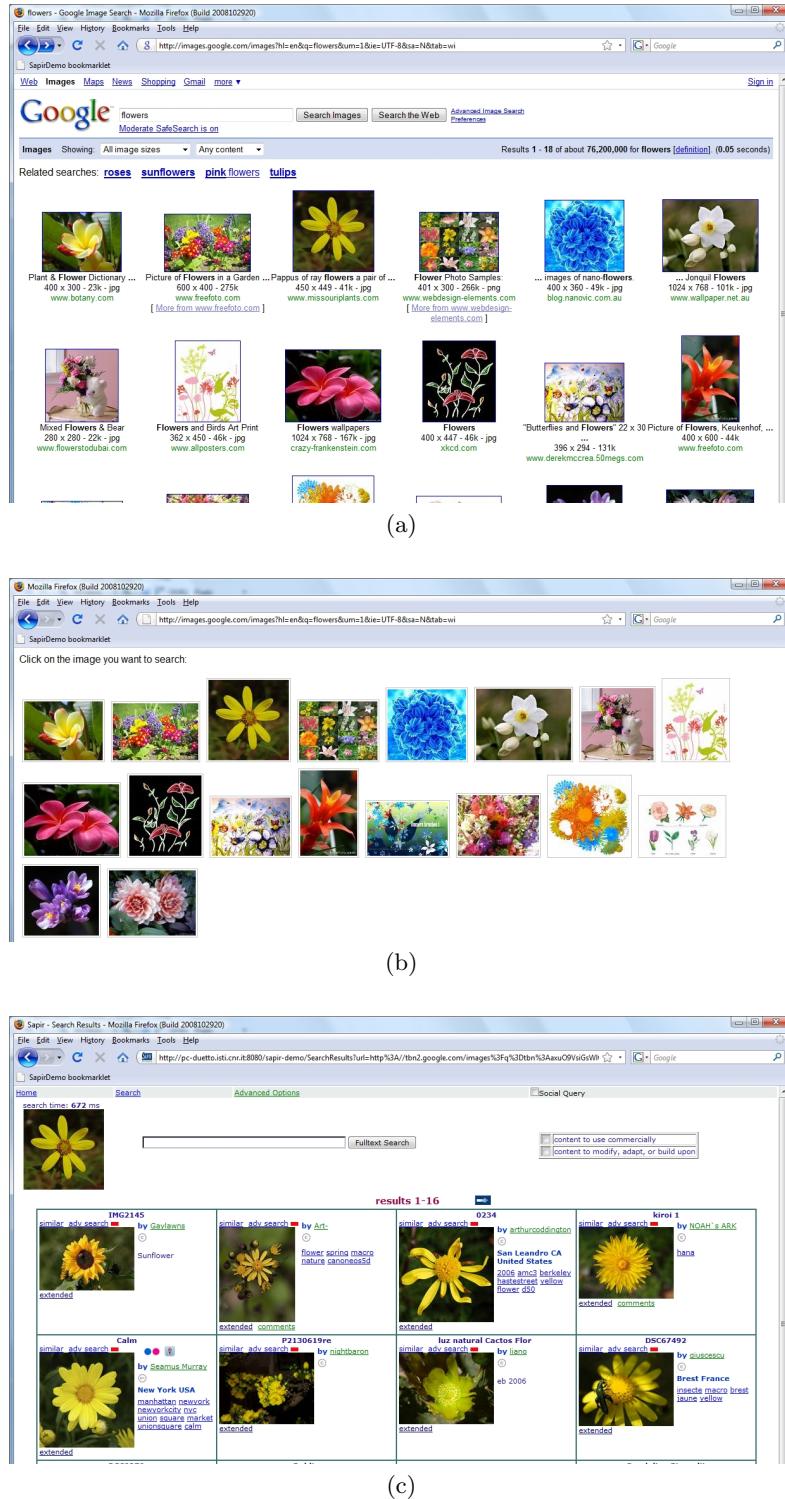


Fig. 3. Bookmarklet usage example

- From the point of view on user experience in searching, it will be the first time that a population of user will have the possibility to make their search using the search by content similarity paradigm (together the usual text search) on a large image collection with realistic response time. We will collect statistics on user behavior (access logs), such as the way they combine text and image content search. This is the first time such experience can be studied in a realistic environment.

Acknowledgments

This work has been partially supported by the SAPIR (Search In Audio Visual Content Using Peer-to-Peer IR) project, funded by the European Commission under IST FP6 (Sixth Framework Programme, Contract no. 45128).

The development and preparation of the demo has involved a large number of people from several SAPIR project partners. In particular we would like to mention: Benjamin Sznajder from IBM Haifa Research Lab; Paolo Bolettieri, Andrea Esuli, Claudio Gennaro, Matteo Mordacchini and Tommaso Piccioli from ISTI-CNR; Michal Batko, Vlastislav Dohnal, David Novak and Jan Sedmidubsky from Masaryk University; Mouna Kacimi and Tom Crecelius from Max-Planck-Institut für Informatik.

References

1. G. Amato, F. Falchi, C. Gennaro, F. Rabitti, P. Savino, and P. Stanchev. Improving image similarity search effectiveness in a multimedia content management system. In *MIS 2004 - 10th International Workshop on Multimedia Information System, College Park, MD, USA, August 25-27*, pages 139–146, 2004.
2. M. Batko, D. Novak, F. Falchi, and P. Zezula. On scalability of the similarity search in the world of peers. In *InfoScale '06: Proceedings of the 1st international conference on Scalable information systems*, page 20, New York, NY, USA, 2006. ACM Press.
3. M. Bender, S. Michel, P. Triantafillou, G. Weikum, and C. Zimmer. MINERVA: Collaborative P2P Search. In *VLDB '05: Proceedings of the 31st international conference on Very large data bases*, pages 1263–1266. VLDB Endowment, 2005.
4. T. Deselaers, T. Weyand, D. Keysers, W. Macherey, and H. Ney. Fire in imageclef 2005: Combining content-based image retrieval with textual information retrieval. In C. Peters, F. C. Gey, J. Gonzalo, H. Müller, G. J. F. Jones, M. Kluck, B. Magnini, and M. de Rijke, editors, *CLEF*, volume 4022 of *Lecture Notes in Computer Science*, pages 652–661. Springer, 2005.
5. R. Fagin, A. Lotem, and M. Naor. Optimal aggregation algorithms for middleware. In *PODS*, pages 102–113, New York, NY, USA, 2001. ACM Press.
6. R. Fagin, A. Lotem, and M. Naor. Optimal Aggregation Algorithms for Middleware. *CoRR*, cs.DB/0204046, 2002.
7. F. Falchi, C. Lucchese, S. Orlando, R. Perego, and F. Rabitti. Caching content-based queries for robust and efficient image retrieval. In *EDBT '09: Proceedings of the 12th International Conference on Extending Database Technology*, pages 780–790, New York, NY, USA, 2009. ACM.

8. F. Falchi, C. Lucchese, S. Orlando, R. Perego, and F. Rabitti. Caching content-based queries for robust and efficient image retrieval. In *EDBT 2009, 12th International Conference on Extending Database Technology, Saint-Petersburg, March 23-26 2009, Proceedings*, ACM International Conference Proceeding Series. ACM, 2009, forthcoming.
9. F. Jing, M. Li, H. Zhang, and B. Zhang. A unified framework for image retrieval using keyword and visual features. *IEEE Transactions on Image Processing*, 14(7):979–989, 2005.
10. J. Mamou, Y. Mass, M. Shmueli-Sheuer, and B. Sznajder. Query language for multimedia content. In *Procedding of the Multimedia Information Retrieval workshop held in conjunction with the 30 th Annual International ACM SIGIR Conference 27 July 2007, Amsterdam*, 2007.
11. A. Marian, N. Bruno, and L. Gravano. Evaluating top- queries over web-accessible databases. *ACM Trans. Database Syst.*, 29(2):319–362, 2004.
12. S. Michel, P. Triantafillou, and G. Weikum. Klee: A framework for distributed top-k query algorithms. In K. Böhm, C. S. Jensen, L. M. Haas, M. L. Kersten, P.-Å. Larson, and B. C. Ooi, editors, *VLDB*, pages 637–648. ACM, 2005.
13. M. Theobald, G. Weikum, and R. Schenkel. Top-k query evaluation with probabilistic guarantees. In *VLDB*, pages 648–659, 2004.
14. P. Zezula, G. Amato, V. Dohnal, and M. Batko. *Similarity Search. The Metric Space Approach*, volume 32 of *Advances in Database Systems*. Springer Science + Business Media, Inc., 233 Spring Street, New York, NY 10013, USA, 2006.

Design of an Information Retrieval System Based on the Peer-to-Peer Paradigm: An Application to Music Retrieval

Emanuele Di Buccio, Nicola Ferro, Massimo Melucci, Riccardo Miotto and
Nicola Orio

Department of Information Engineering, University of Padua, Italy
`{dibuccio,ferro,melo,miottori,orio}@dei.unipd.it`

Abstract. The Peer-To-Peer (P2P) paradigm is a good choice for providing federated search capabilities to collections of documents spread across Internet or to Digital Libraries. In traditional P2P networks, search operations focus on properly labeled files, and the search is often limited to textual metadata. The explosive growth of available multimedia documents in recent years called for more flexible search capabilities, namely search by content. In this paper we present a novel P2P architecture to provide a distributed content-based multimedia search engine where the search operations exploit some features related to the content of the documents rather than their metadata. The proposed system aims at retrieving music and text documents.

1 Introduction

Peer-To-Peer (P2P) networks integrate autonomous computing resources without requiring a central authority: the basic rationale is that peers are entities able to work as both client and server. P2P is a good choice to provide federated search capabilities to collections of documents spread across Internet or to Digital Libraries. In the beginning, the popularity of the P2P paradigm was due to the diffusion of applications for distributing and sharing digital documents, in particular music files. Among all the proposed systems in late 90s', Napster can be certainly considered the most famous one. In general, the weaknesses of this kind of systems concerned both the scalability and the search capabilities and during the last decade the problem of Information Retrieval (IR) across P2P networks was widely investigated. Although, a considerable part of the usage performed in P2P networks concerned music files sharing, the great part of the proposed solutions allowed to search music information by metadata only. In different scenarios, metadata could be either not suitable, or unreliable or even missing. Moreover, as it is well known, providing metadata for large collections is an extremely time consuming task and could also entail the problem of multilingual access to the documents.

The design of a system for Music Information Retrieval (MIR) across P2P networks should be performed both at modeling and architectural level. In [1] a

weighing framework for addressing the design of a P2P-IR system at modeling level was proposed. The following experimental evaluation reported in [2] showed that the scheme helps the retrieving of a significant proportion of relevant data after traversing only a small portion of a P2P hierarchical network in a depth-first manner. In [3], then, the same problem was addressed at an architectural level by proposing the Superimposed Peer Infrastructure for iNformation Access (SPINA) software architecture which allows to index and retrieve unstructured documents distributed across a P2P network.

MIR approaches [4] provide different methodologies of music processing and retrieval by exploiting some features related to the music content rather than the documents metadata. These approaches are very challenging and aimed at satisfying the need of the users which, for instance, could prefer to retrieve music documents by humming a little part of the melody or by submitting as query a fragment of an audio file.

In this paper we describe an ongoing capability of the SPINA architecture in order to automatically index and retrieve music documents by content. The automatic content-based retrieval of music documents is gaining increasing interest because it can provide new tools for music accessing and distribution. These can be exploited in several contexts such as recommendation systems, digital libraries population, Web searches, detection of copyright infringement and so on. In the last years, different music identification approaches have been proposed and, in particular, more recently in [5], [6] and [7]. In our work, in particular, the general ideas proposed in [7] has been extended to index and retrieve music files by content in a P2P network.

In the following sections, we describe the components of the system, ranging from the architecture designing to the music content indexing.

2 Design of a P2P-IR System

In the last decade the problem of text retrieval across P2P networks was widely investigated [8, 9]. On the contrary, there are only few solutions for the problem of P2P-MIR.

A Distributed Hash Table (DHT)-based system to retrieve music documents was proposed in [10]. The system exploits both manually specified tag-like information – e.g. artist, album, title – and automatic feature extraction techniques to search by content. Even though structured networks — e.g. DHT-based — enable for efficient query routing, they require an high degree of collaboration among peers, requirement which is not suitable for networks that are highly dynamic, heterogeneous, or protective of intellectual property. This kind of networks is well-matched by *unstructured* overlays. P2P-MIR systems for unstructured networks were proposed in [11] and in [12], but both approaches lack in terms of network traffic. A more efficient solution may be achieved by decreasing decentralization. The approach proposed in [13] exploits a centralized coordinator that stores the identifiers of all the peers in the network together with its own *PC feature*, which is a feature describing the music content of the peer, used to

select the most promising peers to answer to the formulated query. As suggested by the authors, the architecture might be improved in terms of efficiency and robustness by increasing the numbers of coordinators.

The type of overlays adopted by the software architecture described in this paper does not require the presence of fixed central entities. Indeed, some peers are *dynamically* elected as ultra-peers or super-peers. Because of the presence of ultra-peers this kind of overlays is named as hybrid. Moreover, the adopted overlay is a particular type of unstructured hybrid network because it is also hierarchical: indeed each peer refers to one and only one ultra-peer. The presence of ultra-peers, which act as hubs, enables to decrease the number of messages during query routing. The adoption of unstructured networks might be a suitable solution also because, as pointed out in [12], in such overlay topologies each peer can share just its own resources without keeping information about the resources of the other peers. This approach may be also exploited to localize peers that share illegal content or files which violate copyright.

The system designed aims at being independent not only from the underlying network infrastructure, but also from the media of the documents stored in the network. An API called SPINA has been designed and implemented [3] in order to achieve a flexible software architecture whose functionalities enable to search text and music documents by content. In the system design, each peer is provided with a local search engine that supplies all the functionalities required to perform the indexing and retrieval operations on the local collection of documents. Each peer locally stores two indexes, one for information about the text collection and the other one for the music documents.

Besides providing these functionalities, the system provides a retrieval process across the P2P network. The rest of the section describes the rationale of the search process and how to represent and select resources at higher levels starting from the statistics locally extracted by the peers about their content.

2.1 Search and Query Routing

Resource selection in P2P systems is related to the task of query routing because of the topology of the network. For this reason, in this section the adopted overlay topology and the query routing mechanism will be briefly described before focusing on the considered solution to the resource selection problem. Figure 1 depicts an instance of the logical layers that SPINA superimposes on top of the underlying network infrastructure. In the figure three layers are considered: starting from the lower one, we can distinguish the document, the peer and the ultra-peer layer. Another characteristic shown in Figure 1 is that, because of the hierarchical nature of the underlying overlay, the network is subdivided in groups – no clustering is adopted – each of which refers to an ultra-peer. For instance, p_a , p_b and p_c belong to the group of peers that refers to the ultra-peer up_d . The notion of **group** is crucial to explain the first part of the retrieval process.

When interacting with the peer, the end user can perform a music or textual search by submitting a query as a bag of features. At present time SPINA supports free text search queries and the possibility to submit a MIDI file in order

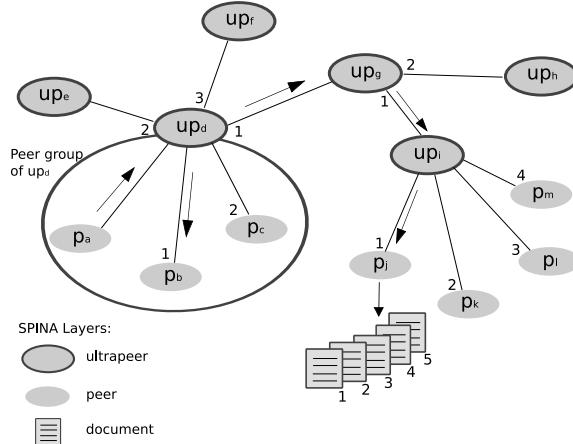


Fig. 1. P2P topology adopted in SPINA and query routing mechanism.

to identify it or to find similar songs. If the query is formulated in a peer, the request is forwarded to the ultra-peer which the peer refers to. But the request can be formulated also in an ultra-peer – remember that also an ultra-peer is a peer and so it is provided by a local engine. Whatever the starting resource is, the resources at higher levels – peers and ultra-peers – are ranked and selected by ultra-peers according to a certain criterion. Then the selected peers are contacted according to the order they appear in the ranked peer list. After ranking its local resources each peer returns to the referring ultra-peer a ranked list of objects in response to the formulated query. For instance, in Figure 1, the query is formulated in peer p_a , then it is forwarded to the referring ultra-peer up_d , and finally to the other peers of the group, that is p_b and p_c . The order in which the peers are contacted, that is first p_b and then p_c , is denoted by the number nearby the edge connecting the peer with its ultra-peer.

Retrieval is not limited to the group a peer belongs to. The ultra-peers communicate among them to form the backbone of an hybrid decentralized network. Before explaining how the search process continues, the notion of **neighbor** has to be clarified. Each ultra-peer stores some information about a subset of the other ultra-peers in the network: this subset of ultra-peers constitutes the set of neighbors of the ultra-peer. According to the information about its neighbors, an ultra-peer can extend the search to the peers of the other groups. In particular, the ultra-peer ranks its neighboring ultra-peers and forwards the query to the most promising neighbors which in their turn search in the groups that serve. For instance, in Figure 1, the ultra-peer up_d contacts first up_g , then up_e and finally up_f . The order in which the ultra-peers are contacted is denoted by the number nearby the edge connecting the ultra-peer which forwards the query and its neighbors. The search process can be extended by propagating the query to

the neighbors of the latest contacted ultra-peers, for a number of hops bounded by the Time-To-Live (TTL) of the formulated query.

Each ultra-peer is able to merge the results obtained by the peers of the group that serves together with the merged list of results returned by its neighbors. The final merged result list is then returned to the requesting peer by following the inverse path of the query.

2.2 Ranking Criterion and High Granularity Indexes

In Section 2.1 we referred to a ranking criterion to rank resources at the different levels of the hierarchy. In particular, an ultra-peer has to rank the peers in its group in order that the query is forwarded to the most promising peers. At the higher level, ultra-peers have to be ranked so that the query can be forwarded to the most promising neighbors of the ultra-peer. Indeed, the resource selection problem in the considered network topology seems to be characterized by a recursive nature exploited in the design of the system, but even before in the weighing framework proposed in [1], that is the Term Weighted Frequency (TWF) Inverse Resource Frequency (IRF) weighing scheme. In this scheme the weight assigned to a resource — peer or ultra-peer — is computed by the statistics about the features extracted from the full content of the documents in the peer’s collections, aggregated according to the level hierarchy. The rationale of this choice is to use the features that occurs in the collection stored in the peer, or in the set of collections stored in a group of peers, and the statistics derived from their frequency of occurrence, to describe respectively a peer or an ultra-peer. The methodology described in Section 3 allows to use the same description of resources at higher levels also in terms of musical features. Since only the exchange of little information – features with their weights – is required to rank resources at different levels, the adopted weighing scheme is a suitable solution in terms of network load.

Each ultra-peer locally stores two indexes, one for each resource level, where the information required by the weighing scheme is stored.

The first index has a “peer-granularity” and it is achieved by aggregating the information about the content of the peers in the group the ultra-peer serves. The aggregation takes place in each single peer that communicates this summary information. The ultra-peer collects and stores this information in a local index, which basically provides the list of peers in the ultra-peer group associated to each feature, as well as the total weight of every feature occurring in every peer. According to this information the peers in the group are ranked by the adopted weighing scheme and then the query is routed to the most promising peers.

The information required to rank and select the neighbors is stored in the second typology of index, that is the “ultra-peer-granularity” index. In this index the list of neighboring ultra-peers is associated to each feature, as well as the total weight of every feature occurring in every neighboring ultra-peer. Each ultra-peer computes its contribution locally by aggregating the information about the peers in its group thus achieving a list of all the features in the peer-granularity

index and the weight such feature has in the ultra-peer itself. Thus each ultra-peer has two indexes – peer and ultra-peer granularity – for each medium – text and audio.

3 Music Content Representation

A general way to represent and index music content is required to integrate a content-based music search component in the proposed architecture. In an IR system, the document indexing is a crucial step because it enables a compact representation of the content of a collection, aimed at an efficient and scalable access and retrieval. General indexing techniques can be extended also to music, providing that significant descriptors are computed from music documents. These descriptors can be defined as the *lexical units* of music, and depend on the dimension that are taken into account – melody, harmony, rhythm, timbre – and are related to the way listeners perceive music. With the aim of creating a common retrieval strategy for different media, the index for music documents has to be consistent with the index for all the other media, following a similar scheme. The basic idea underlying the approach is that a music document can be effectively described by excerpts of its melodic features [14]. The main goal then becomes the automatic extraction of relevant excerpts from an unstructured flow of notes, stored in different digital formats. With this aim we can introduce some general terminology. In particular, according to [7], we can define as **features** the characteristic that describes subsequent notes in a music documents. Features can be of different types, such as the pitch, the pitch interval with the previous note, the duration, the chroma features [15] and so on. In our approach they are mostly related to pitch and rhythm which can also be treated independently and can be considered valid descriptors for our purposes. A sequence of features is defined as **string**. For example, any sub-sequence of notes in a melody can be considered as a string. A string repeated at least twice in a music flow is defined as a **pattern**. The repetition can be due to the presence of different choruses in the flow or by the use of the same melodic material. Patterns can be considered as the descriptor of the music documents and can carry different information about their content. The structure of the patterns can be related to the textual case where patterns play the same role of words of a document. Thus the most common indexing textual techniques can be exploited.

