MANENT: An Infrastructure for Integrating, Structuring and Searching Digital Libraries

Angela Locoro, Daniele Grignani, and Viviana Mascardi

Abstract. Digital Libraries represent the commitment of research communities to preserve authoritative and well structured sources of knowledge, and to share archival organisations, methods and resources thanks to systems relying on standard metadata formats. This chapter describes some natural language processing techniques exploited for automatically extracting structural information from documents stored in Digital Libraries, based on the exposed metadata. The most prominent results achieved in this area are surveyed and discussed. As an example of an infrastructure for integrating, structuring and searching Digital Libraries based on natural language processing and semantic web techniques, we discuss the MANENT system. MANENT is a working prototype offering services of Digital Library content management and record classification and retrieval. It is hosted on a server at the Computer Science Department of Genova University and, starting from 2011, it will become publicly available. 475,000 records drawn from 138 repositories that all over the world expose OAI-PMH services have been downloaded, stored, and their automatic classification is under way.

1 Motivation

Scientific outcomes rely on institutional networks of researchers, leveraged by the Web in their intertwined activity that "help them in criticising and rectifying their findings and preserving the acquired knowledge by transmission to others" [35]. The community does not simply represent an aggregate of individuals based on social

Angela Locoro · Viviana Mascardi
Computer Science Department, University of Genova, Via Dodecaneso 35,
16146 Genova, Italy
e-mail: {locoro,mascardi}@disi.unige.it

Daniele Grignani
Department of Modern and Contemporary History, University of Genova, Via Balbi 6,
16126 Genova, Italy
e-mail: daniele.grignani@gmail.com

M. Biba and F. Xhafa (Eds.): Learning Structure and Schemas from Documents, SCI 375, pp. 315–341.
springerlink.com

relations through which information flow, but it is instead a community of practice, whose coherent behaviour is defined by the commitments of all its members [44]. In this scenario the presence of even more efficient tools for storing, filtering, sharing and retrieving all the needed theoretical and analytical knowledge becomes crucial. Although the Web seems to represent the ultimate technology for transforming the process of knowledge proliferation and availability, it is more and more clear that this technology is "a source of unprecedented amounts of information. In a content-rich environment where much material is no longer evaluated by traditional gatekeepers such as editors before it has the potential to reach large audiences, the ability to find trustworthy content online is an essential skill" [22].

Organisational criteria for storing resources should reflect the specific goals of structured information. If the goal is that of building an archive, then the most relevant element is the faithful adherence of the documents to their original source, obtained by strictly relating the document to the documental base of origin. In more operative scenarios the organisational criterion is the relational one.

On the one hand, as digitisation is a time consuming and costly effort, a careful analysis of the sources complexity should drive the design of effective devices oriented to a clear separation between the standard archival layer and the relational layer, tailored for specific enquiries. Only in this way changes occurring during activities can be done and remain at the operative level without any impact on the consolidated structure of the documental base.

On the other hand managing fragmented information turns out to cause an irreducible selection of some properties and the loss of other ones, which is typical of a process where the user selects pieces of information and re-contextualises them by creating, de facto, a new source of information as a result of researches [24]. A logical separation between the documental base and the data manipulated by the user can be obtained by setting up a work environment where the researcher may customise and create new metadata structures by following specific research projects criteria. The outcome of this process will generate a new source binded to the documental base on which it depends.

Moreover in the digitisation era every information object is available and accessible worldwide. The dynamic nature of a networked environment where such artifacts are created or their virtual surrogates are placed, outbursts the importance of how and to which extent the information and its context should be delivered to the final user. The role of metadata attached to any information source is then twofold: their schema represents both the high-level document structure and its semantic references to their contextual structure. The first role models the document itself, the second role encompasses the original conditions in which it was created and released as information source. These conditions are characterised by the piece of world knowledge strictly related to the document itself, that is a kind of information that surrounds the document content at one step of inference from all the other

objects that it implicitly or explicitly refers to (i.e. authors as people, which are part of a community, with their authoritative power and reputation as well as references that link the document to other documents, and so on).