Following the terminology introduced previously, each document has a number of patterns of different length and with different multiplicity. Among all the computed patterns, some of them could have little or no musical meaning. For instance, a pattern that is repeated only two or three times in a document is likely to be computed by chance just because the combination of features is repeated in some notes combination. Moreover, some patterns related are likely to appear in almost all documents and hence to be poor discriminant among documents. In general, the degree by which a pattern is a good index may vary depending on the pattern and on the document. This is a typical scenario of textual information retrieval where words may describe a document to a differ-

ent extent. For this reason, a weighing scheme based on the *TF-IDF* measure might be a viable solution for MIR too as reported in [14]. Thus, at the end of the process, the music documents will be described by a sequence of patterns together with their weight (which represents the number of times a pattern is repeated in the document) and their positions (in *ms*) along the music flow. In the designed system each peer stores this information in its local audio index.

As stated previously, each pattern is composed by a sequence of features which describes the music content. Features computation is a difficult task which strictly depends on the digital format in which music documents are stored. Indeed, according to [16], some digital formats can represent music scores, whereas others can represent music performances. At the current state of the art, the most common formats are MIDI for music scores and both WAVE and MP3 for music performances.

According to the digital formats, thus, the algorithms of content extraction, even if with the same objective, might be completely different.

4 Conclusions

In this paper the current status of the design of a P2P content-based search engine is reported. The architecture has been thought to be flexible enough to handle different types of media, both in terms of representation and retrieval.

At the present time, the network infrastructure liable to the communication among peers has been implemented together with the functionalities for the search engine. A weighing scheme to rank resources at different level hierarchy and the indexes which stored the information used to compute the weights have been already developed. Each peer of the network stores music and textual documents which can be retrieved by the engine. The implemented functionalities enable to perform content-based retrieval in a P2P networks, where a generic user can submit as query both a sequence of words and a musical file.

Some questions are still open and under investigation. Concerning the P2P architecture, the major issues pertain to the dynamism of both the peers and the statistics. A first point concerns the set of the policies which handle the composition of the peers groups. It is important to define some rules in order to manage the behavior of the network, especially when an ultra-peer is shut down and when a new peer joins the network and has to be associated with an ultra-peer. Other studies will be aimed at the definition of some policies to manage the dynamics due to the change of the documents stored in a peer.

Concerning the documents representation, the major issue is aimed at handling the music recordings, in particular stored in the MP3 format. The music content representation, both of scores and recordings, then has to be deeply tested in order to formally evaluate the retrieval efficiency of the system. In particular, some standard music collections should be exploited in order to make the achieved results comparable with other systems.

Finally, a valid strategy which enables an ultra-peer to merge the results returned by the peers during a search process will be investigated.

Acknowledgments

The authors are grateful to Maristella Agosti and Giorgio Maria Di Nunzio for the fruitful discussions on the topic of this paper. The work reported in this paper has been partially supported by the SAPIR project, as a part of the Information Society Technologies (IST) Program of the European Commission (Contract IST-045128).

References

1. Castiglion, R., Melucci, M.: An evaluation of a recursive weighing scheme for information retrieval in peer-to-peer networks. In: Proceedings of P2PIR 2005, Bremen, Germany (2005) 9–16
2. Melucci, M., Poggiani, A.: A study of a weighting scheme for information retrieval in hierarchical peer-to-peer networks. In: Proceedings of ECIR 2007, Rome, Italy (2007) 136–147
3. Di Buccio, E., Ferro, N., Melucci, M.: Content-based Information Retrieval in SPINA. In: Proceedings of IRCDL 2008, Padua, Italy (2008) 89–92
4. Orio, N.: Music retrieval: a tutorial and review. *Foundations and Trends in Information Retrieval* **1**(1) (2006) 1–96
5. Cano, P., Batlle, E., Kalker, T., Haitsma, J.: A Review of Audio Fingerprinting. *Journal of VLSI Signal Processing Systems* **41**(3) (2005) 271–284
6. Miotto, R., Orio, N.: A methodology for the segmentation and identification of music works. In: Proceedings of ISMIR 2007, Vienna, Austria (2007) 271–284
7. Neve, G., Orio, N.: Indexing and retrieval of music documents through pattern analysis and data fusion techniques. In: Proceedings of ISMIR 2004, Barcelona, Spain (2004) 216–223
8. Lu, J., Callan, J.: Full-text federated search of text-based digital libraries in peer-to-peer networks. *Information Retrieval* **9**(4) (2006) 477–498
9. Nottelmann, H., Fuhr, N.: Comparing Different Architectures for Query Routing in Peer-to-Peer Networks. In: Proceedings of the ECIR 2006, London, UK (2006) 253–264
10. Tzanetakis, G., Gao, J., Steenkiste, P.: A scalable peer-to-peer system for music information retrieval. *Computer Music Journal* **28**(2) (2004) 24–33
11. Yang, C.: Peer-to-peer architecture for content-based music retrieval on acoustic data. In: Proceedings of WWW2003, Budapest, Hungary (2003) 376–383
12. Karydis, I., Nanopoulos, A., Papadopoulos, A.N., Manolopoulos, Y.: Musical retrieval in p2p networks under the warping distance. In: Proceedings of ICEIS 2005, Miami, USA (2005) 100–107
13. Wang, C., Li, J., Shi, S.: A Kind of Content-Based Music Information Retrieval Method in Peer-to-peer Environment. In: Proceedings of ISMIR 2002, Paris, France (2002) 178–186
14. Melucci, M., Orio, N.: Musical information retrieval using melodic surface. In: Proceedings of DL'99, Berkeley, California, United States (1999) 152–160
15. Peeters, G.: Chroma-based estimation of musical key from audio-signal analysis. In: Proceedings of ISMIR 2007, Victoria, Canada (2006) 115–120
16. Orio, N.: Alignment of performances with scores aimed at content-based music access and retrieval. In: Proceedings of ECDL 2002, Rome, Italy (2002) 479–492

A Hybrid Strategy for Italian Word Sense Disambiguation

Pierpaolo Basile¹, Marco de Gemmis¹, Pasquale Lops¹ and Giovanni Semeraro¹

Department of Computer Science
University of Bari “Aldo Moro”
Via E. Orabona, 4 - 70126 Bari, Italy
`{basilepp,degemmis,lops,semearo}@di.uniba.it`

Abstract. Word Sense Disambiguation (WSD) is the problem of selecting a sense for a word from a set of predefined possibilities. Currently, there are two main methodological streams in this area: knowledge-based and corpus-based methods. Knowledge-based methods use external knowledge sources (e.g. *Machine Readable Dictionaries* or *Thesauri*) for assigning the correct sense to a word in a text. Corpus-based methods exploit machine learning techniques to induce models of word usage from large text collections.

This paper presents a WSD strategy which combines a knowledge-based method that exploits sense definitions in a dictionary and relations among senses in a semantic network, with supervised learning methods on annotated corpora. The idea behind the proposed approach is that the knowledge-based method can cope with the possible lack of training data, while supervised learning can improve the precision of a knowledge-based method when training data are available.

In order to evaluate the effectiveness of the proposed approach, experimental sessions were carried out on the dataset used for the WSD task in the EVALITA 2007 initiative, devoted to the evaluation of Natural Language Processing tools for Italian.

The general conclusion of the experimental session is that the most effective hybrid WSD strategy is the one that integrates the knowledge-based approach into the supervised learning method, which outperforms both methods taken singularly.

1 Background and Motivations

The inherent ambiguity of human language is a greatly debated problem in many research areas, such as Information Retrieval and Text Categorization, since the presence of polysemous words often causes a wrong relevance judgment or classification of documents. These problems call for alternative methods that work not only at the lexical level of the documents, but also at the *meaning* level. The task of Word Sense Disambiguation (WSD) consists in assigning the most appropriate meaning to a polysemous word within a given context. Applications such as machine translation, knowledge acquisition, common sense reasoning

and others, require knowledge about word meanings, and WSD is essential for all these applications.

The assignment of senses to words is accomplished by using two major sources of information [19]:

1. the *context* of the word to be disambiguated, e.g. information contained within the text in which the word appears;
2. *external knowledge sources*, including lexical resources, as well as hand-devised knowledge sources, which provide data useful to associate words with senses.

All disambiguation work involves matching the context of the instance of the word to be disambiguated with either information from an external knowledge source (*knowledge-driven* WSD), or information about the contexts of previously disambiguated instances of the word derived from corpora (*data-driven* or *corpus-based* WSD).

Corpus-based WSD exploits semantically annotated corpora to train machine learning algorithms to decide which word sense to choose in which context. Words in such annotated corpora are tagged manually using semantic classes chosen from a particular lexical semantic resource (e.g. WORDNET [9]). Each sense-tagged occurrence of a particular word is transformed into a feature vector, which is then used in an automatic learning process. The applicability of such supervised algorithms is limited to those few words for which sense tagged data are available, and their accuracy is strongly influenced by the amount of labeled data available.

Knowledge-based WSD has the advantage of avoiding the need of sense-annotated data, rather it exploits lexical knowledge stored in machine-readable dictionaries or thesauri. Systems adopting this approach have proved to be ready-to-use and scalable, but in general they reach lower precision than corpus-based WSD systems.

Our hypothesis is that the combination of both types of strategies can improve WSD effectiveness, because knowledge-based methods can cope with the possible lack of training data, while supervised learning can improve the precision of knowledge-based methods when training data are available.

This paper presents a method for solving the semantic ambiguity of *all words* contained in a text¹. We propose a hybrid WSD algorithm that combines a knowledge-based WSD algorithm, called JIGSAW [3], which we designed to work by exploiting WORDNET-like dictionaries as sense repository, with a supervised machine learning algorithm (K-Nearest Neighbor classifier). WORDNET-like dictionaries are used because they combine the characteristics of both a dictionary and a structured semantic network, supplying definitions for the different senses of words and defining groups of synonymous words by means of *synsets*, which represent distinct lexical concepts. WORDNET also organizes synsets in a conceptual structure by defining a number of semantic relationship (IS-A, PART-OF, etc.) among them.

¹ *all words* task tries to disambiguate all the words in a text, while *lexical sample* task tries to disambiguate only specific words

Mainly, the paper concentrates on two investigations:

1. First, corpus-based WSD is applied to words for which training examples are provided, then JIGSAW is applied to words not covered in the first step, with the advantage of knowing the senses of the context words already disambiguated in the first step;
2. First, JIGSAW is applied to assign the most appropriate sense to those words that can be disambiguated with a high level of confidence (by setting a specific parameter in the algorithm), then the remaining words are disambiguated by the corpus-based method.

The paper is organized as follows: After a brief discussion about the main works related to our research, Section 3 gives the main ideas underlying the proposed hybrid WSD strategy. More details about the K-NN classification algorithm and JIGSAW, on which the hybrid WSD approach is based, are provided in Section 4 and Section 5, respectively. Experimental sessions have been carried out in order to evaluate the proposed approach in the critical situation when training data are not much reliable, as for Italian. Section 6 presents new results with respect to those already presented in [4], obtained by solving some problems in the part of speech tagging and lemmatization procedures. Conclusions and future work close the paper.

2 Related Work

For some Natural Language Processing (NLP) tasks, such as part of speech tagging or named entity recognition, there is a consensus on what makes a successful algorithm, regardless of the approach considered. Instead, no such consensus has been reached yet for the task of WSD, and previous work has considered a range of knowledge sources, such as local collocational clues, common membership in semantically or topically related word classes, semantic density, and others. In recent SENSEVAL-3 evaluations², the most successful approaches for *all words* WSD relied on information drawn from annotated corpora. The system developed by Decadt [7] uses two cascaded memory-based classifiers, combined with the use of a genetic algorithm for joint parameter optimization and feature selection. A separate word expert is learned for each ambiguous word, using a concatenated corpus of English sense tagged texts, including SemCor, SENSEVAL datasets, and a corpus built from WORDNET examples. The performance of this system on the SENSEVAL-3 English all words dataset was evaluated at 65.2%. Another top ranked system is the one developed by Yuret [23], which combines two Naïve Bayes statistical models, one based on surrounding collocations and another one based on a bag of words around the target word. The statistical models are built based on SemCor and WORDNET, for an overall disambiguation accuracy of 64.1%. All previous systems use supervised methods, thus requiring a large amount of human intervention to annotate the training

² <http://www.senseval.org>.

data. In the context of the current multilingual society, this strong requirement is even increased, since the so-called “sense-tagged data bottleneck problem” is emphasized.

To address this problem, different methods have been proposed. This includes the automatic generation of sense-tagged data using monosemous relatives [12], automatically bootstrapped disambiguation patterns [14], parallel texts as a way to point out word senses bearing different translations in a second language [8], and the use of volunteer contributions over the Web [17]. More recently, Wikipedia has been used as a source of sense annotations for building a sense annotated corpus which can be used to train accurate sense classifiers [16]. Even though the Wikipedia-based sense annotations were found reliable, leading to accurate sense classifiers, one of the limitations of the approach is that definitions and annotations in Wikipedia are available almost exclusively for nouns.

On the other hand, the increasing availability of large-scale rich (lexical) knowledge resources seems to provide new challenges to knowledge-based approaches [20, 15]. Our hypothesis is that the complementarity of knowledge-based methods and corpus-based ones is the key to improve WSD effectiveness. The aim of the paper is to define a cascade hybrid method able to exploit both linguistic information coming from WORDNET-like dictionaries and statistical information coming from sense-annotated corpora.

3 A Hybrid Strategy for WSD

The goal of WSD algorithms consists in assigning a word w_i occurring in a document d with its appropriate meaning or sense s . The sense s is selected from a predefined set of possibilities, usually known as *sense inventory*. We adopt ITALWORDNET [21] as sense repository. The algorithm is composed by two procedures:

1. **JIGSAW** - It is a knowledge-based WSD algorithm based on the assumption that the adoption of different strategies depending on Part-of-Speech (PoS) is better than using always the same strategy. A brief description of JIGSAW is given in Section 5, more details are reported in [3, 5, 22].
2. **Supervised learning procedure** - A K-NN classifier [18], trained on MultiSemCor corpus³[6] is adopted. Details are given in Section 4. MultiSemCor is an English/Italian parallel corpus, aligned at the word level and annotated with PoS, lemma and word senses. The parallel corpus is created by exploiting the SemCor corpus⁴, which is a subset of the English Brown corpus containing about 700,000 running words. In SemCor, all the words are tagged by PoS, and more than 200,000 content words are also lemmatized and sense-tagged with reference to the WORDNET lexical database. SemCor has been used in several supervised WSD algorithms for English with good

³ <http://multisemcor.itc.it/>

⁴ <http://www.cs.unt.edu/~rada/downloads.html#semcor>

results. MultiSemCor contains less annotations than SemCor, thus the accuracy and the coverage of the supervised learning for Italian might be affected by poor training data.

The idea is to combine both procedures in a hybrid WSD approach. A first choice might be the adoption of the supervised method as first attempt, then JIGSAW could be applied to words not covered in the first step. Differently, JIGSAW might be applied first, then leaving the supervised approach to disambiguate the remaining words. An investigation is required in order to choose the most effective combination.

4 A Supervised Learning Method

The goal of supervised methods is to use a set of annotated data as little as possible, and at the same time to make the algorithm general enough to be able to disambiguate all content words in a text. We use MultiSemCor as annotated corpus, since at present it is the only available semantic annotated resource for Italian. The algorithm starts with a preprocessing stage, where the text is tokenized, stemmed, lemmatized and annotated with PoS.

Also, the collocations are identified using a sliding window approach, where a collocation is considered to be a sequence of words that forms a compound concept defined in ITALWORDNET (e.g. artificial intelligence). In the training step, a semantic model is learned for each PoS, starting with the annotated corpus. These models are then used to disambiguate words in the test corpus by annotating them with their corresponding meaning. The models can only handle words that were previously seen in the training corpus, and therefore their coverage is not 100%. Starting with an annotated corpus formed by all annotated files in MultiSemCor, a separate training dataset is built for each PoS. For each open-class word in the training corpus, a feature vector is built and added to the corresponding training set. The following features are used to describe an occurrence of a word in the training corpus as in [10]:

- **Nouns** - Two features are included in feature vector: the first noun, verb, or adjective before the target noun, within a window of at most three words to the left, and its PoS;
- **Verbs** - Four features are included in feature vector: the first word before and the first word after the target verb, and their PoS;
- **Adjectives** - all the nouns occurring in two windows, each one of six words (before and after the target adjective) are included in the feature vector;
- **Adverbs** - the same as for adjectives, but vectors contain adjectives rather than nouns.

The label of each feature vector consists of the target word and the corresponding sense, represented as *word#sense*. Table 1 describes the number of vectors for each PoS.

PoS	#feature vectors
Noun	38,546
Verb	18,688
Adjective	6,253
Adverb	1,576

Table 1. Number of feature vectors.

To annotate (disambiguate) new text, similar vectors are built for all content-words in the text to be analyzed. Consider the target word *bank*, used as a noun. The algorithm catches all the feature vectors of *bank* as a noun from the training model, and builds the feature vector v_f for the target word. Then, the algorithm computes the similarity between each training vector and v_f and ranks the training vectors in decreasing order according to the similarity value.

The similarity is computed as Euclidean distance between vectors, where PoS distance is set to 1, if PoS tags are different, otherwise it is set to 0. Word distances are computed by using the *Levenshtein* metric, that measures the amount of difference between two strings as the minimum number of operations needed to transform one string into the other, where an operation is an insertion, deletion, or substitution of a single character [13]. Finally, the target word is labeled with the most frequent sense in the first K vectors.

5 JIGSAW - A Knowledge-based Approach

JIGSAW is a WSD algorithm based on the idea of combining three different strategies to disambiguate nouns, verbs and adjectives/adverbs. The main motivation behind our approach is that the effectiveness of a WSD algorithm is strongly influenced by the PoS tag of the target word.

JIGSAW takes as input a document $d = (w_1, w_2, \dots, w_h)$ and returns a list of synsets $X = (s_1, s_2, \dots, s_k)$ in which each element s_i is obtained by disambiguating the *target word* w_i based on the information obtained from the sense repository about a few immediately surrounding words. We define the *context* C of the target word to be a window of n words to the left and another n words to the right, for a total of $2n$ surrounding words. The algorithm is based on three different procedures for nouns, verbs, adverbs and adjectives, called $JIGSAW_{nouns}$, $JIGSAW_{verbs}$, $JIGSAW_{others}$, respectively.

JIGSAW_{nouns} - Given a set of nouns $W = \{w_1, w_2, \dots, w_n\}$, obtained from document d , with each w_i having an associated sense inventory $S_i = \{s_{i1}, s_{i2}, \dots, s_{ik}\}$ of possible senses, the goal is assigning each w_i with the most appropriate sense $s_{ih} \in S_i$, according to the *similarity* of w_i with the other words in W (the context for w_i). The idea is to define a function $\varphi(w_i, s_{ij})$, $w_i \in W$, $s_{ij} \in S_i$, that computes a value in $[0, 1]$ representing the confidence with which word w_i can be assigned with sense s_{ij} . In order to measure the relatedness of two words we adopted a modified version of the Leacock-Chodorow [11] measure, which computes the length of the path between two concepts in a hierarchy by

passing through their *Most Specific Subsumer* (MSS). We introduced a constant factor *depth* which limits the search for the MSS to *depth* ancestors, in order to avoid “poorly informative” MSSs. Moreover, in the similarity computation, we introduced both a Gaussian factor $G(pos(w_i), pos(w_j))$, which takes into account the distance between the position of the words in the text to be disambiguated, and a factor $R(k)$, which assigns s_{ik} with a numerical value, according to the frequency score in ITALWORDNET.

JIGSAW_{verbs} - Before describing the *JIGSAW_{verbs}* procedure, the *description* of a synset must be defined. We define the *description* of a synset as the string obtained by concatenating the gloss and the sentences that ITALWORDNET uses to explain the usage of a synset. *JIGSAW_{verbs}* includes, in the context C for the target verb w_i , all the nouns in the window of $2n$ words surrounding w_i . For each candidate synset s_{ik} of w_i , the algorithm computes $nouns(i, k)$, that is the set of nouns in the description for s_{ik} . Then, for each w_j in C and each synset s_{ik} , the following value is computed:

$$max_{jk} = max_{w_l \in nouns(i, k)} \{ \text{sim}(w_j, w_l, depth) \} \quad (1)$$

where $\text{sim}(w_j, w_l, depth)$ is the same similarity measure adopted by *JIGSAW_{nouns}*. Finally, an overall similarity score among s_{ik} and the whole context C is computed:

$$\varphi(i, k) = R(k) \cdot \frac{\sum_{w_j \in C} G(pos(w_i), pos(w_j)) \cdot max_{jk}}{\sum_h G(pos(w_i), pos(w_h))} \quad (2)$$

where both $R(k)$ and $G(pos(w_i), pos(w_j))$, that gives a higher weight to words closer to the target word, are defined as in *JIGSAW_{nouns}*. The synset assigned to w_i is the one with the highest φ value.

JIGSAW_{others} - This procedure is based on the WSD algorithm proposed in [2]. The idea is to compare the glosses of each candidate sense for the target word to the glosses of all the words in its context.

6 Experiments

The main goal of our investigation is to study the behavior of the hybrid algorithm when available training resources are not much reliable, e.g. when a lower number of sense descriptions is available, as for Italian. The hypothesis we want to evaluate is that corpus-based methods and knowledge-based ones can be combined to improve the accuracy of each single strategy.

Experiments have been performed on a standard test collection in the context of the *All-Words-Task*, in which WSD algorithms attempt to disambiguate all words in a text. Specifically, we used the EVALITA WSD All-Words-Task dataset⁵, which consists of about 5,000 words labeled with ITALWORDNET synsets. An important concern for the evaluation of WSD systems is the agreement rate between human annotators on word sense assignment.

⁵ <http://evalita.fbk.eu/2007/tasks/wsd.html>

While for natural language subtasks like part of speech tagging, there are relatively well defined and agreed-upon criteria of what it means to have the “correct” part of speech assigned to a word, this is not the case for word sense assignment. Two human annotators may genuinely disagree on their sense assignment to a word in a context, since the distinction between the different senses for a commonly used word in a dictionary like WORDNET tend to be rather fine.

What we would like to underline here is that it is important that human agreement on an annotated corpus is carefully measured, in order to set an upper bound to the performance measures: it would be futile to expect computers to agree more with the reference corpus than human annotators among them. For example, the inter-annotator agreement rate during the preparation of the SENSEVAL-3 WSD English All-Words-Task dataset [1] was approximately 72.5%. Unfortunately, for EVALITA dataset, the inter-annotator agreement has not been measured, one of the reasons why the evaluation for Italian WSD is very hard. In our experiments, we reasonably selected different baselines to compare the performance of the proposed hybrid algorithm.

6.1 Integrating JIGSAW into a supervised learning method

The design of the experiment is as follows: firstly, corpus-based WSD is applied to words for which training examples are provided, then JIGSAW is applied to words not covered by the first step, with the advantage of knowing the senses of the context words already disambiguated in the first step. The performance of the hybrid method was measured in terms of precision (P), recall (R), F-measure (F) and the percentage A of disambiguation attempts, computed by counting the words for which a disambiguation attempt is made (the words with no training examples or sense definitions cannot be disambiguated). Table 2 shows the baselines chosen to compare the hybrid WSD algorithm on the All-Words-Task experiments.

The simplest baseline consists in assigning a random sense to each word (*Random*), another common baseline in Word Sense Disambiguation is first sense (*1st sense*): each word is tagged using the first sense in ITALWORDNET that is the most commonly (frequent) used sense. The other baselines are the two methods combined in the hybrid WSD, taken separately, namely JIGSAW and K-NN, and the basic hybrid algorithm “K-NN + 1st sense”, which applies the supervised method, and then adopts the first sense heuristic for the words without examples into training data. The K-NN baseline achieves the highest precision, but the lowest recall due to the low coverage in the training data (42.2%), and makes this method useless for all practical purposes. Notice that JIGSAW was the only participant to EVALITA WSD All-Words-Task, therefore it currently represents the only available system performing WSD All-Words task for the Italian language.

Table 3 reports the results obtained by the hybrid method on the EVALITA dataset. We study the behavior of the hybrid approach with respect to that of JIGSAW, since this specific experiment aims at evaluating the potential improvement due to the inclusion of JIGSAW into K-NN. Different runs of the

Setting	P	R	F	A
1 st sense	0.648	0.614	0.631	94.7
Random	0.483	0.458	0.470	94.7
JIGSAW	0.639	0.606	0.622	94.7
K-NN	0.797	0.336	0.473	42.2
K-NN + 1 st sense	0.640	0.606	0.623	97.7

Table 2. Baselines for Italian All-Words-Task

Setting	P	R	F	A
K-NN + JIGSAW	0.624	0.591	0.607	94.7
K-NN + JIGSAW ($\varphi \geq 0.80$)	0.693	0.337	0.453	48.6
K-NN + JIGSAW ($\varphi \geq 0.60$)	0.680	0.410	0.512	60.3
K-NN + JIGSAW ($\varphi \geq 0.40$)	0.652	0.452	0.534	69.3
K-NN + JIGSAW ($\varphi \geq 0.20$)	0.652	0.452	0.534	69.3

Table 3. Experimental results of K-NN+JIGSAW

hybrid method have been performed, each run corresponding to setting a specific value for φ (the confidence with which a word w_i is correctly disambiguated by JIGSAW). In each different run, the disambiguation carried out by JIGSAW is considered reliable only when φ values exceed a certain threshold, otherwise any sense is assigned to the target word (this is the reason why A decreases by setting higher values for φ). A positive effect on precision can be noticed by varying φ between 0.20 and 0.80. It tends to grow and overcomes all the baselines, but a corresponding decrease of recall is observed, as a consequence of more severe constraints set on φ .

Better results are achieved when no restriction is set on φ (K-NN+JIGSAW in Table 3): the recall is significantly higher than that obtained in the other runs. On the other hand, the precision reached in this run is lower than in the others, but it is still acceptable.

To sum up, two main conclusions can be drawn from the experiments:

- when no constraint is set on the knowledge-based method, the hybrid algorithm K-NN+JIGSAW in general has the same performance of the knowledge-based approach JIGSAW;
- when thresholding is introduced on φ , improvement is observed with respect to K-NN+JIGSAW in terms of precision.

A deeper analysis of results revealed that lower recall was achieved for verbs and adjectives rather than for nouns. Indeed, disambiguation of Italian verbs and adjectives is very hard, but the lower recall is probability due also to the fact that *JIGSAW* uses glosses for verbs and adjectives disambiguation. As a consequence, the performance depends on the accuracy of word descriptions in the glosses, while for nouns the algorithm relies only the semantic relations between synsets.

6.2 Integrating supervised learning into JIGSAW

In this experiment we test whether the supervised algorithm can help JIGSAW to disambiguate more accurately. The experiment has been organized as follows: JIGSAW is applied to assign the most appropriate sense to the words which can be disambiguated with a high level of confidence (by setting the φ threshold), then the remaining words are disambiguated by the K-NN classifier. The dataset and the baselines are the same as in Section 6.1.

Notice that, differently from the experiments described in Table 3, the run JIGSAW+K-NN has not been reported since JIGSAW covered all the target words in the first step of the cascade hybrid method, then the K-NN method is not applied at all. Therefore, for this run, results obtained by JIGSAW+K-NN correspond to those get by JIGSAW alone (reported in Table 2).

Table 4 reports the results of all the runs. Results are very similar to those obtained in the runs K-NN+JIGSAW with the same settings on φ . Precision tends to grow, while a corresponding decrease in recall is observed. The main outcome is that the overall accuracy of the best combination JIGSAW+K-NN ($\varphi \geq 0.80$, F and P values highlighted in bold in Table 4) is outperformed by K-NN+JIGSAW. Indeed, this result was largely expected because the small size of the training set does not allow to cover words not disambiguated by JIGSAW.

The general conclusion of the experimental session is that the most effective hybrid WSD strategy is the one that integrates JIGSAW into K-NN.

Setting	P	R	F	A
JIGSAW ($\varphi \geq 0.80$) + K-NN	0.715	0.392	0.556	55.6
JIGSAW ($\varphi \geq 0.60$) + K-NN	0.688	0.440	0.537	64.0
JIGSAW ($\varphi \geq 0.40$) + K-NN	0.651	0.484	0.555	74.4

Table 4. Experimental results of JIGSAW+K-NN

Even if K-NN+JIGSAW is not able to achieve the baselines set on the 1st *sense* heuristic (first and last row in Table 2), we can conclude that a step toward these hard baselines has been moved. The main outcome of the study is that the best hybrid method on which further investigations are possible is K-NN+JIGSAW. It is important to underline that it is very hard to outperform the 1st *sense* baseline. Indeed, as regards English WSD All-Words task, only few supervised systems⁶ are able to reach this goal. We reasonably think that this could be considered true also for Italian, even if, unfortunately, we do not have the possibility to compare our system with other ones.

⁶ English supervised systems can rely on larger training corpus than Italian WSD ones.

7 Conclusions and Future Work

This paper presented a method for solving the semantic ambiguity of *all words* contained in a text. We proposed a hybrid WSD algorithm that combines a knowledge-based WSD algorithm, called JIGSAW, which we designed to work by exploiting WORDNET-like dictionaries as sense repository, with a supervised machine learning algorithm (K-Nearest Neighbor classifier). The idea behind the proposed approach is that JIGSAW can cope with the possible lack of training data, while K-NN can improve the precision of JIGSAW method when training data are available. This makes the proposed method suitable for disambiguation of languages for which the available resources are lacking in training data or sense definitions, such as Italian.

Extensive experimental sessions were performed on the EVALITA WSD All-Words-Task dataset, the only dataset available for the evaluation of WSD systems for the Italian language. An investigation was carried out in order to evaluate several combinations of JIGSAW and K-NN. The main outcome is that the most effective hybrid WSD strategy is the one that runs JIGSAW after K-NN, which outperforms both JIGSAW and K-NN taken singularly. Future work includes 1) the investigation of other ways of combining different algorithms, for example the *JIGSAW* output could be used as a feature into a supervised system or 2) the use of different supervised algorithms.

References

1. E. Agirre, B. Magnini, O. Lopez de Lacalle, A. Otegi, G. Rigau, and Vossen. SemEval-2007 Task 1: Evaluating WSD on Cross-Language Information Retrieval. In *Proceedings of SemEval-2007*. Association for Computational Linguistics, 2007.
2. S. Banerjee and T. Pedersen. An Adapted Lesk Algorithm for Word Sense Disambiguation Using WordNet. In *CICLing '02: Proceedings of the Third International Conference on Computational Linguistics and Intelligent Text Processing*, pages 136–145, London, UK, 2002. Springer-Verlag.
3. P. Basile, M. de Gemmis, A.L. Gentile, P. Lops, and G. Semeraro. JIGSAW algorithm for Word Sense Disambiguation. In *SemEval-2007: 4th International Workshop on Semantic Evaluations*, pages 398–401. ACL Press, 2007.
4. P. Basile, M. de Gemmis, P. Lops, and G. Semeraro. Combining Knowledge-based Methods and Supervised Learning for Effective Italian Word Sense Disambiguation. In Johan Bos and Rodolfo Delmonte, editors, *Semantics in Text Processing - STEP 2008 Conference Proceedings*, volume 1 of *Research in Computational Semantics*, pages 5–16. College Publications, 2008.
5. P. Basile, M. Degennaris, A. L. Gentile, P. Lops, and G. Semeraro. The JIGSAW Algorithm for Word Sense Disambiguation and Semantic Indexing of Documents. In Roberto Basili and Maria Teresa Pazienza, editors, *AI*IA*, volume 4733 of *Lecture Notes in Computer Science*, pages 314–325. Springer, 2007.
6. L. Bentivogli and E. Pianta. Exploiting parallel texts in the creation of multilingual semantically annotated resources: the MultiSemCor Corpus. *Natural Language Engineering*, 11(03):247–261, 2005.

7. B. Decadt, V. Hoste, W. Daelemans, and A. Van den Bosch. Gambl, Genetic Algorithm optimization of Memory-based WSD. In *Senseval-3: 3rd International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*, 2004.
8. M. Diab. Relieving the Data Acquisition Bottleneck in Word Sense Disambiguation. In *Proceedings of ACL*, 2004. Barcelona, Spain.
9. C. Fellbaum. *WordNet: An Electronic Lexical Database*. MIT Press, 1998.
10. V. Hoste, W. Daelemans, I. Hendrickx, and A. van den Bosch. Evaluating the results of a memory-based word-expert approach to unrestricted word sense disambiguation. *Proceedings of the ACL-02 workshop on Word sense disambiguation: recent successes and future directions-Volume 8*, pages 95–101, 2002.
11. C. Leacock and M. Chodorow. *Combining local context and WordNet similarity for word sense identification*, pages 305–332. MIT Press, 1998.
12. C. Leacock, M. Chodorow, and G. Miller. Using corpus statistics and WordNet relations for sense identification. *Computational Linguistics*, 24(1):147–165, 1998.
13. V. I. Levenshtein. Binary codes capable of correcting deletions, insertions and reversals. *Soviet Physics Doklady*, 10:707–710, 1966.
14. R. Mihalcea. Bootstrapping large sense tagged corpora. In *Proceedings of the 3rd International Conference on Language Resources and Evaluations*, 2002.
15. R. Mihalcea. Unsupervised large-vocabulary word sense disambiguation with graph-based algorithms for sequence data labeling. In *HLT '05: Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 411–418, Morristown, NJ, USA, 2005. Association for Computational Linguistics.
16. R. Mihalcea. Using Wikipedia for Automatic Word Sense Disambiguation. In *Proceedings of the North American Chapter of the Association for Computational Linguistics*, 2007.
17. R. Mihalcea and T. Chklovski. Open Mind Word Expert: Creating Large Annotated Data Collections with Web Users' Help. In *Proceedings of the EACL Workshop on Linguistically Annotated Corpora, Budapest*, 2003.
18. T. Mitchell. *Machine Learning*. McGraw-Hill, New York, 1997.
19. I. Nancy and J. Véronis. Introduction to the special issue on word sense disambiguation: The state of the art. *Computational Linguistics*, 24(1):1–40, 1998.
20. R. Navigli and P. Velardi. Structural semantic interconnections: A knowledge-based approach to word sense disambiguation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27 (7):1075–1086, 2005.
21. A. Roventini, A. Alonge, F. Bertagna, N. Calzolari, J. Cincila, C. Girardi, B. Magnini, R. Marinelli, M. Speranza, and A. Zampolli. ItalWordNet: building a large semantic database for the automatic treatment of Italian. *Computational Linguistics in Pisa - Linguistica Computazionale a Pisa. Linguistica Computazionale, Special Issue, XVIII-XIX*, Tomo II:745–791, 2003.
22. G. Semeraro, M. Degennaris, P. Lops, and P. Basile. Combining learning and word sense disambiguation for intelligent user profiling. In *Proceedings of the Twentieth International Joint Conference on Artificial Intelligence IJCAI-07*, pages 2856–2861, 2007. M. Kaufmann, San Francisco, California. ISBN: 978-I-57735-298-3.
23. D. Yuret. Some experiments with a naive bayes WSD system. In *Senseval-3: 3rd Internat. Workshop on the Evaluation of Systems for the Semantic Analysis of Text*, 2004.

Searching and Browsing Digital Library Catalogues: A Combined Log Analysis for The European Library

Maristella Agosti¹, Franco Crivellari¹, Giorgio Maria Di Nunzio¹, Yannis Ioannidis², Elefterios Stamatogiannakis², Mei Li Triantafyllidi², and Maria Vayanou²

¹ Department of Information Engineering – University of Padua
Via Gradenigo, 6/a – 35131 Padova – Italy
{agosti,crive,dinunzio}@dei.unipd.it

² Department of Informatics and Telecommunications – University of Athens
Panepistimiopolis, Ilissia – 15784 Athens – Greece
{yannis,estama,meili,vayanou}@di.uoa.gr

Abstract. The interaction between a user and a digital library catalogue can be analyzed and studied in order to gather user preferences and learn what the user likes the most and use this information to present the results in a personalized way. Since research on personalization of contents relies on actions performed by people, it is essential to identify user sessions in order to capture user intentions and preferences in a particular instant of time.

In this paper, we present a combined analysis of two different logs, action log and HTTP logs, of The European Library which is a free service giving access to resources of the 48 national libraries of Europe. A user study was conducted in order to collect into HTTP logs enough data to study the browsing activity and analyze possible relations between explicit preferences collected by online questionnaires and implicit actions recorded in the logs.

1 Introduction

Library catalogues, and in particular National library catalogues, are important sources for the preservation and diffusion of cultural material which can be stored on different media, for example books, files, pictures, and so on. The transformation of library catalogues into digital library catalogues has given the possibility to access library material remotely in an integrated way, and often includes a library catalogue module as the public interface to the system's inventory. The user is usually allowed to find items according to a set of choices, among them: author, title, subject.

The interaction between the user and the system can be analyzed and studied in order to gather user preferences and “learn” what the user likes the most, and use this information to present the results in different ways for each user. User preferences can be learned explicitly, for example asking the user to fill-in

questionnaires, or implicitly, studying the actions of the user which are recorded in the action log of a system. The second choice is certainly much less intrusive but requires more effort to reconstruct each search session a user made in order to learn his/her preferences.

The problem of identifying user sessions may become easier if authentication was mandatory for a user who wants to access a service. Authentication would provide the exact time of a session intended as a sequence of events like: login, actions and browsing, logout. However, authentication is not always required, especially when dealing with online free services. In these cases, since research on personalization of contents relays on actions performed by people, it is essential to identify user sessions in order to capture user intentions and preferences in a particular instant of time. Organizing the requests in a single session permits to have a better view of the actions performed by visitors. A procedure, named “session reconstruction”, may be used in order to map the list of activities performed by every single user to the visitors of the site. A possible approach to isolate a single session of a user is the use of the pair IP address and user agent [1], and permits only a fixed gap of time between two consecutive requests [2]. This type of heuristic has been widely used in literature and works pretty well; it is important, however, to bear in mind that with this approach it is not possible to identify the specific instant in which a client leaves the site, therefore statistics like time spent on the Web site, or average of a session, are biased.

The main focus of this paper is to analyze the logs, for personalization purposes mentioned above, of The European Library (TEL)¹. The European Library is a free service of the Conference of the European National Librarians (CENL)² for anyone world-wide seeking for library material. It was launched in 2005 to offer a common access point to the distributed resources of the 48 national libraries of Europe in 20 languages (for a comprehensive reference to The European Library and its history from project to operational service, please refer to [3]).

TELplus³ is a project that is being taken forward by a consortium of 26 partners, national libraries and research centres to form a pool of research and innovation to provide value-adding services and products for The European Library. The project provides a major research and innovation hub for The European Library. Each work package aims at either significantly raising the amount of digital content in The European Library, or improving access to content and the overall usability of the service. Log file analysis can provide a rich source of information that can feed into improving the overall usability of the portal, creating particular services for specific user groups, offering true personalized search and retrieval according to a user individual profile [4]. The work here reported has been conducted in the context of Work Package 5, which is the work

¹ <http://www.theeuropeanlibrary.org/>

² <http://www.cenl.org/>

³ <http://www.theeuropeanlibrary.org/telplus/>

package of the project devoted to the conceiving of user personalization services also through logs analysis⁴.

The European Library portal does support authentication but in a soft way: a user can register and login if she wants, but she is free to use almost all the search services without authentication. However, there are pieces of information in The European Library we can use to partially reconstruct a session of a user: this information is stored in the cookie the The European Library Web server sends to the client when the first HTTP request is sent to the portal. In fact, the cookie contains an identifier, called TEL session identifier (or TELsessid), which should identify a session. Actually, this identifier stores information about user preferences and actions on a particular computer, the one a user is interacting with, and does not contain information about sessions. To clarify this point we describe some scenarios, supposing that nobody authenticates in the Web site:

- the user browses the TEL portal for a while, then closes the browser, leaves and reopens the browser for another search later. There are two distinct sessions but only one TELsessid;
- a user finishes his browsing activity and after a while another user starts to browse the portal on the same computer. There are two different users and sessions, but only one TELsessid;
- the user browses and performs some searches on the TEL portal, then for privacy issues he deletes the cookies stored in the browser and continues browsing. There is a single session, but two TELsessid are issued by the TEL Web server.

There could be many other examples of this kind, which demonstrates that this task is difficult and requires many different strategies to try to get close to the real session a user made.

To present the work conducted and the reached results on this difficult task the paper is organized as follows: Section 2 describes The European Library environment and the main results reached through log analysis and user studies are presented in Sections 3 and 4 respectively. Section 5 reports on results about the portal's general usage while specific focus is paid on the searching behavior in Section 6. Finally, collection related issues are investigated in Section 7 and significant correlations are highlighted, drawing a variety of conclusions. Final remarks are given in Section 8.

2 The European Library Environment

The European Library environment consists of three different components that must be analyzed together to produce a complete traffic overview: The European Library portal, action logs and HTTP logs.

In The European Library portal's home page, a user can initiate a simple keyword search with a default predefined collection list presenting catalogues from

⁴ <http://www.theeuropeanlibrary.org/telplus/workplan.php>

national libraries. From the same page, a user may click *advanced search* link where she can enrich her query with Boolean operators and limit search to specific fields like author, language, or ISBN. Alternatively from the *simple search* page, she may change the searched collection by checking the theme categories below the search box. After search button is clicked the result page appears, where results are classified by collections and the results of the top collection in the list are presented with brief descriptions. We will call the presentation of result records description, a *result list* page. At this point, a user may choose to see *result lists* of other collections or move to the next page of records of current collection's results. While viewing a *result list* page a user may also click on a specific record to see detailed information about the specific record. Depending on the selected collection and record viewed, additional services may be available e.g. links leading to a library site. Additionally, while still being at the result page, a user can perform a new search or search within the current *result list*.

The European Library environment keeps detailed logs of users' actions within the portal. Action logs are stored in a table of a relational database, where a table record represents a user action. The most significant columns of the table are:

- *userid*: A numeric id, for identifying registered users or 'guest' otherwise.
- *userip*: User's IP address.
- *sessid*: An automatically generated alphanumeric, identifying sequential actions of the same user (sessions).
- *query*: Query contents.
- *action*: Name of the action that a user performed.
- *colid*: The corresponding collection's alphanumeric id.
- *date* : Date and time of the action's occurrence.