This chapter surveys some methods, tools, and results relevant for the area of Digital Library integration, structuring, and searching. The approaches we are mainly interested in, are those based on semantic web and natural language processing techniques. Indeed, these are the founding techniques upon which MANENT roots. MANENT performs harvesting of metadata exposed in standard format from real world digital libraries and automatically classifies documents by topic, according to the WordNet Domain structure. This structure has been reproduced into a "WordNet Domain Ontology" that allows MANENT to easily represent and exploit semantic relations among domains. MANENT allows the user to search documents by expressing queries in natural language. It supports a "topic-based" search of documents relevant for the user query, based on the WordNet Domain Ontology. A more sophisticated "text-based" search, usually used for refining the topic-based one, exploits text semantic similarity between the user query and text that appears in documents metadata.

The chapter is organised as follows: Section 2 overviews standards, techniques and tools upon which MANENT roots. Section 3 describes MANENT and Section 4 reports the experiments we conducted with it and the results we obtained. Section 5 analyses the related work. Section 6 concludes and outlines the future developments of our research.

2 Background

MANENT and the infrastructures for Digital Library integration and structuring we will describe in this chapter rely on semantic web and natural language processing approaches. In this section we briefly recall the most recent standards, tools, and techniques relevant for designing and implementing such infrastructures. We assume a basic knowledge of XML [48], RDF [47], RDFS [46], and of WordNet [33].

2.1 Standards for Digital Library Access and Description

The Open Archives Initiative Protocol for Metadata Harvesting, OAI-PMH [34], provides an application-independent interoperability framework based on metadata harvesting. A OAI-PMH framework involves *data providers*, who administer systems that support the OAI-PMH as a means of exposing metadata, and *service providers* who use metadata harvested via the OAI-PMH as a basis for building value-added services.

OAI-PMH is hence based on a client-server architecture, in which "harvesters" request information on updated records from repositories. Requests for data can be

based on a datestamp range, and can be restricted to named sets defined by the provider. Data providers are required to provide XML metadata according to the Open Archives Initiative Protocol for Metadata Harvesting [36] and the Guidelines for Repository Implementers [37].

2.2 Ontologies and Related Languages and Tools

According to T. Gruber's definition *"in the context of computer and information sciences, an ontology defines a set of representational primitives with which to model a domain of knowledge or discourse. The representational primitives are typically classes (or sets), attributes (or properties), and relationships (or relations among class members). The definitions of the representational primitives include information about their meaning and constraints on their logically consistent application"* [21].

One of the most widespread languages for describing ontologies is OWL [45], a semantic markup language that extends the vocabulary of RDF.

Protégé is a widely adopted open source ontology editor and knowledge acquisition system developed by Stanford Center for Biomedical Informatics Research at the Stanford University School of Medicine [38].

2.3 WordNet Domains and the WordNet Domains Ontology

WordNet Domains[1] [27, 7] is a project developed by the Fondazione Bruno Kessler (FBK), Trento, with the purpose of providing WordNet synsets with a syntagmatic layer beside the existing paradigmatic layer (represented by semantic relations such as for example hyperonymy and meronymy). The project has the scope of better characterising a word meaning within its use in the language and in texts, hence reducing ambiguity.

The theory underlying WordNet Domains is that of Semantic Domains [28], stating that words in the lexicon can be grouped together into sets of strongly associated concepts, determined by the lexical coherence property. Based on such property some set of words tend to highly co-occur in texts, which means that an underlying semantic layer of associations among them exists and thus can be modelled accordingly.

The operation of tagging WordNet synsets with domain labels has been conducted by the FBK team partially by hand (by annotating a subset of synsets) and by automatically extending such labels to related synsets through the WordNet hierarchy, fixing the automatic procedure with corrections where necessary. The domain labels are those of the Dewey Decimal Classification system[2], a standard largely

[1] `http://wndomains.fbk.eu/index.html`.

[2] `http://www.oclc.org/dewey/versions/default.htm`.

adopted by library systems. The task of labelling synsets has been conducted on WordNet version 2.0. For the more recent versions of WordNet, mappings files from the oldest version to the newest exist and are freely available.

Table 1 shows an excerpt, for the first 32 top-ranked domain labels, of the number of WordNet 3.0 synsets which have been tagged by domain labels. Each synset can be labelled with more than one label and the whole number of synsets upon which this statistics was conducted is 115.248.

Table 1 WordNet Domains labels and number of synsets tagged with each label. The *factotum* label refers to words whose use in the language is not related to any specific domain of discourse.