A detailed description of the possible values of action field is listed in Table 1. Action logs' analysis, presented in this paper covers the period from 1st January 2007 until 18th July 2008.

HTTP log files contain the records of all the HTTP requests made by clients to the Web server of The European Library. Here, we present the choice of those fields included in the log files we have analyzed:

- *date*: Date, in the form of yyyy-mm-dd.
- *time*: Time, in the form of hh:mm:ss.
- *cs-method*: The requested action. Usually GET for common users.
- *cs-uri-stem*: The URI-Stem of the request.
- *cs-uri-query*: The URI-Query, where requested.
- *c-ip*: The IP address of the client.
- *cs(User-Agent)*: User-Agent of the Client. For a standard user this means the browser and other information about operating system.
- *cs(Referrer)*: The site where the link followed by the user was located.

HTTP log analysis has been carried out on the data collected during the user studies described in Section 4.

Table 1: Description of The European Library logging actions and classification in action types.

Action type	Action value	Description
<i>Search:</i> all search related actions	search_sim	simple search
	search_adv	advanced search
	search_res	search initiated from results page
	search_url	search initiated from url query string
<i>Result List Browsing:</i> actions related to presentation of search result lists	view_brief	display result list page
	page_brief	navigation between result list pages using "next" or "previous" buttons
	jump_to_page	jump to result list page with user specified sequential number
<i>Result clicking:</i> actions indicating selection and view of an object's full record information	view_full	display/navigate between result record pages
<i>Collection selection:</i> actions to choose collections to be searched	col_set_theme	<i>Theme collection selection:</i> Collections chosen from theme list (checkbox in homepage) or from country list (dropdown list in homepage)
	col_set_theme_country	
	col_set_country	<i>Advanced collection selection:</i> Save link clicked after selecting collections in collections tab, that are browsed by country or subject or searched by description
	col_set_subj	
	col_set_desc	
	col_set_default	collections default list reinstated
<i>Result retention:</i> actions denoting a user's wish to retain information about a specified object.	option_save_session_favorite	result record saved in favorites
	option_send_mail	sent by email result record
	option_print	print result record
	options_save_reference	record saved for reference manager use
<i>Outgoing:</i> actions resulting in user redirection towards library sites or services	service_X	full record service link used of country X
	service_all	full record service link used
Show help file	show_help_X	displays help file X
<i>Result clicking / Outgoing</i>	available_at	"Available at Library" link clicked to view record in native interface
<i>Result clicking / Outgoing</i>	see_online	"See online" link clicked to see object in native interface

Table 2: Distribution of registered users declared profession.

Profession	Number of users	Profession	Number of users
Student - Undergraduate	522	Librarian - Administrator	93
Other	514	High School Teacher	65
Student - Postgraduate	304	Media/Journalist	51
Researcher	269	Librarian - Faculty Liaison	33
University Teacher/Professor	221	Librarian - Corporate	23
Professional/Practitioner	159	Bookseller	19
Librarian - Other	146	Total	2419

3 Registered Users

The European Library service offers users the opportunity to register and benefit from personalized services, e.g. registered users may store favorites for later sessions and save search sessions in their history. Until July 2008, the user registration procedure has been performed 2,419 times. However this number does not correspond to actual registered users, since there are firm indications, that some users have registered more than once. TEL's low registration numbers suggest that users are not motivated enough to register. Lack of user motivation is also supported by the registered user log-in frequency. In detail, approximately 84% of registered users have performed one single registered session since the beginning of 2007.

Apart from uniquely defining users, registration procedure constitutes also a valuable source of explicit user feedback. The optional and mandatory information, filled in the registration form, reveals the addressed user group characteristics. Approximately half of the registered users come from academic community and most of them are undergraduate or postgraduate students. The second most popular value chosen for the *profession* field is *Other* (21.2%), indicating that TEL is accessed by people, whose type of work is not reflected in choices available within the drop-down list. The distribution among the available professions is depicted in Table 2. Approximately half of the registered users come from the academic community and most of them are undergraduate or postgraduate students.

Also the results on The European Library portal's usage by registered users are reported in Section 5.

4 User Studies

Log data are not the only form of data which can capture user preferences; there are also explicit ways of gathering these preferences, such an example are user surveys. With the aim of gaining insights by analyzing log data together with

data from controlled studies, a user's study was conducted at the University of Padua, Italy, in order to study specific aspects of The European Library portal, focusing mainly on: parts of the Web site, clarity of the Web site, level of satisfaction.

This study was conducted in a controlled setting during the end of 2007 and the beginning of 2008, in the computer laboratories of different faculties of the university, with at least one person of the same university, present in the laboratory together with users that were all students of one of the faculties of the university. All students were attending one course that was taught by one person of the university involved in the TELplus project. This means that the students were all attending a course in some way pertinent to the contents of the questionnaire and the use of The European Library portal was coherent with the aim of the course's part that was taught at the time of the use of the portal and of the filling in of the questionnaire. In order to track the activity of each user and find him in the HTTP logs, we asked to change the user agent with a string which would have been easy to be identified in the logs. This identification played an important role because it gave us the possibility to reconstruct each user session in a clear way, without the use of heuristics.

At the end, a total of 216 students participated in these studies, but it was possible to clearly recognize in the logs only 151 of them. This is due to errors during the typing of the user agent which made it impossible to find user in the logs, or two identical user agents with overlapping time, in this case we preferred to discard this information so as to not introduce more bias.

It is important to underline that the sample of students cannot be considered as a significant sample of all the user of The European Library service; however, the results of the analysis on this group of students are useful to understand the behavior of undergraduate students of Humanities, specifically from Italy, and in particular from University of Padua. These are users who can be interested in using the portal and searching of bibliographic records and, for this reason, their judgements about the interface and the services have to be seriously taken into consideration for future improvements.

5 General Usage

In this Section, we analyze The European Library portal's general usage from the following points of view: action logs, HTTP logs, and HTTP Logs of User Studies.

5.1 Action Logs

The average session duration is estimated to approximately 6 minutes while median duration is limited to 2. In terms of actions, the average session length is almost 8 actions and median is 4 actions. Comparing with registered user sessions, where the median session duration reaches almost 5 minutes and median

Table 3: Number of actions per action category.

Action type	#	Type of action	#
Search	584,687	Outgoing actions	55,516
Result list browsing	478,585	Collection selection	200,009
Result clicking	626,007	Show help files	2,196
Result retention	13,154	Total actions	1,916,134

length is 8, one can safely conclude that registered users' sessions indicate a higher level of users' expertise than in the general case.

We have thematically classified actions into 7 categories, namely:

- *Search*,
- *Result list browsing*,
- *Result clicking*,
- *Result retention*,
- *Outgoing actions*,
- *Collection selection*, and
- *Show help files* actions.

Table 3 depicts the frequency of each action type. It is worth noting that some actions belong to more than one category, hence the sum of actions in Table 3 exceeds the total number of logged actions. The category label assigned to each individual action is reported in the *Action type* field of Table 1.

Ideally actions of type *result list browsing* should follow *search* actions, being at least the same in number as *Search* actions. The discrepancy's explanation is that users abandoned their searches before The European Library service returned any search results, probably due to The European Library service response time.

5.2 HTTP Logs

The HTTP log data were collected over the period of time which goes from January 2007 to June 2008. Table 4 reports descriptive statistics computed on the collected HTTP log data.

As reported in Table 4, in the analyzed period of time the number of unique visitors was more than 475 thousands, with an average of 870 visitors per day. More than 700 thousands visits where recorded with a daily average of about 1,300; these visits produced around 58 millions of hits (HTTP requests to the Web server) which corresponds to a daily average of 106 thousands of contacts. The number of Web pages requested were more than 19 millions, which corresponds to almost 35 thousands of pages per day. "Not viewed" traffic, both hits and bandwidth, includes traffic generated by robots, worms, or replies with special HTTP status codes.

Table 4: HTTP Log Synthetic Descriptive Statistics.

	Total	Daily Average	Per visit
No. of visitors	475,333	868.98	Not applicable
No. of visits	709,922	1,297.85	1.49 (per visitor)
Hits	58,053,293	106,130.3	81.77
No. of accessed pages	19,047,263	34,821.32	26.837
Bandwidth	1,059.51 GB	1.94 GB	1.53 MB
Not Viewed Hits	14,229,379	106,130.3	20.04
Not Viewed Bandwidth	2,876.63 GB	5.26 GB	4.15 MB

The two plots reported in Figure 1 show the trend of the total number of visitors of The European Library for each month of the period under analysis. In the first plot, the difference between the maximum value (January 2008) is reported, while in the second plot the numbers of visitors are reported.

If we consider the length of the sessions of people who connected to The European Library, we can see that the majority of the sessions last no more than two minutes. The distribution of the length of sessions can be summarized with the Pareto diagram reported in Figure 2.

The analysis of log data for each month, or even a shorter period like a week or a day, is fundamental for understanding specific and punctual analysis on a particular variable of interest, for example the study of the actions of a particular group of users during the day of a week. On the other hand, aggregated data are more suitable for presenting the results of long periods of analysis. In order to analyze the trend of the traffic of one year and a half of accesses to The European Library portal, we decided to aggregate the log data in group of three months to follow the possible seasonal variations. An example of this type of analysis is reported in Figure 3 where the variation of seasonal trends of the length of sessions is reported. The great majority of visits do not last more than a couple of minutes. It is interesting to study the trend of the percentage of visits duration compared to the total.

5.3 HTTP Logs of User Studies

We have analyzed the HTTP requests made by the 151 students of the University of Padua which have been recognized in the HTTP logs. Given the fact that this was intentionally a controlled experiment the number of sessions is obviously equal to the number of students. The analysis carried out were the following: statistics on the number of requests made by each user, statistics on the length of a session in terms of filling-in the questionnaire and of browsing the portal.

In Table 5 a summary of the statistics of the number of records is presented. Figures show that the distribution of these values is not symmetric: values are skewed to the lowest values, the median, is in fact closer to the first quartile. The mean, a measure highly affected by extreme values, is equal to 228.3 requests, this happens because there are a number of users who performed a high number

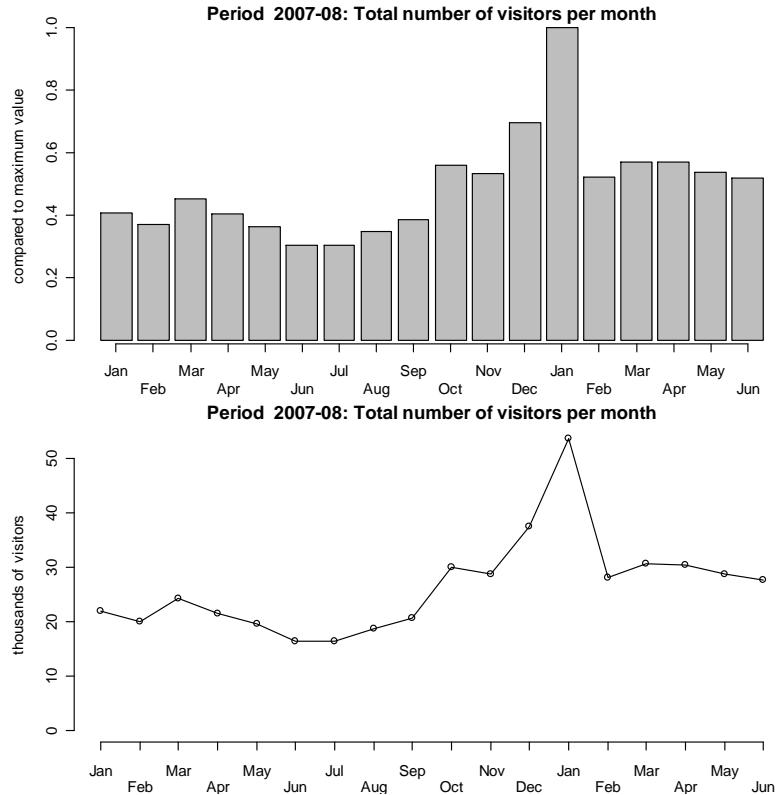


Fig. 1: Number of visitors per month.

of requests per session. The maximum number of requests was 866, while the minimum was 4 corresponding to a student who, at the end of the questionnaire, cleaned the cache of the browser and incidentally reopened the page of The European Library portal, which produced a new TEL session identifier and a sequence of 4 HTTP requests.

HTTP requests can also be divided in classes of frequencies to study modes of the distribution. It is interesting to note that about 60% of the HTTP sessions are classified sessions with less than 250 requests. If we consider that each single click on a link of a Web page produces multiple HTTP requests, this result means that the majority of sessions are very short in terms of browsing activity. This behavior reflects what has been found in previous studies on The European Library portal usage: sessions tend to be short and with few requests [5].

A study on the relation between the time a user spent to compile the questionnaire during the survey and the time a user spent on browsing the portal was carried out. For the first measure there were time constraint due to the available time slots of University computer labs. In Table 6 the statistics for this time

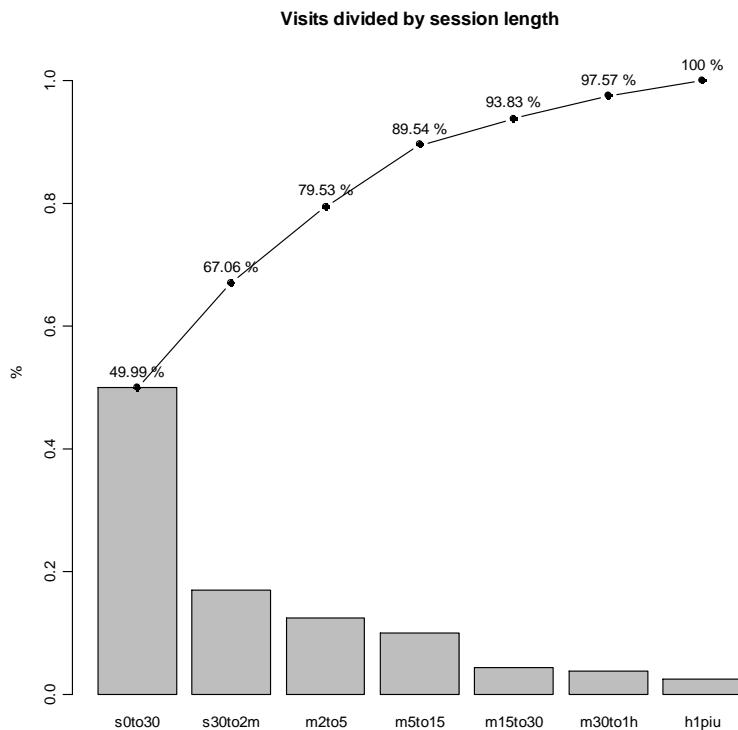


Fig. 2: Distribution of session lengths with a Pareto diagram.

are shown. The time is clearly bounded between 30 and 45 minutes which were the usual slots. In the same table, we also reported the number of requests per minute.

In Table 7 another statistic is shown, the one relative to the time spent for browsing calculating as the difference between the first and the last HTTP request and the average number of requests per minute. The numbers are similar to those calculated for the questionnaire in Table 6, this means that the students browsed the portal while filling in the questionnaire. These results are interesting when compared to the average time of a search session presented in Section 5.1 which presented a median of about 2 minutes and an average of 6 minutes.

6 Searching in The European Library

The European Library service provides users with two interfaces for formulating queries, namely simple search and advanced search. Simple search is similar to a traditional web search interface, where the user inputs terms separated

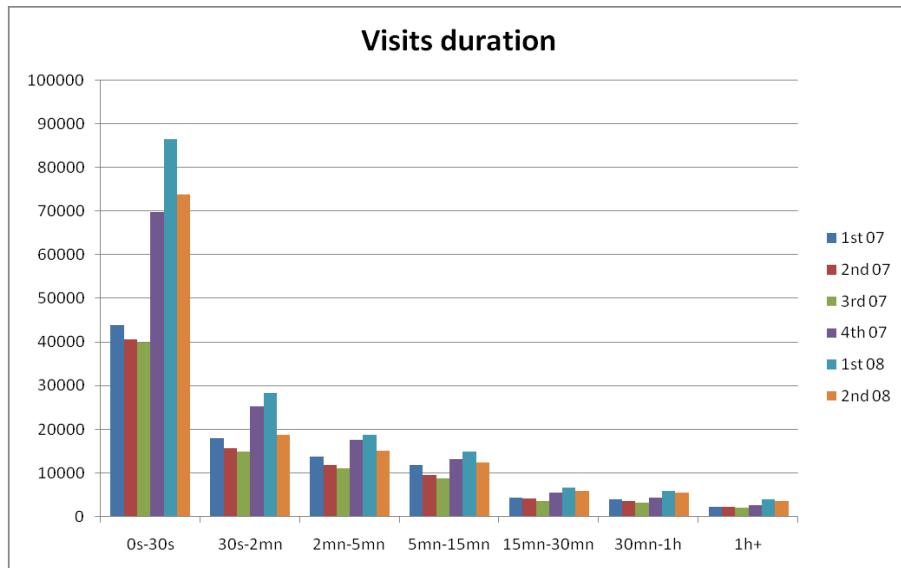


Fig. 3: Visit duration grouped quarterly.

Table 5: Summary of statistics for the number of HTTP requests per user.

Minimum	1st Quartile	Median	Mean	3rd Quartile	Maximum
4.0	99.0	175.0	228.3	311.5	866.0

by space, having an implied conjunction between them. In advanced search, The European Library service provides attribute fields (Title, Creator, Subject, Type, Language, ISBN, ISSN) for query formulation, along with a set of Boolean operators (AND, OR, NOT). A user can chose from a drop-down list specific fields that should be searched for, meeting conditions provided as user input next to the list. We refer to queries containing operators or attribute fields as advanced queries, discriminating them from simple keyword queries.

A session in average contains 2.3 queries, having an average query length of 2.6 terms per query.

Approximately 50% of overall logged queries are unique queries while 10% of the query terms have occurred only once, thus emphasizing the need to handle queries at a term level, so as to efficiently leverage past users' search interactions. Furthermore, approximately 3% of simple keyword queries are empty.

Table 8 shows syntactic elements found in The European Library *advanced queries*, alongside their number of occurrences. Noteworthy is that *advanced queries* constitute 13.2% of all queries, firmly suggesting user preference for simple keyword search.

Turning to keyword queries, Table 9 depicts the frequencies of most popular queries alongside frequencies of most popular search terms.

Table 6: Summary of statistics for the time spent to fill-in the questionnaire during the user survey (first row), and the number of requests per minute (second row).

	Minimum	1st Quartile	Median	Mean	3rd Quartile	Maximum
minutes	7.0	28.0	31.0	33.55	40.0	60.0
requests/min	0.14	2.85	5.56	7.43	9.92	56.0

Table 7: Summary of statistics for the time spent for browsing the portal (first row), and the number of requests per minute (second row).

	Minimum	1st Quartile	Median	Mean	3rd Quartile	Maximum
minutes	0.03	25.58	30.25	31.80	38.90	57.38
requests/min	0.28	3.05	6.17	8.35	11.12	120.00

7 Collection Selection in The European Library

Since The European Library catalogues contain more than 300 collections, providing users with collection selection options is a necessity.

We have classified collection selection options in the following categories (descriptions of related actions are shown in Table 1):

- *Default collection selection* category consists of sessions not containing collection selection actions, thus using TEL’s default collection selection. Overall, 65% of sessions fall in this category.
- *Theme collection selection* category, consists of sessions containing only *theme collection selection* actions. 33% of sessions follow the aforementioned collection selection pattern.
- *Advanced collection selection* category consists of sessions containing at least one *advanced collection selection* action. Overall, 2% of sessions are in this category.

Extensive use of theme collection selection feature is justified by its simplicity, requiring minimum user effort while being accessible within the search page.

Investigating collection selection methods other than those already provided by The European Library, we correlated nationalities of collections accessed by users with their geographic location (using IP geolocation). Figure 4 depicts a representative sample of this distribution. Strong preference of users towards collections having the same nationality as their location, is evident. Another noteworthy user behavior pattern is that users in Spain, Italy, France and Canada favor French collections against German ones, while users in Germany, Poland, Hungary and Croatia have the exact opposite behavior.

8 Conclusions

Studies on log files are essential for personalization purposes, since they implicitly capture user intentions and preferences in a particular instant of time.

Table 8: Frequencies of The European Library syntactic elements.

Syntax	#	Syntax	#
AND	26229	Type	880
OR	1268	Language	3722
NOT	415	ISBN	4695
Title	37024	ISSN	1171
Creator	31029	Advanced queries	77057
Subject	7002	Total queries	583422

Table 9: Frequencies of most popular queries (first column) and most popular terms (second column).

Query	#	Term	#
mozart	6634	mozart	7438
meisje met de parel	2001	history	2739
harry potter	1873	european	2549
van gogh	1730	journal	2349
pink floyd	717	potter	2308
nuremberg	657	harry	2271
rembrandt	609	international	2065
einstein	563	meisje	2032

The study carried out and presented in this paper took into account two different kinds of log files: action logs and HTTP logs. Action logs allowed to track registered users of the portal and, more in general, analyze search preferences. A user study was conducted in order to collect into HTTP logs enough data to study the browsing activity and analyze possible relations between explicit preferences collected by online questionnaires and implicit actions recorded in the logs.

In general, user sessions are short, on average less than ten minutes according to sessions reconstructed from both action logs and HTTP logs. There are at most a couple of search actions with a limited browsing activity of the results and of the Web site itself.

A possible comment to the length of a session could be that a user, that visits the portal for the first time, most of the times starts to conduct a Google-type search using the search form. Since the answer from the portal is different from a Google-type answer, the user goes away and never returns (or most of the times he does not return). This is in part reflected by the answers collected by questionnaires during user studies, where users were not satisfied by the presentation of the results and by the content of the results.

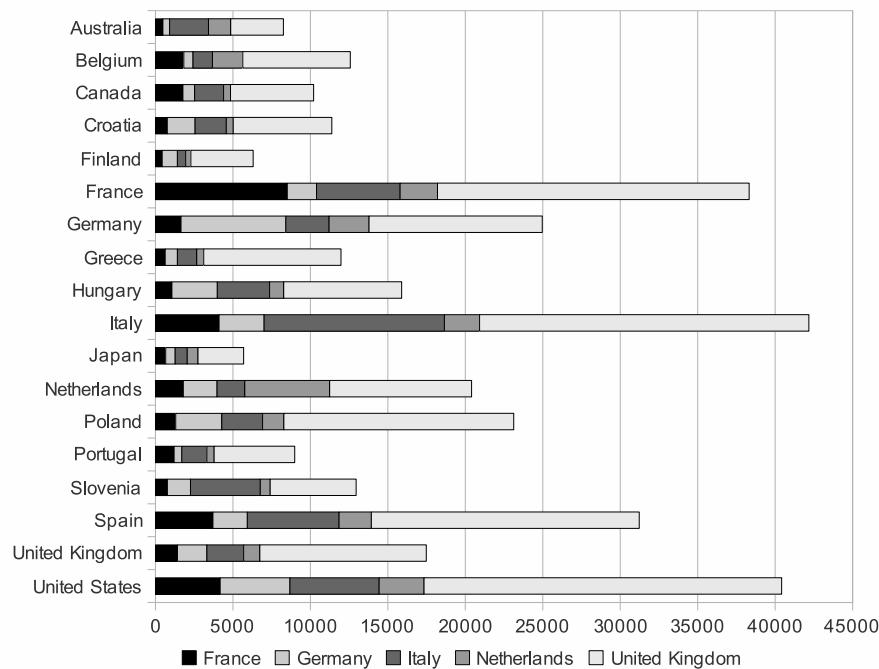


Fig. 4: Distribution of TEL traffic, group by originated country (country extracted from user IP geolocation) on 5 of most popular collection countries.

Some other evidences captured by log files is that there is a strong preference of users towards collections having the same nationality as their location. At the same time, users are willing to browse and search collections in different languages.

Acknowledgements

The authors would like to thank the staff of The European Library Office for their systematic support during the analysis of logs which is reported in the paper.

The work reported has been partially supported by the TELplus Targeted Project for digital libraries, as part of the eContentplus Program of the European Commission (Contract ECP-2006-DILI-510003).

References

1. Nicholas, D., Huntington, P., Watkinson, A.: Scholarly Journal Usage: the Results of Deep Log Analysis. *Journal of Documentation* **61**(2) (2005) 248–280

2. Berendt, B., Mobasher, B., Nakagawa, M., Spiliopoulou, M.: The Impact of Site Structure and User Environment on Session Reconstruction in Web Usage Analysis. In Zaïane, O.R., Srivastava, J., Spiliopoulou, M., Masand, B.M., eds.: Knowledge Discovery on the Web (WEBKDD) - MiningWeb Data for Discovering Usage Patterns and Profiles. Volume 2703 of Lecture Notes in Computer Science., Springer (2003) 159–179
3. Cousins, J., Chambers, S., van der Meulen, E.: Uncovering cultural heritage through collaboration. *Int. J. on Digital Libraries* **9**(2) (2008) 125–138
4. Angelaki, G.: Research and Innovation in The European Library; the TELplus Project. Globalisation and the Management of Information Resources <http://slim.emporia.edu/globenet/Sofia2008/>, Sofia, Bulgaria (Nov 2008)
5. Agosti, M., Di Nunzio, G.: Web Log Mining: A study of user sessions. In: Proc. 10th DELOS Thematic Workshop on Personalized Access, Profile Management, and Context Awareness in Digital Libraries (PersDL 2007), Corfu, Greece (Jun 2007)

The On-TIME project

Tiziana Catarci¹, Alan Dix², Raffaele Giuliano, Marco Piva,
Antonella Poggi¹, Fabio Terella, Emanuele Tracanna

¹SAPIENZA Università di Roma, Italy

{catarci, poggi}@dis.uniroma1.it

²Lancaster University, United Kingdom

alan@hcibook.com

Abstract. The On-TIME project was born as follow-up of a DELOS task within the User Interface and Visualization workpackage. Its general aim is to provide the user with a system that allows her to focus on tasks rather than just on managing her personal information, as in traditional personal information management systems. To achieve this goal, user data and tasks are described in terms of explicit semantics that the user can share, i.e., by means of a Personal Ontology, reflecting the user's view of the world and her personal interests. In this paper, we describe the main current achievements of the on-going On-TIME project, namely the actual architecture and the effective user interface of a first working prototype.

Keywords: Personal Information Management, Task Management, User Interface.

1 Introduction

Today's personal desktops have become from far the most commonly/frequently used personal digital library. However, personal desktops mostly consist of a disconnected set of generic tools, that the user interacts with, often by manually repeating similar tasks and copying for this the same data from one application to the other several times.

This has motivated the arising of a specific research task within the DELOS Network Of Excellence, more precisely within the User Interface and Visualization workpackage. One of the outcomes of this task was the definition of a new paradigm of system, called Personal Interaction Management System (PIMS) [4], that allows the user to focus on the tasks they have to perform rather than just managing their personal information. The main feature of a PIMS is to rely on the use of a Personal Ontology reflecting the user's view of the world and her personal interests, to describe both user data and tasks in terms of explicit semantics that the user can share. This represents a challenge since tasks need to be explicitly represented both in their static aspects, i.e. the kind of information that they manipulate, the kind of programs involved in these manipulations, security and authentication issues that may arise, and in their dynamic ones, i.e. the sequences of actions that they require, the alternative

choices that are given to the user, pre-conditions, post-conditions, and invariants for the various actions that are involved in the task.

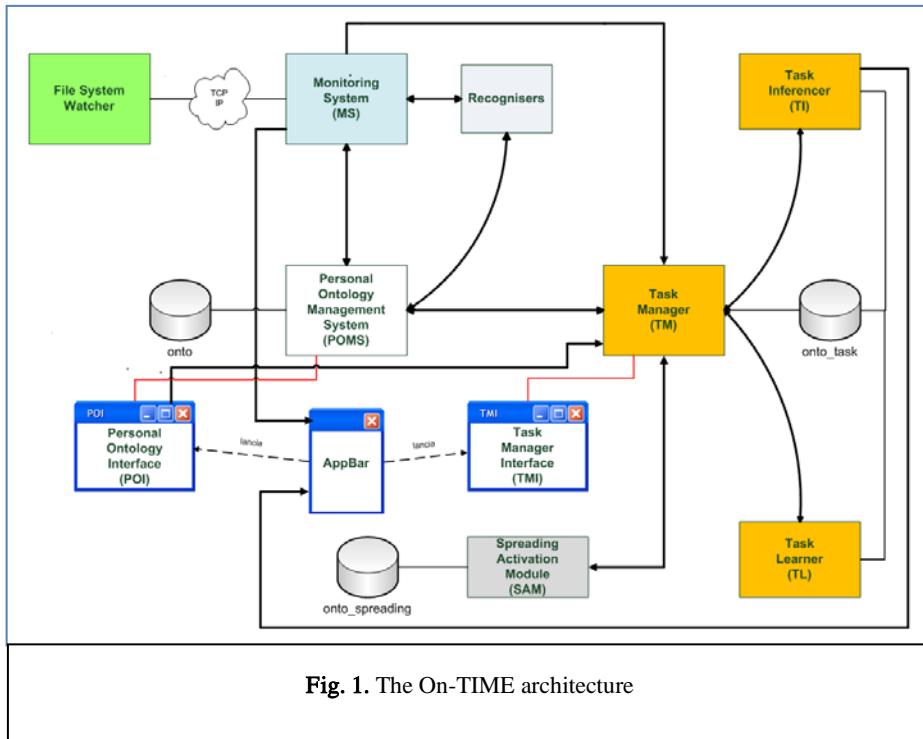
Other outcomes of the abovementioned DELOS task were (i) the definition of a “core” Personal Ontology to be further extended depending on the user stereotype and (ii) the design of a preliminary architecture based on the use of an existing web bookmarking tool, called Snip!t [5], whose distinguishing feature is to “intelligently” suggest appropriate actions to perform starting from a Web page content.

The On-TIME project was born as follow-up of the DELOS task. Specifically, On-TIME carried out an actual refinement of the PIMS architecture and led to a first working PIMS prototype [3], implementing and integrating all DELOS task outcomes. The ultimate goal of the On-TIME project is to provide a system that effectively supports the user while performing day-to-day tasks. This requires in particular to (i) design an effective user interface, (ii) be able to learn most common tasks, to infer what is the next task to be performed, and finally to execute (semi-automatically) tasks by linking them to user’s data.

In this paper, we present the main current achievements of the On-TIME project. In particular, in Section 2, while describing the On-TIME architecture, we provide some insights on the main features of the system. In Section 3, we illustrate the On-TIME user interface. Then, we conclude the papers by discussing the main project open issues.

2 On-TIME architecture

In this section, we describe the On-TIME architecture. Specifically, we discuss each of the main modules of the system illustrated in Fig.1. Note that two traditional relational repositories of personal data and tasks definitions are used, named *onto* and *onto_task* respectively, that are uniquely accessed through dedicated modules. The main modules of the user interface, represented in the figure as application-like windows, will be discussed in the next section.



File system watcher. This module is responsible for monitoring the file system. More precisely, it can be configured to make it detect events such as the creation, modification, renaming, or removal of files depending on their extension, and/or their location. Furthermore, it can monitor specific email clients, such as Windows Live Mail. For each event that is detected, the file system watcher sends an XML message via a TCP socket, containing both the description of the event and the element of the file system concerned by the event itself.

Monitoring system. As this module receives an XML message from the *File system watcher*, it processes the message and, possibly, uses the *Recognisers* to analyse its content. Then, depending on the event that was detected, it issues an update to the *Personal Ontology Management System*. For instance, if the event was the creation of a new file named On-TIME_Proposal.doc, then it issues the addition of an instance of the concept “File” of the Personal Ontology, with appropriate name, author and date of creation.

Recognisers. *Recognisers* are modules that detect the occurrence of complex types of data, e.g. person names, addresses, based on the use of specific regular expressions, as well as of dictionaries. Then, as a new value is detected, recognisers check whether such value already occurs in the Personal Ontology, and, if not, propose the user to add an instance of an appropriate concept, associated to the value.

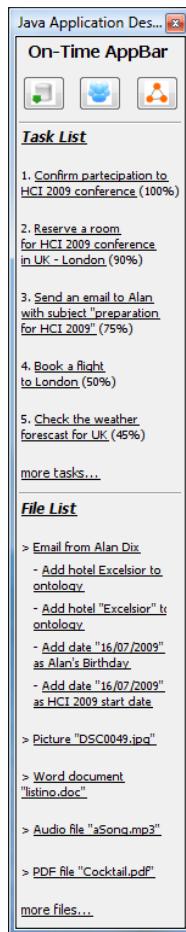
Personal ontology management system. The *Personal Ontology Management System* is responsible to maintain the Personal Ontology, and to manage the underlying reasoning services over the Personal Ontology, namely query answering and semantic update/erasure of instances. To this aim, we use the Personal Ontology designed within the DELOS task. We also use the QuOnto system as underlying reasoning engine [2].

Spreading activation module. This module processes instances of the Personal Ontology, and associates to each instance a value that represents its “contextual relevance”, also called *activation level*. This value is computed on the basis of studies of the mechanisms of the human memory [6]. For example, instances that have recently been part of the user’s activities or are related to them are flagged as “hot”. The spreading activation value associated to instances is crucial, e.g. to establish which data should be considered as input for next tasks.

Task manager. The *Task manager* maintains tasks definition, and executes tasks on the basis of the *Task inferencer* output. To this aim, a formal language with a well-defined semantics was devised, based on previous work on the topic[1], that, intuitively, allows to define tasks in terms of pre-conditions and post-conditions over the Personal Ontology.

Task inferencer. This module is responsible to infer what is the task that is more likely to be performed next, on the basis of tasks pre-conditions as well as current data activation levels. For instance, suppose that the task **Confirm participation to a conference** requires the existence of a forthcoming event related to an on-going project. Then, the *Task inferencer* will propose the user to perform all steps required to organize a mission in the occasion of a specific event, as soon as both the specific event is flagged as “hot”, e.g. because the user recently updated the details.

Task Learner. The Task learner is the module in charge of learning tasks definition by monitoring the user interaction with her desktop, taking into account the semantics of personal data handled within the interaction.



3 The On-TIME user interface

In this section, we describe the On-TIME user interface. Specifically, we discuss the three main modules of the interface, namely the *AppBar* module, the *Personal Ontology interface* and the *Task manager interface* through the presentation of a typical user interaction with the On-TIME system interface.

Suppose that Antonella works in a research lab that has recently installed a new fax. She needs to update all the documents where she used to indicate the fax number. This can be easily achieved with On-TIME, by simply updating the Personal Ontology. Hence, the starting point of her interaction with On-TIME is the *AppBar* [Fig. 2] installed on her desktop.

The *AppBar* component is the core of the system interface and the first component that appears to the user. It is a side bar (like the widget bar in Windows Vista) that can easily be accessed by the user with the mouse. This bar is divided into three sections. The first section contains a link, represented as a button, for each possible view of the *Personal Ontology Interface* (see below). The second section contains a list of tasks that according to the *Task manager*, are more likely to be performed. The third section contains the list of updates suggested by the *Monitoring system*. More precisely, if candidates instances, or instance attributes, that are not currently in the ontology, were found by the *Monitoring system* within the newly saved documents, the user would be able to update the Personal Ontology associating to those instances the right semantics, e.g., a date could be her best friend’s date of birth or an important event date.

Fig. 2 - AppBar

The *Personal Ontology Interface* is a component providing three coordinated views[6] over the Personal Ontology, having a common section that is an indented list showing the concepts hierarchy tree. While the common section allows the user to select a specific concept quickly, for example to view its instances in details, the three views have each a different purpose, and are coordinated so that when the user switches from one view to the other, she keeps focusing on the same particular instance/concept. Specifically, the *Structure View* allows the user to investigate the details of the ontology structure, the *Instance View* to select an instance of a specific ontology concept and to visualize and edit its details, and finally the *Navigation View* to browse the ontology through both its instances and concepts. Let us describe the latter in more details. It shows always two graphs, representing respectively connections among instances and concepts. One of the two graphs is always foreground and the other is visible in a frame positioned in the upper right corner of the screen, with the facility of switching between them. In order to optimize instance browsing, we use the focus plus context technique which allows the user to focus her/his attention on a specific instance or concept, by moving it on the centre of the screen, highlighting neighboring nodes (at a certain distance from the focus) and displaying the names of relations connecting them. This approach allows for maintaining the graph visualization as thinner as possible, succeeding in managing huge graphs. Moreover, the user can navigate the graph arbitrarily, and then return to the focused instance, by clicking on an anchor that is always visible.

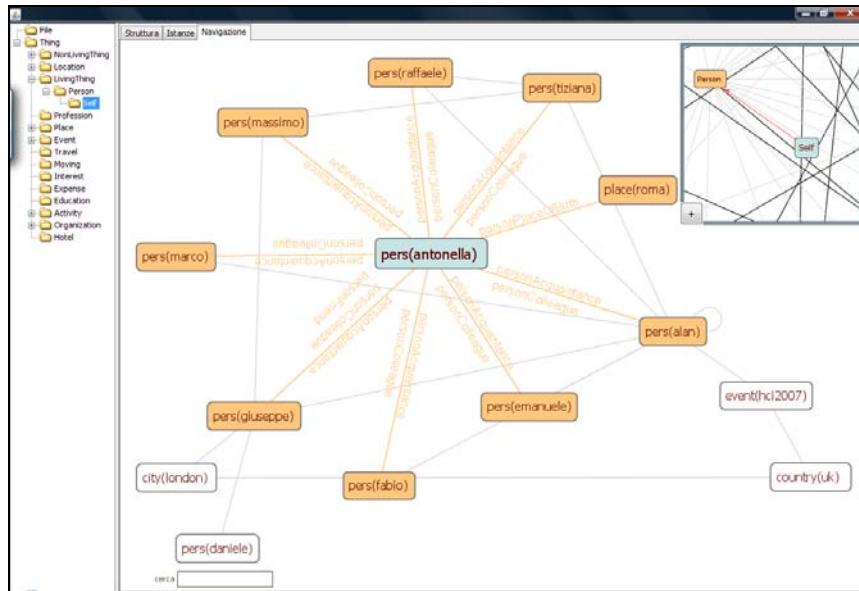


Fig. 3 - Graph view

Coming back to our scenario, by clicking on the *AppBar* appropriate button, Antonella accesses the Navigation View to browse the Personal Ontology. Doing so, the ontology graph appears foreground, centered on the instance **pers(antonella)** [Fig. 3] denoting herself, (i.e., the system owner). Then, she verifies all relations

connecting her with other ontology instances. In particular, she now switches to the Instance View, that according to the coordinated multiple views paradigm, is focused on `pers(antonella)`. Thus, Antonella immediately accesses the attribute `personFax` denoting the fax number that she is using, and she updates it. Note that, while saving the changes, the *Personal Ontology management system* ensures that the information currently contained in the Personal Ontology is not contradictory.

Suppose now that Antonella moves her attention back to the *AppBar*. She would then notice that On-TIME suggests the task **Confirm participation to HCI2009 conference**. Actually, the *Task inferencer* proposes such a task, because Antonella just received an email request of confirmation by the HCI2009 organizers, which was detected by the *Monitoring system*, and made increase the activation level of the instance `event(hci2009)` denoting the event HCI2009. Antonella takes then advantage of this smart suggestion by the system, and executes the task by clicking on the relative link in the *AppBar* which opens a new window that is part of the *Task manager interface*, i.e. the component that, by means of appropriate wizards, allows the user to execute a task, by interacting with the *Task manager*. In our scenario, the selected task just requires Antonella to confirm the reservation details, automatically returned by the *Personal Ontology management system*. Then, the *Task manager* completes the task execution by updating the ontology with the information that Antonella is going to participate to the HCI2009 conference. Specifically, this will be achieved by adding the couple of instances (`event(hci2009)`, `pers(antonella)`) to the relation `personConference` existing between the concepts Person and Events.

4 Conclusions

We presented the on-going On-TIME project. In particular, we described the main results achieved so far within the project that led to the first working prototype of a Personal Interaction Management System. Special emphasis was given to the On-TIME user interface, that, besides providing an effective access to the On-TIME system, also represents the first interface allowing the user both to browse and update an ontology, based on its semantics instead of its syntax.

Several issues need to be addressed to reach the ultimate goal of the On-TIME project. Concerning the interface, more tests need to be performed. Also, the interface should be personalized, and personalization should be as much as possible automatic making the Personal Ontology play the role of user profile. Finally, we want to investigate how to visualize more general ontologies, e.g., where an instance may belong to distinct concepts that need not to be in a hierarchy.

Another challenging feature is the design of task learning algorithms and the design of a bunch of personal ontologies, aiming at covering actual users stereotypes.

References

1. Calvanese, D., De Giacomo, G., Lembo, D., Lenzerini, M., Rosati, R.: DL-Lite: Tractable Description Logics for Ontologies. In: 20th Nat. Conf. on Artificial Intelligence AAAI'05 (2005)
2. Calvanese, D., De Giacomo, G., Lenzerini, M., Rosati, R.: Actions and Programs over Description Logic Ontologies. In: DL workshop (2007)
3. Catarci, T., Giuliano, R., Piva, M., Poggi, A., Terella, F: The On-TIME User Interface. In: CHItaly. Rome (2009)
4. Catarci, T., Dix, A., Katifori, A., Lepouras, G., Poggi, A.: Task-Centered Information Management. In: First International DELOS Conference, Pisa, Italy, February 13-14, 2007, Revised Selected Papers, C. Thanos, F. Borri and L. Candela (eds.). LNCS 4877, pp. 197-206. Springer (2007)
5. Dix A., Beale R. and Wood A. 2000. Architectures to make Simple Visualisations using Simple Systems. In Proceedings of Advanced Visual Interfaces (AVI2000), pp. 51--60. ACM Press (2000)
6. Dix, A., Katifori, A., Lepouras, G., Vassilakis, C.: Spreading Activation Over Ontologies: From Personal Context To Web Scale Reasoning. (Submitted for publication) (2009)
7. North, C. and Shneiderman, B.: A taxonomy of multiple window coordinations. Technical Report #CS-TR-3854, University of Maryland, College Park, Dept of Computer Science (1997)

Major Preservation Projects under the 6th Framework Program

Vittore Casarosa

ISTI-CNR, Pisa, Italy and HATII at University of Glasgow
casarosa@isti.cnr.it

1 Introduction

Preservation of traditional (physical) objects is (was ?) mainly a management function, whose objective is “to ensure that information survives (unaltered) in usable form for as long as it is wanted”. Preservation of digital information may be different, due to the presence of (rapidly changing) technology needed to access the information. Therefore, while traditionally preserving things meant keeping them unchanged, the ubiquitous digital environment has fundamentally changed the concepts of preservation requirements. If we hold on to digital information without modifications, accessing the information will become increasingly difficult, if not impossible.”

The problem in preservation therefore is not that of just “maintaining the bits” (which is relatively easy), but is the one of ensuring that the digital information can be used (rendered) at any time in the future.

The 6th Framework Program of the European Commission has funded three major projects dealing with the preservation of digital objects, namely DPE (Digital Preservation Europe), Planets (Preservation and Long-term Access through NETworked Services) and CASPAR (Cultural, Artistic and Scientific Knowledge for Preservation, Access and Retrieval).

DPE, which ended in March 2009, was a Coordinated Action fostering collaboration and synergies between many existing national and international initiatives across Europe. It was addressing the need to improve coordination, cooperation and consistency in practices and activities for the effective preservation of digital materials.

Planets and CASPAR are two large Integrated Projects, both addressing the preservation issues in order to provide a “complete solution” to memory institutions. Although sometimes the two projects can be seen as competing, a deeper examination of their approaches reveals that most often they are complementary. In the following we will briefly highlight those two projects.

2 PLANETS

Planets is a four-year project co-funded by the European Commission under the 6th Framework Programme (IST-033789). It started in June 2006 and will end in May 2010, with a total budget of 14 million Euro, about half of which contributed by FP6. The project's goal is to provide long-term access to digital scientific and cultural assets. Planets Partners are:

Libraries The British Library The National Library of the Netherlands Austrian National Library The Royal Library, Denmark State and University Library, Denmark	Archives The National Archives of the Netherlands The National Archives of England, Wales and the United Kingdom Swiss Federal Archives
Universities University at Cologne University of Freiburg HATII at the University of Glasgow Vienna University of Technology	Technology companies ARC Seibersdorf research GmbH IBM Netherlands B.V. Microsoft Research Limited Tessella Support Services Plc

Planets activities are focusing on five major technical areas, plus the Dissemination, Take-up and Training activities, which are customary for European projects, as sketched in Fig. 1.

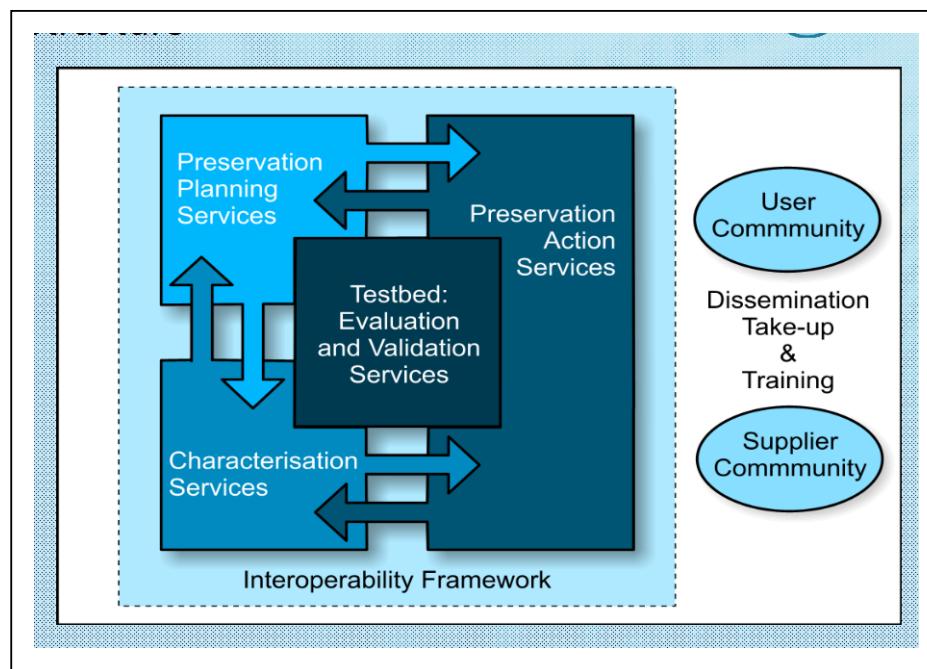


Figure 1

Preservation Planning is a set of services that enables an organisation to identify risks to its digital collection, come up with a set of alternative treatment plans to address the problem, evaluate the plans to determine the most effective one, and then execute the plan with as much automation as possible. In addition to technical information about the collection(s) to be preserved, the planning services require information also about the organisation's policies, goals, and constraints; the environments in which end-users interact with the content, and the goals that the end-users have. The preservation planner uses two key sets of Planets services: content characterisation and preservation action.

Content characterisation services identify and extract features of the content in the collection that are relevant to the planning process. Examples include basic technical features such as the pixel-depth, height, width of images, rendered features such as the number of pages in a document, and other aspects such as whether a database uses stored procedures. Planets has developed two XML based languages, namely XCEL to specify how to extract properties from a digital object, and XCDL, to describe the properties of a digital object. A tool to compare the properties of two digital objects is also available, as an aid in checking that the migration process does not change them.

Preservation action services determine what can be done. For example, one preservation action might take a specific type of digital object and convert it to a new format; another might produce an emulation environment that enables users to interact with the original digital object using the original software application; another might provide a viewing tool that provides limited interaction with a digital object, but can be readily executed on a user's portable device.

Although some preservation plans will be simple, others will comprise complex multi-step work-flows that involve extracting content from a repository, characterising it, using the results to select one or more services to treat, transform, or encapsulate the content, and then either returning the result to the repository with a detailed record of treatment, or providing a capability that can be used in a delivery environment so that end-users can get appropriate access. The Preservation actions will rely on a Registry of available tools and services, either developed by Planets or provided by different sources, such as open-source projects, commercial vendors, and third-party service providers.

The Interoperability framework enables the intimate relationship between preservation planning, preservation actions, and content characterisation as well as extensibility with a plug-in approach. It provides essential shared services such as auditing and security, as well as an extensibility mechanism, by which composite services and workflows can be defined by combining existing services available in the Registry.

The Testbed application provides a controlled environment that enables researchers and practitioners to conduct experiments and gather data on a scientific basis. They will be able to define, schedule and execute an experiment, collecting data to analyze the

result or to reproduce it later. The data collected in the testbed can provide information about the strengths, weakness, and properties of transformation tools on different types of content, can provide a basis for selecting better preservation plans, can enable the users to compare different approaches to preservation planning.

Complete information about Planets can be found at the project web site:
<http://www.planets-project.eu/>

3 GASPAR

CASPAR is 42 months project co-funded by the European Commission under the 6th Framework Programme (IST- 033572). It started in April 2006 and will end in September 2010, with a total budget of 15 million Euro, of which about 9 million Euro are contributed by FP6. The project's main goals are to implement, extend, and validate the OAIS reference model and to enhance the technology for capturing preservation related information for digital objects. Planets Partners are:

Academic HATII at the University of Glasgow, UK University of Urbino, Italy University of Leeds, UK	Research council Science and Technology Facilities Council, UK Consiglio Nazionale delle Ricerche – Institute of Information Science and Technologies, Italy Centre National de la Recherche Scientifique, France
Industrial Engineering Ingegneria Informatica S.p.A., Italy IBM Haifa Research Laboratory, Israel	International Organizations European Space Agency, ESRIN, Italy UNESCO, France
Small and Medium Enterprises Advanced Computer Systems S.p.A., Italy Asemantics S.r.l., Italy Metaware S.p.A., Italy	National Organizations Institut National de l'Audiovisuel, France Foundation for Research and Technology Hellas, Greece Institut de Recherche et Coordination Acoustique/Musique, France International Centre for Art and New Technologies, Czech Republic

The CASPAR research and development activities have been focused on the implementation of the OAIS reference model (Open Archival Information System, ISO:14721:2002), and on testing this architecture on three application domains, namely cultural data, contemporary performing arts and scientific data. An OAIS is an archive

consisting of an organization of people and systems, that has accepted the responsibility to preserve information and make it available for a Designated user community. The information being maintained has been deemed to need indefinite preservation, even if the OAIS itself is not permanent.

An overall pictorial view of the CASPAR Framework is depicted in Fig. 2.

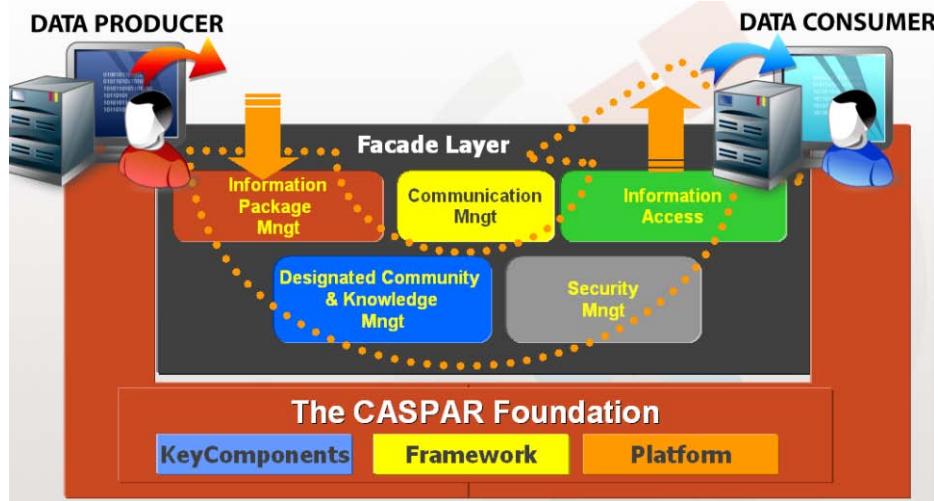


Figure 2

As shown in Figure 2, the CASPAR Foundation provides OAIS-compliant services with many features for supporting Data Producer and Consumer in their preservation activities. From an overall point of view, it is possible to group the features together in 5 main blocks:

1. Information Package Management
2. Information Access
3. Designated Community and Knowledge Management
4. Communication Management
5. Security Management.

Based on the OAIS Reference and Functional Models, CASPAR has defined the basic infrastructure for supporting those main features and providing digital preservation services, called the CASPAR Foundation, which is composed by 11 Key Components. The CASPAR Framework guarantees portability and interoperability (i.e. compliance to WS-I open standard) with existing systems and platforms.

The Key components are:

registry (REG), knowledge (KM), orchestration (POM), representation information (REPINF), preservation datastore (PDS), data access and security (DAMS), digital rights (DRM), finding aids (FIND), virtualisation (VIRT), packaging (PACK) and

authenticity (AUTH). The overall architecture of the CASPAR Foundation is depicted in Fig. 3.

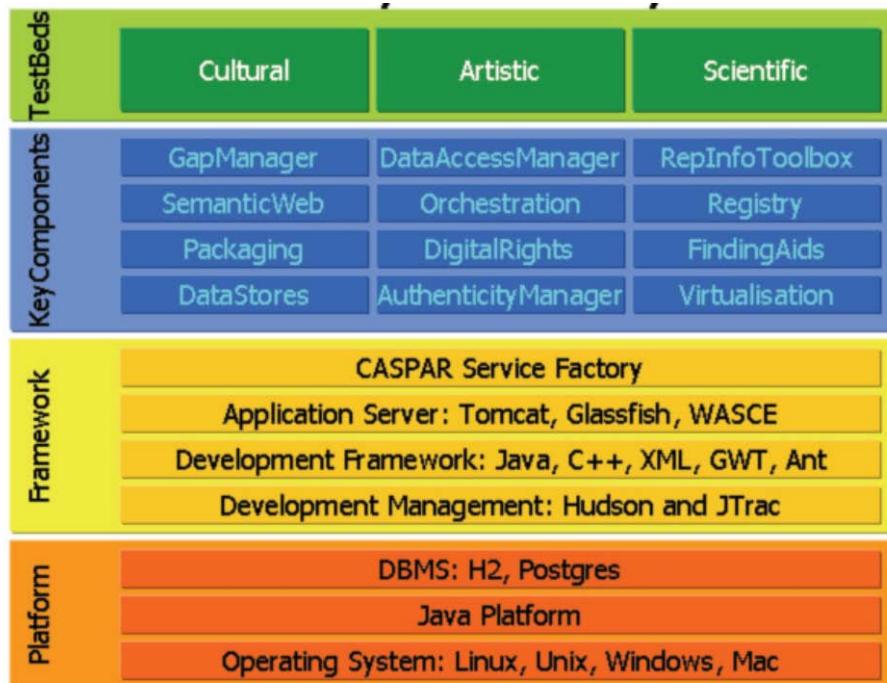


Figure 3

On top of the key components, CASPAR is testing and demonstrating preservation scenarios and strategies in order to validate its conceptual model and architectural solutions within different testbed domains: cultural, artistic and scientific.

Cultural data testbed

This testbed focuses on the preservation of all data necessary to document, visualise and model heritage sites and will provide a valuable resource to assist conservation experts in restoring the associated site while keeping its original integrity. The documentation, visualisation, and modelling of natural and cultural heritage sites is a complex task that requires large amounts of data and information. The objective of this testbed is to test the solutions developed throughout CASPAR, including virtual digital objects, spanning between processing and display.

Contemporary art testbed

This testbed is focused on the implementation for contemporary arts: contemporary music, performing arts, and other forms of technology-enhanced arts like video games. These objects also include dynamic interactive digital objects oriented towards presentation and replay. Contemporary music, as well as performing arts and video games produce very complex objects including specific hardware, instructions, and

specific equipment. These electronic models often contain highly complex extensions such as interaction devices and instructions. Interaction devices include specific sensors, instruments, and equipment while instructions often include score, software components, and audio files. The objective of this testbed is to study these specific extensions, their relationship with the generic models, procedures, and services, and to start the analysis and set-up of both a generic and specific infrastructure.

Science data testbed

This testbed is aimed at satisfying the Earth Science community requirements, by developing the necessary specific services and prototyping an Earth Observation science data preservation. It also aims at the preservation life cycle of scientific data and the preparation of specific tools that will allow visualisation and navigation of the complex metadata associated with science data complex objects. In terms of preservation, these are often very high volume, complex digital objects, oriented towards processing.

Complete information about CASPAR can be found at the project web site:

<http://www.casparpreserves.eu/>

Digital Repository Infrastructure Vision for European Research - DRIVER -¹

Sophia Jones¹, Paolo Manghi²,

¹ Greenfield Medical Library, University of Nottingham,
Nottingham, United Kingdom
sophia.jones@nottingham.ac.uk

² Istituto di Scienza e Tecnologie dell'Informazione,
Consiglio Nazionale delle Ricerche
Pisa, Italy
paolo.manghi@isti.cnr.it

Abstract. This paper describes the DRIVER European Project mission, in terms of its organizational and technological goals and results.

Keywords: repository infrastructure, digital libraries, European project, Open Access

1 Introduction

The EU-funded project DRIVER (“Digital Repository Infrastructure Vision for European Research”)² has two main aims: (*i*) setting-up a European *Confederation* for advocating and promoting EC Open Access mandates across European researchers and institutions, and (*ii*) enabling a *technical infrastructure* of European Institutional Repositories aggregating and making accessible Open Access content throughout Europe.

¹ This paper was originally published as an article in BELIEF Magazine 2009

² www.driver-community.eu

1.1 The DRIVER Confederation

The DRIVER confederation is an organization of people working to achieve a formal establishment of an (European) digital repository community. The Confederation mirrors strategic alliances that move towards a global, interoperable, trusted, long-term repository infrastructure for which DRIVER has built the nucleus in Europe. It aims to encourage a combined effort of repository development between federations within a network of content providers. The Confederation partners represent European and international repository communities, subject based communities, repository system providers, service providers, as well as political, research, and funding organisations who share the DRIVER vision to allow all research institutions in Europe and worldwide to make all their research publications openly accessible through institutional repositories. DRIVER liaises with institutions and initiatives from the majority of European countries, the U.S., Canada, Latin America, China, Japan, India and Africa.

Closely related to the theme of interoperability and specifically in relation to this project, are the DRIVER Guidelines, as they create a common ground by achieving interoperability on two layers: (i) syntactical (use of OAI-PMH and OAI_DC), and (ii) semantic (use of vocabularies). The data in the technical infrastructure is based on locally hosted resources that are collected in digital repositories and harvested and aggregated by DRIVER. In order to ensure a high quality of aggregation, the DRIVER Guidelines have been developed to make it possible to harmonise and validate the data. DRIVER makes its aggregated data available for re-use via OAI-PMH to all partners in the DRIVER network of content providers whilst respecting the provenance of resources by “branding” them with information of the local repository. The DRIVER Guidelines provide orientation for managers of new repositories to define their local data-management policies, for managers of existing repositories to take steps towards improved services and for developers of repository platforms to add supportive functionalities in future versions. By following the Guidelines repositories can become part of the DRIVER network and can re-use DRIVER data for the development of local services. In essence then, the DRIVER Guidelines assist repository managers to make their material more widely available. Interoperability in the sense

of the DRIVER Guidelines means standardised metadata of the harvested records, based on the use of standards.

1.2 The DRIVER Technical Infrastructure

An important outcome of the DRIVER project is *D-Net*,³ a software toolkit capable of enabling a running environment where data and service providers can (*i*) find the tools to aggregate heterogeneous OAI-PMH compliant Institutional Repositories into uniform shared Information Spaces, and (*ii*) dynamically build, modify and customize their Digital Library (DL) applications to operate over such Spaces. The technology supports a Service-oriented e-infrastructure, where distributed and shared resources are implemented as standard Web Services and applications consist of sets of interacting services. The current implementation of D-Net offers functionality services required to build distributed aggregation systems and DL end-user applications. Aggregation systems enable the construction of uniform Information Spaces of metadata records to be harvested from possibly heterogeneous Institutional Repositories. Important in this context are Store Services, Index Services and Aggregation Services, which offer advanced tools for OAI-PMH harvesting, cleaning and integrating metadata records according to target metadata record formats. The resulting Information Spaces can then be accessed via an arbitrary number of DL applications built by service providers by combining D-Net services such as Recommendation, Collection, Browsing, User Interfaces and others.

Key features of D-Net are the *scalability* and the *openness* of its infrastructure instances. Specifically, an instance of the DRIVER infrastructure can scale up to arbitrary numbers of service instances, applications and organizations while the underlying application framework is open to the introduction of services providing new functionality so as to extend the D-Net toolkit.

Since July 2008, the DRIVER project maintains a running instance of D-Net which hosts one main aggregation system, integrating Open

³ http://www.driver-repository.eu/D-NET_release: developed at Istituto di Scienza e Tecnologie dell'Informazione, CNR (Pisa, IT), ICM (Warsaw, PL), Department of Informatics and Telecommunications of the University of Athens (Athens, GR) and University of Bielefeld Library (Bielefeld, DE).

Access metadata records from a growing number of European Institutional Repositories. At present, the infrastructure runs 36 services distributed over 9 partner sites; as a result of the Confederation efforts, the resulting Information Space numbers 1,000,000+ records out of 200+ repositories across 27 countries, and the number of repositories willing to join is still growing. Currently, the space is accessed by three DL applications: the Belgium national repository portal, offering search over the Belgium Repository Federation subset; Recolecta national repository portal, offering search on the Spanish Repository Federation subset; and the main DRIVER portal, providing access and advanced functionality over the whole space.

Europeana: Towards The European Digital Library

Nicola Aloia, Cesare Concordia, Carlo Meghini

Institute of Information Science and Technologies (ISTI)
Area della ricerca CNR, via G. Moruzzi 1, 56124 PISA, Italy
{Nicola.Aloia, Cesare.Concordia, Carlo.Meghini}@isti.cnr.it

Abstract. This paper briefly describes the process that will lead to Europeana, the European Digital Library. This process is currently running, so that it is possible to give only an account of its inception, involved actors and projects, and current status. The paper concludes by quickly outlining the role that CNR has in the making of Europeana.

1 Europeana: the inception

In October 2004 Google launched Book Search a tool that searches the full text of books. The service was formerly known as Google Print when it was introduced at the Frankfurt Book Fair. In reaction to that, in 2005 Jacques Chirac called for a European Digital Library in order to affirm the ‘cultural identity’ of Europe and to spread its heritage.

In 2006, The European Commission takes on board the duty of creating the European Information space (i2010) and in particular elaborates a plan to bring on-line the European Culture on 24 August 2006, the European Commission presented its vision about the European Digital Library:

A common multilingual access point would make it possible to search Europe's distributed – that is to say, held in different places by different organisations – digital cultural heritage online. Such an access point would increase its visibility and underline common features. The access point should build on existing initiatives such as The European Library (TEL), in which Europe's libraries already cooperate. It should where possible closely associate private holders of rights in cultural material and all interested stakeholders. A strong commitment by the Member states and cultural institutions to arrive at such an access point should be encouraged [2].

2 Building Europeana: actors

Several actors are involved in building the Europeana Digital Library: stakeholders, experts in the various scientific/technological areas, system developers, business developers, and managers. Stakeholders are the institutions who provide content, such as libraries, archives, museums, audiovisual archives, and others; they have two main goals: to provide metadata to the Europeana Digital Library about their artifacts, and to help in defining standards for interoperability. Scientific actors have as main goal to define the interoperability framework, and to specify the functionality of Europeana. System developers have the task of designing and implementing the Digital Library System (DLS) [1], based on the functional specifications. Business developers have the task of ensuring the long-term sustainability of Europeana.

3 Building Europeana: the projects cluster

The Europeana Digital Library will be the result of a number of projects run by different cultural heritage institutions; amongst these projects, we have:

- Athena, an aggregator that helps museums bringing their content to Europeana.
- APENet, a Best Practice Network whose objective is to build an Internet Gateway for Documents and Archives in Europe.
- EuropeanaLocal, aiming at improving the interoperability of the digital content held by regional and local institutions.
- European Film Gateway, whose goal is to find solutions for providing integrated access to the Europe's cinematographic heritage.
- Judeica, whose aim is to establish a comprehensive map of European Jewish Cultural Heritage.
- EuropeanaConnect, that will provide the technologies and resources to semantically enrich the digital content in Europeana.
- Europeana V1.0 that will implement the technological platform.

All projects are part-funded by the European Commission's eContentplus program. The full implementation of the Europeana Digital Library System is the goal of the two "core" projects: EuropeanaConnect and Europeana V1.0. The complete list of the projects contributing to Europeana is at the URL: <http://group.europeana.eu/>.

4 Building Europeana: the EDLNet

The EDLnet Thematic Network is an eContentPlus program recently concluded, that has prepared the ground for the European Digital Library. The project main topic was to improve the cross-domain accessibility to cultural content, a pillar of the European Commission's i2010 Digital Libraries initiative. EDLnet Thematic Network has brought on board the key European stakeholders to build consensus on creating the European Digital Library. EDLnet project main results are:

- The creation of a large visible community of archivists, librarians and museum people committed to making content available in an interoperable way.
- The production of clear and usable summary reports and recommendations on each of the main areas of interoperability addressed in the EDLnet.
- The definition of a roadmap showing how the component parts interlink and what needs to be achieved when to realise the aim of the European Digital Library.
- The implementation of a fully working prototype, with interoperable multilingual access, covering over 4,5 million digital items.
- The definition of a proposal for funding to create a fully operational European Digital Library service.

As of April 09 the Europeana prototype (www.europeana.eu) contains data from 54 cultural institutions from 24 countries.

5 The System View

In the wide public, Europeana is primarily perceived as a Portal exposing a great amount of cultural heritage information. Even though this perception is not entirely misleading, one of the goals of Europeana is to build an open services platform enabling users and cultural institutions to create and access a large collection of objects representing digital and digitised content via an Application Program Interface (API). Europeana API will enable cultural institutions and users to:

- Access Europeana content
- Provide content to Europeana
- Build applications using Europeana functionalities for their own use.
- Use Europeana services for their own Digital libraries

From this more general point of view, the Europeana Portal is just a component of the system, more specifically it can be viewed as a web application using the Europeana API to offer services, in particular discovery-based access to, the Europeana Digital Library.

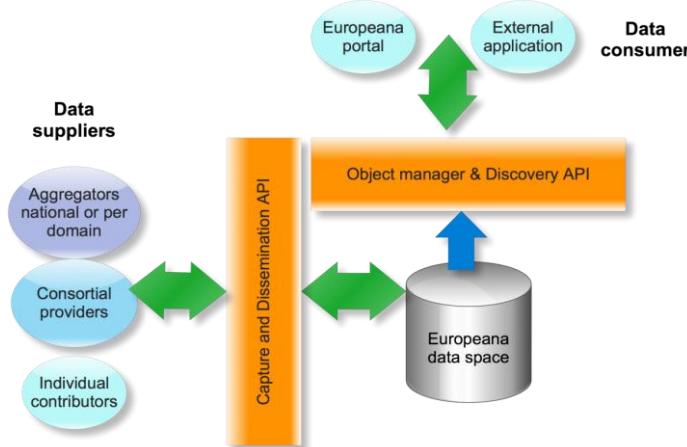


Fig.1 the data flow in Europeana

6 ISTI and Europeana

The CNR-ISTI has been invited to join the EDLnet in July 2007 as a bridge between the DELOS community and Europeana, and also based on the significant previous experience of the Institute in digital library projects. In EDLNet, CNR-ISTI had a key role in the definition of the functional architecture of Europeana [3], in the design and development of the advanced search functionalities of the Europeana prototype and in the implementation of the data ingestion functionality for populating the Europeana data space. Moreover the development tools and test servers have been hosted on ISTI computers. The ISTI-CNR is currently a partner of the Europeana V1.0 project, the main tasks in which are involved are: the definition of a data model for Europeana, the implementation of the search functionalities of the DLS and the set-up and management of EuropeanaLabs, which will play to Europeana the role that GoogleLabs plays to Googleland; as it has been for EDLNet the ISTI will provide technical support to the development infrastructure by hosting test servers and tools for distributed team management.

References

1. Candela L., et al: Setting the Foundations of Digital Libraries. The DELOS Manifesto. D-Lib Magazine, March/April 2007, Volume 13 Number 3/4.
2. Commission Recommendation of 24 August 2006 on the digitisation and online accessibility of cultural material and digital preservation (2006/585/EC)
3. Dekkers M., Gradmann S., Meghini C.: D2.5 Europeana Outline Functional Specification, March 2009

TrebleCLEF: Evaluation, Best Practices and Collaboration for Multilingual Information Access

Carol Peters

Istituto di Scienza e Tecnologie dell'Informazione (ISTI-CNR), Pisa, Italy
carol.peters@isti.cnr.it

Abstract. The objectives of the TrebleCLEF Coordination Action for multilingual information access are outlined, the results obtained in the first twelve months are presented, and the activities planned for 2009 are described.

1 Introduction

The popularity of Internet and the consequent global availability of networked information sources and digital libraries has led to a strong demand for multilingual access and communication technologies. Such technologies should support the timely and cost-effective provision of knowledge-intensive services for all members of linguistically and culturally diverse communities. This is particularly true in the multilingual setting of Europe. However, despite recent research advances, there are still very few operational systems in existence, and these are limited to the most widely used languages. The challenge to be faced is how to best transfer the research results to the wider market place. The objective of the two-year TrebleCLEF project is to build on and extend the results already achieved by the existing Cross Language Evaluation Forum. The aim is not only to support the development and consolidation of expertise in the multidisciplinary research area of multilingual information access (MLIA) but also to disseminate this knowhow to the application communities.

Treble-CLEF thus has three main goals:

1. **Organisation of Annual Evaluation Campaigns and Analysis of Results:**
Promoting the academic excellence represented by the Cross Language Evaluation Forum (CLEF) by coordinating research aimed at finding solutions for key problems in the multilingual information access area and providing an infrastructure to evaluate the results and to conduct in-depth analyses.
2. **Technology Transfer and Best Practices:** Launching a concerted action of technology transfer and dissemination of knowhow, tools, resources and best practice guidelines aimed at system developers and application communities.
3. **Dissemination, Community-Building and Training:** Encouraging community building and collaboration through the provision of a discussion forum, by making the scientific data and results produced during the evaluation campaigns publicly available, and by organizing workshops, and training activities.

TrebleCLEF began in January 2008. In the following sections we report on the results obtained during the first twelve months and the activities planned for the second year.

2 Evaluation Campaign

The CLEF 2008 evaluation campaign was a great success with a large number of participants. There were seven main evaluation tracks plus two pilot tasks. The aim, as usual, was to test the performance of a wide range of MLIA systems or system components. Bearing in mind the objectives of TrebleCLEF, in addition to research-oriented tasks, such as the new VideoCLEF track, tasks were organised (i) to meet the needs of a specific application community, e.g. digital libraries with the TEL task using the data of The European Library, the medical image processing sector with ImageCLEFmed activities, and (ii) to examine user behaviour, e.g. the interactive iCLEF track which used a multilingual search game, based on the Flickr database of images, to study user interaction in this context via questionnaires and log analyses. 100 groups, mainly but not only from academia, participated in the campaign. Most of the groups were from Europe but there was also a good contingent from North America and Asia plus a few participants from South America and Africa. In 2008, the evaluation infrastructure was considerably strengthened with the further design and development of the DIRECT system, providing participants with new means of access to the test collections and the campaign results [1]. Full details regarding the design of the tracks, the methodologies used for evaluation, and the results obtained can be found in the on-line Working Notes [2] or the post-campaign Proceedings [3].

The results of the campaign were presented at a two-and-a-half day workshop held in Aarhus, Denmark, 17-19 September 2008 and attended by 150 researchers and system developers. The annual workshop, held in conjunction with the European Conference on Digital Libraries, plays an important role by providing the opportunity for all the groups that have participated in the evaluation campaign to get together comparing approaches and exchanging ideas.

The focus on the user, on log analysis, and on the needs of application communities is being increased in 2009. In line with the TrebleCLEF philosophy, CLEF2009 includes three new tracks focused on analysing user behaviour in a multilingual context (LogCLEF), on studying the requirements of multilingual patent search (CLEF-IP), and on improving our understanding of MLIA systems and their behaviour with respect to languages (GridCLEF) [4].

2 Best Practices and Guidelines

Three best practices white papers will be produced by the project. The first focuses on recommendations for the language resources needed in the development of MLIA systems. A preliminary version of this report was released at the end of 2008 and a final version in Spring 2009 [5]. It provides a survey of the state-of-the-art and presents an assessment of the priority requirements in this sector resulting from consultations with system developers, language industry and communication players. The first part of the report focuses on the description of a generic MLIA system, describing the required modules (morphological analyzers, part-of-speech taggers, syntactic parsers, cross-language components, etc.) An important part of the deliverable is dedicated to a survey of existing language resources that are used and/or are useable for MLIA. On

the basis of a series of inventories and community feedbacks, the report proposes an action plan for developing a set of Language Resources for all technologies related to MLIA. This minimal set of resources and basic tools/modules constitutes our contribution to the setting up of a Basic LAnguage Resource Kit (BLARK) for MLIA.

The survey also focused on new trends to improve the existing resources for MLIA, both in terms of quality and language coverage. A section is devoted to the description of needs as expressed by key developers. Many developers are working on exploiting the combination of various modalities (e.g. text and image, audio-visual documents, etc.) as well as deploying MLIA for specific domains (e.g. patents). The languages of interest to most respondents were identified: the major European languages are the most cited but less-used European languages, Chinese, Japanese, Arabic, Russian and several others are also named.

The other two white papers will regard "Best Practices in System and User-oriented MLIA" and "Best Practices for Test Collection Creation". In order to obtain input for the first one, which will be released in June 2009, two workshops were organised on the needs of MLIA system developers and MLIA users in 2008 [6,7]. Issues raised at the system developers workshop included the realism of the evaluation resources with respect to use cases and scenarios, performance measures, verifiability and robustness of results, presentation of multilingual results and user-oriented evaluation, and the transferability of results to operational settings. At the User Communities workshop consensus emerged around (i) features that any MLIA system should have from a user's perspective, and (ii) possible strategies to bring MLIA research closer to real world needs and to effectively transfer MLIA technologies to society. The white paper will aim at providing recommendations for successful MLIA implementations, generalizing as much as possible from the mostly academic papers and experiences as there are few "packaged" cross-language offerings in the marketplace. It is hoped that the recommendations will be found to be useful for commercial developers/implementers of complex search applications, who intend to build their own components tailored to their specific applications.

A TrebleCLEF Workshop focused on novel methodologies for evaluation in information retrieval held at ECIR (European Conference on Information Retrieval) on March 30, 2008 in Glasgow, United Kingdom [8] and a first draft survey of test collection evaluation practises in information retrieval have produced input for the Best Practices for Test Collection Creation, Evaluation Methodologies and Language Processing Technologies deliverable due for release in October 2009.

In addition to completing the best practice guidelines, TrebleCLEF will also build a portal with pointers to relevant resources and tools to aid implementers and developers in adopting the recommendations. We hope to influence future academic research, addressing the gaps which have been identified during these studies.

3 Dissemination and Training

Dissemination activities in this first year have mainly concentrated on the scientific community with the organisation of the workshops mentioned above, a good number of presentations at conferences and invited talks, and a large number of publications. In

2009 the dissemination training activities will be directed more consistently towards the application communities with the presentation of the Best Practice Guidelines and at the MLIA Technology Transfer Day, planned for December 2009. A major training activity will be the TrebleCLEF Summer School to be held in June 2009. The aim of the Summer School is to give participants a grounding in the core topics that constitute the multidisciplinary area of Multilingual Information Access. Both theoretical and practical issues are addressed. The focus of the school is on "How to build effective multilingual information retrieval systems and how to evaluate them" [9].

4 Conclusions

It is still too early to assess to what extent TrebleCLEF will be able to realise all its objectives. While the continually increasing impact of the evaluation campaign activity is demonstrated by the rise in numbers of participants and the increase in complexity of the tasks, this is still mainly concentrated in the academic circles. It is more difficult to breakthrough and influence industry. However, the response to our workshops and surveys has been encouraging and from the input received it is clear that the need for guidelines on how to put MLIA theory into commercial practice is becoming a more widely acknowledged reality.

References

1. Ferro, N. Operational Scientific Digital Library. TrebleCLEF D2.2 <http://www.trebleclef.eu/publications.php>
2. http://clef-campaign.org/2008/working_notes/CLEF2008WN-Contents.html
3. Peters, C., Deselaers, T., Ferro, N., Gonzalo, J., Jones, G.J.F., Kurimo, M., Mandle, T., Peñas, A., Petras, V. Evaluating Systems for Multilingual and Multimodal Information Access, 9th Workshop of the Cross-Language Evaluation Forum, Aarhus, Denmark, September 17-19, 2008, Springer LNCS 5706, in print.
4. <http://www.clef-campaign.org/2009.html>
5. Moreau, N. Best Practices in language Resources for MLIA. TrebleCLEF D5.2 <http://www.clef-campaign.org/>
6. Braschler, M., Clough, P.D. Bringing Multilingual Information Access to Operational Systems. TrebleCLEF System Developers Workshop. D3.3 <http://www.trebleclef.eu/publications.php>
7. Gonzalo, J., Peñas, A., Verdejo, F., Peters, C. Workshop on Best Practices for the Development of Multilingual Information Access Systems: the User Perspective TrebleCLEF User Communities Workshop D3.2 <http://www.trebleclef.eu/publications.php>
8. Sanderson, M. Workshop on Novel Methodologies for Evaluation in Information Retrieval. In SIGIR Forum, June 2009, 43(1), pp59-62. <http://www.sigir.org/forum/2009J/2009j-sigirforum-sanderson.pdf>
9. <http://www.trebleclef.eu/summerschool.php>

MultiMatch: Multilingual / Multimedia Access to Cultural Heritage

Giuseppe Amato, Franca Debole, Carol Peters, Pasquale Savino

Institute of Information Science and Technologies - CNR
Pisa, Italy
{Franca.Debole}@isti.cnr.it

Abstract. Our shared cultural heritage (CH) is an essential part of our European identity, transcending cultural and language barriers. The aim of the MultiMatch project is to enable users to explore and interact with online internet-accessible CH content, across media types and language boundaries, in ways that do justice to the multitude of existing perspectives. This has been achieved through the development of a search engine targeted for the access, organisation and personalized presentation of cultural heritage information. MultiMatch aims at complex, heterogeneous digital object retrieval and presentation.

1 INTRODUCTION

Online Cultural Heritage (CH) content is being produced in many countries by organisations such as national libraries, museums, galleries and audiovisual archives. Additionally, there are increasing amounts of CH relevant content available more generally on the World Wide Web. While some of this material concerns national or regional content only of local interest, much material relates to items involving multiple nations and languages, for example concerning events in Europe or Asia. In order to gain a full understanding of such events, including details contained in different collections and exploring different cultural perspectives requires effective multilingual search technologies. The EU FP6 MultiMatch project is concerned with information access for multimedia and multilingual content for a range of European languages. MultiMatch tries to offer “complex object retrieval” through a combination of focused crawling, and semantic enrichment that exploits the vast amounts of metadata available in the cultural heritage domain. The MultiMatch search engine was developed with specialised search facilities for multilingual access to cultural heritage material in diverse media. The aim was to present the user with the “detailed picture” of complex CH objects. The overall goal of the project was to build a fully operational system prototype, designed and refined according to the requirements of diverse user classes. Users can search information using their preferred language, searching for all types of digital objects, accessing only the sites that contain information potentially relevant to their request, retrieving mainly relevant items, and viewing the query results in an organized structured fashion. Standard and

ontology-based descriptions of content are used, thus providing an interoperable semantic framework for intelligent multimedia object delivery. Metadata automatically extracted from CH material was mapped onto this framework. The system has been demonstrated for the main languages of the cultural heritage institutions in the consortium: Dutch, Italian, Spanish, English and also German and Polish, but it was extendible to other languages.

2 MOTIVATION

Europe's vast collections of unique and exciting cultural content are an important asset of our society. On the web, cultural heritage (CH) content is everywhere, in traditional environments such as libraries, museums, galleries and audiovisual archives, but also reviews in popular magazines and newspapers, in multiple languages and multiple media. CH objects on the web are no longer isolated objects, but situated, richly connected entities, equipped with very heterogeneous metadata, and with information from a broad spectrum of sources, some with authoritative views and some with highly personal views. The aim of the MultiMatch project is to enable users to explore and interact with online internet-accessible CH content, across media types and language boundaries, in ways that do justice to the multitude of existing perspectives. This has been achieved through the development of a search engine targeted for the access, organisation and personalized presentation of cultural heritage information. The main source of information stored in the MultiMatch prototype system is composed of cultural heritage objects obtained through crawling and indexing of material obtained from cultural heritage sites, web encyclopedias (e.g. Wikipedia), digital libraries of specific cultural heritage organizations, OAI compliant digital resources, and RSS feeds from cultural web sites. The cultural heritage search and navigation facilities envisaged by MultiMatch cater for these information needs by presenting users with a composite picture of complex CH objects. For instance, in reply to a user's request for information on Van Gogh, the MultiMatch engine can present information on Van Gogh from multiple museums around Europe, in multiple languages; it could complement this with pointers to Van Gogh's contemporaries, with links to exhibitions on Van Gogh, to reviews of these exhibitions, to blog entries by visitors to these exhibitions, and to background information taken from online resources or dedicated sites. The MultiMatch search engine has been developed with specialised search facilities for multilingual access to cultural heritage material in diverse media.

3 THE SYSTEM

The MultiMatch search engine is able to:

- identify relevant material via an in-depth crawling of selected CH institutions, accepting and processing any semantic web encoding of the information retrieved;

- crawl the Internet to identify websites with CH information, locating relevant texts, images and videos, regardless of the source and target languages used to write the query and/or describe the results;
- automatically classify the results on the basis of a document's content, its metadata, its context, and on the occurrence of relevant CH concepts;
- automatically extract relevant information which will then be used to create cross-links between related material, such as the biography of an artist, exhibitions of his/her work, critical analysis, etc.;
- organise and further analyse the material crawled to serve focused queries generated from information needs formulated by the user;
- interact with the user to obtain a more specific definition of initial information requirements;
- the search results are organized in an integrated, user-friendly manner, allowing users to access and exploit the information retrieved regardless of language barriers.

The MultiMatch search engine enables the user to retrieve cultural objects through different modalities:

1. The simplest one is a traditional free text search. This search mode is similar to that provided by general purpose search engines, such as Google, but MultiMatch provides more precise results and with support for multilingual searches (English, Italian, Spanish, Dutch, German, and Polish). Multilingual searches are performed either through machine translation or by using a general purpose dictionary extended to include terms which are CH specific.
2. Multimedia searches, based on similarity matching and on automatic information extraction techniques.
3. Metadata based searches.
4. A browsing capability allows users to navigate the MultiMatch collection using, among others, a web directory-like structure based on the MultiMatch ontology.

Searches can be made at three main levels of interaction: (a) Default search mode, (b) Specialized search mode, (c) Composite search mode. The *default* search level is provided for generic users. In this way, given a general query, MultiMatch retrieves all the cultural objects, web pages and multimedia content that best suit the query. Merging, ranking and classification of these results are also performed. Users with a more precise knowledge of system functionality, and with specific search needs, may use one of the specialized interaction levels available. These allow the user to query specific search services. In this way, MultiMatch includes standalone image, video and metadata-based searches, each with its own search fields, display and refinement options. It also includes a set of browsing capabilities to explore MultiMatch content. The “composite search mode” supports queries where multiple elements can be combined. For example, it is possible to search using the metadata fields associated with each document, but combining this restriction with free text and/or image similarity searches. In the following subsections we describe the MultiMatch approach to support the specialized search:

- creators search. The general idea of this specialized search level is that, for a given type of cultural entity (creator/creation), the user can query the MultiMatch system to retrieve all the information available about it (e.g. the user can query about Van Gogh and then retrieve all the information available about the painter). MultiMatch creates relations between cultural objects to allow the user to browse and discover information related with his current search.
- audiovisual search. This type of information can be considered as multimodal, which implies that pure visual contents (images and videos) are also related with spoken contents, associated metadata, and texts describing the contents. The MultiMatch system provides three different specialized searches on multimedia contents:
 - image search. MultiMatch offers the possibility of retrieving still images and video keyframes based on text and image queries using multimodal searching.
 - video search. MultiMatch offers the possibility to search for video contents using text queries and also image queries.
 - audio search. Users is able to perform audio search to retrieve audio documents and also video documents by way of their speech tracks. An index built from these transcripts makes audio search possible. The user submits a free text query. Audiovisual documents relevant to this query (i.e. containing the query in the speech recognition transcript) is shown, allowing the user to start playing the audio or video document before the occurrence of the first query word.

4 CONCLUSION

The project was completed on October 31st, 2008 and all planned objectives were achieved. The MultiMatch project involved 11 partners: Istituto di Scienza e Tecnologie dell'Informazione - Consiglio Nazionale delle Ricerche (ISTI-CNR), University of Sheffield, Dublin City University, University of Amsterdam, University of Geneva, Universidad Nacional de Educacion a Distancia, Fratelli Alianari Istituto Edizioni Artistiche SpA, Netherland Institute for Sound and Vision, Biblioteca Virtual Miguel de Cervantes, OCLC PICA, WIND Telecomunicazioni SpA. The project website is active (<http://www.multimatch.org>) and an online demo of the MultiMatch prototype can be accessed by registered users (free registration available).

Acknowledgement

Work partially supported by European Community under the Information Society Technologies (IST) programme of the 6th FP for RTD - project Multi-MATCH contract IST- 033104. The authors are solely responsible for the content of this paper. It does not represent the opinion of the European Community, and the European Community is not responsible for any use that might be made of data appearing therein.

D4Science: an e-Infrastructure for Supporting Virtual Research Environments

Leonardo Candela, Donatella Castelli, and Pasquale Pagano

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

Abstract. e-Infrastructure is the term coined for innovative research environments that provide modern scientists with seamless access to shared, distributed and heterogeneous resources. *Virtual Research Environments* (VREs) are applications whose constituents are dynamically borrowed from the e-Infrastructure, bound (and deployed) instantly, just at the time and for the period they are needed. In this paper we describe D4Science, an e-Infrastructure supporting the design and deployment of VREs.

1 Introduction

Today research activities require collaborations among parties that are widely dispersed and autonomous. Collaborations are often cross-discipline and require innovative research environments that make available data, processing and interaction intensive workflows to produce new knowledge able to stimulate further research.

To support such a demanding scenario a very promising approach is based on *e-Infrastructures*. By definition, an e-Infrastructure is a framework enabling secure, cost-effective and on-demand *resource sharing* [1] across organisation boundaries. A resource is here intended as a generic entity, physical (e.g. storage and computing resources) or digital (e.g. software, processes, data), that can be shared and interact with other resources to synergistically provide some functions serving its clients, either human or inanimate. Thus, an e-Infrastructure poses as a “*mediator*” in a market of resources having the role to accommodate the needs of resource providers and consumers. The infrastructure layer gives support to: (*i*) resource providers, in “*selling*” their resources through it; (*ii*) resource consumers, in “*buying*” and orchestrating such resources to build their applications. Further, it provides organizations with logistic and technical aids for application building, maintenance, and monitoring. A well-known instance of such an e-Infrastructure is represented by the Grid [2], where a service-based paradigm is adopted to share and reuse low-level physical resources. Application-specific e-Infrastructures are in their turn inspired by the generic e-Infrastructure framework and bring this vision into specific application domains by enriching the infrastructural *resource model* with specific *service* resources, i.e. software units that deliver functionality or content by exploiting available physical resources.

This potentially unlimited market of resources allows a new development paradigm that becomes reality through *Virtual Research Environments*, i.e. integrated environments providing seamless access to resources and facilities for communication, collaboration and any kind of interaction among scientists and researchers. A Virtual Research

Environment is built by aggregating the needed constituents after hiring them through the e-Infrastructure and can be considered as an organised ‘view’ built atop the pool of available assets, ranging from computers and servers to collections and services.

This paper presents the realisation of these two very challenging approaches in the context of the D4Science EU project.

2 D4Science Overview

D4Science¹ (DIstributed colLaboratories Infrastructure on Grid ENabled Technology 4 Science - Jan 2008-Dec 2009) is a project co-funded by European Commission’s Seventh Framework Programme for Research and Technological Development involving 11 participating organizations. It continues the path that GÉANT², EGEE³ and DILIGENT [3] projects have initiated towards establishing networked, grid-based, and data-centric e-Infrastructures that accelerate multidisciplinary research by overcoming barriers such as those related to heterogeneity, sustainability and scalability.

In particular, D4Science is currently operating an infrastructure consisting of heterogeneous resources ranging from *hardware resources*, i.e. machines acting as computing and storage resources providers (in part borrowed from the EGEE infrastructure) or hosting environment supporting dynamic software deployment, to *software resources*, i.e. software packages implementing specific functions, *services*, i.e. running instances of software resources providing functions, and *data resources*, i.e. collection of compound information objects representing various kind of information.

This infrastructure is currently supporting the operation of two very large and challenging scientific communities: Environmental Monitoring and Fisheries and Aquaculture Resources Management. These scientific communities are served through three virtual organizations (VOs): Environmental Monitoring VO, Fishery Country Profiles Production System VO and Integrated Capture Information System VO. These VOs are dynamic group of individuals and/or institutions defined around a set of sharing rules in which resource providers and consumers specify clearly what is shared, who is allowed to share, and the conditions under which sharing occurs to serve the needs of a specific community. These VOs consists of various resources including collection of Earth images, satellite products, species distribution maps, reports, statistical data, and tools for processing and analyzing them.

The development and operation of the D4Science infrastructure is supported by the gCube software system [4]. gCube is a distributed system for the operation of large-scale scientific infrastructures. It has been designed from the ground up to support the full life-cycle of modern scientific enquiry, with particular emphasis on application-level requirements of information and knowledge management. To this end, it interfaces pan-European Grid middleware (gLite⁴) for shared access to high-end computational and storage resources, but complements it with a rich array of services that collate, describe, annotate, merge, transform, index, search, and present information for a variety of multidisciplinary and international communities. Services, information, and machines are infrastructural resources that communities select, share, and consume in the scope of collaborative Virtual Research Environments.

¹ <http://www.d4science.eu>

² <http://www.geant.net>

³ <http://public.eu-egee.org/>

⁴ <http://glite.web.cern.ch/glite/>

3 Building and Operating Virtual Research Environments

The D4Science e-Infrastructure supports the creation and management of VREs by offering mechanisms for the VRE definition, deployment and operation [5].

The definition process is organized in steps each allowing the designer to characterize different aspects of the expected VRE. These steps enable to specify the features of the expected environment and allow the D4Science enabling system to put in operation it by automatically allocating and deploying the most appropriate resources selected between the ones available in the VO. The identification of these steps and the dependences between them have been strongly influenced by the digital library model presented in the DELOS Digital Library Reference Model [6]. In particular, these steps aim at capturing VRE constituent elements belonging the Content, Functionality, Users, and Architecture dimensions. Once the specification is completed, the VRE generation logic implemented by the D4Science infrastructure analyses it and derives an optimal deployment plan aiming at maximizing existing resources usage and eventually including dynamic resource generation. The infrastructure guarantees an optimal consumption of the available resources by selecting the minimal amount of them sufficient to meet its established performance and robustness criteria. By using these mechanisms four VREs have been created serving very different application domains:

Fishery Country Profiles Production System (FCPPS) supports scientists in the generation of fisheries and aquaculture reports. The production of country profiles requires complex aggregation and editing of continuously evolving multi-lingual data from a large number of heterogeneous data sources. Availability of the FCPPS VRE permits to the scientists producing them to update and web-publish these vital reports as frequently as the community requires, while also having access to additional resources when needed.

Integrated Capture Information System (ICIS) supports scientists in integrating regional and global capture and distribution information of aquatic species, from a number of Regional Fishery Management Organisations and international organisations (FAO, WorldFish Center) into a common system. The VRE provides not only access to the necessary data. Rather it organises a number of services for providing an harmonised view of catch statistics and allowing the community to overlay them according to pre-defined reallocation rules.

Global Ocean Chlorophyll Monitoring (GCM) offers to scientists an environment that integrates satellite data of microscopic marine plants and sea surface temperature. This environment supports research on biodiversity by facilitating process like the measuring, distribution, monitoring, and modelling of phytoplankton (microscopic marine plants), the provision of forecasts of sea state and currents, the monitoring of algal blooms and marine pollution, and the measuring of changes in the ocean productivity.

Global Land Vegetation Monitoring (GVM) provides a virtual environment that integrates satellite images of vegetative land cover. It facilitates specific research on how climate changes and land cover influence environmental resources. By having access to the data and tool of this VRE scientists can determine important measures like the total green leaf area for a given ground area, how much water will be stored and released by an ecosystem, how much leaf litter it will generate, and how much photosynthesis is going on.

As it should emerge from the brief description above, these four VREs offer innovative, user-tailored, and dynamically created collaboration environment even by exploiting the same enabling technology. In this environment scientists addressing a specific problems can access a number of geographically disperse cross-domain resources of different nature and operate with them as if these resources were belonging to their own organization (although in the limits imposed by the resources regulating policies).

4 Concluding Remarks

e-Infrastructures that provide application services for a range of user communities cannot ignore the diversity and specificity of their requirements. In this paper, we have argued that such requirements can be conveniently met by implementing a development approach based on Virtual Research Environments and briefly presented how this approach has been put in place in the context of the D4Science project.

Acknowledgments This work is partially funded by the INFRA-2007-1.2.2 - Deployment of eInfrastructures for scientific communities - Research Infrastructures - Seventh Framework Programme of the European Commission in the context of the D4Science project (Grant Agreement no. 212488).

References

1. Foster, I., Kesselman, C., Tuecke, S.: The Anatomy of the Grid: Enabling Scalable Virtual Organization. *The International Journal of High Performance Computing Applications* **15** (2001) 200–222
2. Foster, I., Kesselman, C., Nick, J., Tuecke, S.: The Physiology of the Grid: An Open Grid Services Architecture for Distributed Systems Integration. *Open Grid Service Infrastructure WG*, Global Grid Forum (2002)
3. 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** (2007) 59–80
4. Pagano, P., Simeoni, F., Simi, M., Candela, L.: Taming development complexity in service-oriented e-infrastructures: the gcore application framework and distribution for gcube. *Zero-In e-Infrastructure News Magazine* **1** (2009) 19 – 21
5. Assante, M., Candela, L., Castelli, D., Frosini, L., Lelii, L., Manghi, P., Manzi, A., Pagano, P., Simi, M.: An Extensible Virtual Digital Libraries Generator. In Christensen-Dalsgaard, B., Castelli, D., Jurik, B.A., Lippincott, J., eds.: 12th European Conference on Research and Advanced Technology for Digital Libraries, ECDL 2008, Aarhus, Denmark, September 14–19. Volume 5173 of Lecture Notes in Computer Science., Springer (2008) 122–134
6. Candela, L., Castelli, D., Ferro, N., Ioannidis, Y., Koutrika, G., Meghini, C., Pagano, P., Ross, S., Soergel, D., Agosti, M., Dobreva, M., Katifori, V., Schuldt, H.: The DELOS Digital Library Reference Model - Foundations for Digital Libraries. *DELOS: a Network of Excellence on Digital Libraries* (2008) ISSN 1818-8044 ISBN 2-912335-37-X.

TELplus: Aimed at Strengthening, Extending and Improving The European Library Service

Maristella Agosti

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

Abstract. The objectives of the TELplus project, which is aimed at strengthening, extending and improving The European Library service, are outlined together with some relevant results obtained in the first part of the project. In particular some results obtained in the work package 5, devoted to User personalization services – log file analysis and use of annotations – are reported.

1 Introduction

TELplus¹ is a project funded by the European Commission under the *eContentplus* Programme, within the domain of *Digital Libraries*. The project is supported by CENL, the Conference of European National Librarians², and is coordinated by the National Library of Estonia.

The overall goal of TELplus is to strengthen, expand and improve The European Library³. Managed by CENL, The European Library is an on-line service providing access to the electronic resources of the National Libraries of Europe. During the past years, The European Library has gradually expanded to offer access to the resources of 48 national libraries of Europe in 35 languages. Resources can be both digital – e.g. books, posters, maps, sound recordings, videos – and bibliographical. Quality and reliability are guaranteed by the 48 collaborating national libraries of Europe, where a national library is a library specifically established by a country to store its information database; national libraries usually host the legal deposit and the bibliographic control centre of a nation.

The European Library has been officially acknowledged by the European Commission⁴ as the organisational structure for the creation of Europeana⁵, the European digital library, museum and archive. Funded under the *eContentplus* Programme, Europeana is developing a prototype website providing access to some two million digital objects, including film material, photos, paintings, sounds, maps, manuscripts, books, newspapers and archival papers. Within this context, the results of

¹ <http://www.theeuropeanlibrary.org/telplus/>

² <http://www.cenl.org/>

³ <http://www.theeuropeanlibrary.org/>

⁴ <http://ec.europa.eu/>

⁵ <http://www.europeana.eu/>

TELplus will make a valuable contribution to the further development of Europeana. The project started in October 2007 and it ends in December 2009.

2 Objectives and Target Users

The main objectives of the TELplus project are:

- To enhance the service offered by The European Library by improving and facilitating access and use of the materials in national libraries in all Member States in a multilingual context and improving standard-based interoperability.
- To strengthen, extend and improve The European Library service by adding digital cultural and scientific/scholarly content and improving access to it in the following ways:
 - Adding and improving content of The European Library by:
 - OCRing more than 20 million pages of important multi-lingual content now available only as images according to current best practice;
 - Making the data that national libraries currently hold in proprietary databases (only searchable via the z39:50 protocol) OAI compliant and therefore harvestable for far greater usability;
 - Adding the National Libraries of Bulgaria and Romania as full members of The European Library.
 - Improving access and usability by:
 - Improving full text indexing and investigating automatic vocabulary mappings;
 - Learning from users through The European Library user group;
 - Creating a modular service infrastructure that enables users to integrate services with the European Library portal and adding a number of new services;
 - Work on user personalization to provide directions towards new services for users;
 - Providing translations of collection descriptions in 26 languages.

Since TELplus is closely related to The European Library, the project's target final user groups which are reached indirectly through TEL coincide with those identified for The European Library service itself. These are:

- Academic research community: facilitators and users of institutions of interest to academic community, university students, schools, educational establishments;
- Professional researchers: librarians and information professionals;
- Creative industry and cultural sector: publishing houses, market research agencies, knowledge and training centres of multinationals, museums, archives;
- Non-professional researchers: "the informed citizens".

For direct dissemination through its own website and its other dissemination tools the TELplus project is mainly reaching out to professionals and academics interested in the technical development of The European Library and of digital libraries in the wider sense.

The issue of the analysis of user requirements is specifically addressed by the Work Package 5 (WP5) of the project, aimed to develop specifications for the design of innovative personalized services for the final users of Europeana. The work undertaken in WP5 by the University of Padua team so far has been mainly focused on the HTTP log data analysis and the use of annotations as shortly reported in the following.

2 User personalisation services

The work on user personalization is necessary to direct the designers of advanced information services giving insights and directions towards new services of interest for the users. The findings serve to guide the design of software components for new added-value services of interest to the final users of digital libraries management systems.

Log File Analysis

The analysis of HTTP log data has been conducted over a period of time of eighteen months which goes from January 2007 to June 2008. The time span of the HTTP log data available, that is eighteen months, represents an exception on present literature and work conducted at international level, because most of the other studies, that have made public, are based on data of shorter time intervals. The availability of an HTTP log data set over such a long period of time is giving us the opportunity of deriving information that can also inform on the tendency of the use of the portal by final users, giving the opportunity of using these results as an aid in the planning of services and use of the portal.

Table 1: Synthetic Descriptive Statistics

	Total	Daily average	Per visit
No. of visitors	475,333	868.98	Not applicable
No. of visits	709,922	1,297.85	1.49 (per visitor)
Hits	58,053,293	106,130.3	81.77
No. of accessed pages	19,047,263	34,821.32	26.83
Bandwidth	1,059.51 GB	1.94 GB	1.53 MB
“Not viewed” hits	14,229,379	106,130.3	20.04
“Not viewed” bandwidth	2,876.63 GB	5.26 GB	4.15 MB

The work of knowledge extraction from the log data available is under way, by now some synthetic descriptive statistics are available and are reported in Table 1 where it is shown that in the analyzed period of time the number of unique visitors was more than 475 thousands, with an average of 870 visitors per day. More than 700 thousands visits

where recorded with a daily average of about 1,300; these visits produced around 58 millions of hits (HTTP requests to the Web server) which corresponds to a daily average of 106 thousands of contacts. The number of Web pages requested were more than 19 millions, which corresponds to almost 35 thousands of pages per day. "Not viewed" traffic, both hits and bandwidth, includes traffic generated by robots, worms, or replies with special HTTP status codes.

Use of Annotations

The work undertaken in the area of annotation sharing has designed and built a demonstrator where users can construct over time an useful hypertext, which relates pieces of information of personal interest - inserted by the final user - to the digital objects which are managed by the Digital Library System (DLS). In fact, the user annotations allow the creation of new relationships among existing digital objects by means of links that connect annotations together with existing objects. In addition, the hypertext between annotations and annotated objects can be exploited not only for providing alternative navigation and browsing capabilities, but also for offering advanced search functionalities, able to retrieve more and better ranked objects in response to a user query by also exploiting the annotations linked to them.

Therefore, annotations can turn out to be an effective way of associating this kind of hypertext to a DLS to enable the active and dynamic use of information resources. In addition, this hypertext can span and cross the boundaries of the single DLS, if users need to interact with the information resources managed by diverse DLS, as it is the case of The European Library where 48 national libraries cooperate in the service also with their internal implementations. This latter possibility is quite innovative, because it offers the means for interconnecting various DLS in a personalized and meaningful way for the end-user, and this is a big challenge for DLS of the next generation, as it is sketched in Figure 1.

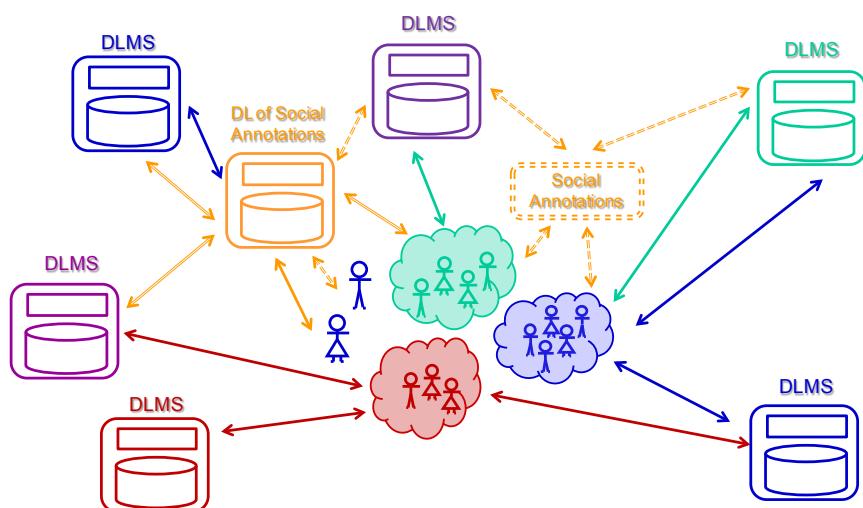


Figure 1

Author Index

Agosti, Maristella	120, 170	Jones, Sophia	150
Aloia, Nicola	154	Lops, Pasquale	108
Amato, Giuseppe	162	Lucchese, Claudio	88
Bache, Richard	60		
Baruzzo, Andrea	34	Manghi, Paolo	1, 150
Basile, Pierpaolo	51, 108	Marenglen, Biba	22
Basile, Teresa M.A.	22	Masiero, Ivano	80
Biba, Marenglen	22	Mass, Yosi	80, 88
Bolettieri, Paolo	88	Meghini, Carlo	154
Candela, Leonardo	1, 166	Melucci, Massimo	80, 100
Casarosa, Vittore	143	Mikuličić, Marko	1
Casoto, Paolo	34	Miotto, Riccardo	80, 100
Castelli, Donatella	1, 166	Orio, Nicola	80, 100
Catarci, Tiziana	136		
Concordia, Cesare	154	Pagano, Pasquale	1, 166
Crestani, Fabio	60	Perego, Raffaele	88
Crivellari, Franco	120	Peters, Carol	72, 158, 162
Dattolo, Antonina	34	Piva, Marco	136
Debole, Franca	162	Poggi, Antonella	136
de Gemmis, Marco	108	Rabitti, Fausto	88
Di Buccio, Emanuele	80, 100	Savino, Pasquale	162
Di Nunzio, Giorgio Maria	120	Semeraro, Giovanni	51, 108
Dix, Alan	136	Shmueli-Scheuer, Michal	88
Esposito, Floriana	22	Silvello, Gianmaria	12
Falchi, Fabrizio	88	Stamatogiannakis, Elefterios	120
Ferilli, Stefano	22	Sznajder, Benjamin	80
Ferro, Nicola	12, 72, 100	Tasso, Carlo	34
Gentile, Anna Lisa	51	Terella, Fabio	136
Giuliano, Raffaele	136	Tracanna, Emanuele	136
Ioannidis, Yannis	120	Triantafyllidi, Mei Li	120
		Vayanou, Maria	120