Domain	♯ Syn	Domain	♯ Syn
biology	23814	law	1887
plants	17849	music	1857
factotum	16099	linguistics	1760
animals	12265	metrology	1673
chemistry	6495	administration	1462
geography	5169	physics	1342
medicine	4223	pharmacy	1339
person	3790	psychological_features	1327
anatomy	3340	transport	1283
religion	2249	geology	1269
gastronomy	2215	food	1197
buildings	2107	fashion	1194
history	2044	economy	1157
military	2033	entomology	1075
politics	2032	physiology	1065
literature	1986	industry	1008

As part of our recent research on automatically discovering the WordNet domain of ontology entities and of entire ontologies, we took the WordNet Domains taxonomy[3] and codified it in OWL using Protégé. We called this new ontology WordNet-Domains.owl. It consists of 160 domain labels divided as follows:

- 11 first level classes which represent the upper layer of domains classification. They are: *applied_science*, *doctrines*, *factotum*, *free_time*, *metrology*, *number*, *person*, *pure _science*, *quality*, *social_science*, *time_period*;
- 42 mid level classes that are used to tag synsets representing concepts used at an intermediate level of generality (e.g. *medicine*, *economy*, *sport*, and so on);
- 107 low level classes that are subclasses of one of the 42 mid level concepts or belong to a further level of specialisation and are also used to tag synsets, which are relative to concepts used in more specialised domains (e.g. *psychiatry*, *banking*, *athletics* and so on).

[3] Available at the official page of the project: `http://wndomains.fbk.eu/hierarchy.html`. Last accessed on 30 June 2010.

In [26] we describe how we used the WordNet Domains ontology to successfully tag ontology concepts with WordNet Domains in an automatic fashion. We tagged each concept from one ontology with the label assigned to the WordNet synsets whose lemmas correspond to the concept itself, expanded the domain label tags assigned to each concept with the top domain labels contained in the WordNet Domains ontology through an inference procedure based on its hierarchy, and computed frequencies on the whole set of tags to determine the most frequent domains at ontology level as well as at concept level. Following this approach, we were able to successfully assign the correct domain among the first-level ones to each of the 17 real ontologies we used as testbed, and the correct domain among the mid level classes to 15 of them.

In MANENT, we exploit our WordNet Domains ontology in a very similar way, and with the same satisfactory results (see Section 4).

2.4 Text Semantic Similarity

The problem of determining the semantic similarity of sentences has been widely discussed in the literature starting from the late sixties, when two pioneer works [41, 39] were published on concrete applications of text similarity measures.

From then on, significant research results were achieved on this topic. Many recent papers compare similarity measures according to different viewpoints [32, 4, 31].

In MANENT we are experimenting the WordNet-based semantic similarity measurement application by T. Simpson and T. Dao[4], licensed under The Code Project Open License (CPOL).

Given two words s and t, Simpson and Dao's algorithm computes *SenseWeight*, a weight calculated according to the frequency of use of the senses assigned to s and t respectively by exploiting the adapted Lesk algorithm [5] and the total of frequency of use of all senses, and *PathLength*, the length of the connection path from s to t.

The similarity of s and t is computed as

$$sim(s, t) = SenseWeight(s)*SenseWeight(t)/PathLength$$

and the similarity of two sentences X and Y is computed based on the similarities of $< s,t >$ such as $s \in X$ and $t \in Y$.

Other sentence semantic similarity measures will be tried in the future: MANENT functionality of refining queries by exploiting text similarity measures is still in an early stage. For demonstrating the feasibility of our approach, however, Simpson and Dao's application was powerful enough, and easy to use.

[4] `http://www.codeproject.com/KB/string/semanticsimilaritywordnet.aspx`.

3 MANENT

In this Section we introduce the MANENT architectural and functional features. In order to keep the chapter readable for both specialists and non-specialists, we avoid the technical details and keep our discussion at a high level of abstraction.

The architectural layer we propose in MANENT aims at fostering research communities of practice and encompasses a digital library infrastructure and an OAI-PHM harvesting service that conveys information exchange and retrieval as well as automatic classification of metadata contents by topic. The key characteristic of MANENT is that of modelling, describing, storing and retrieving information objects by means of an OWL ontology derived from EAD (Encoded Archival Description [13]) as a formal and standard conceptualisation of archival rationales maintained by the Library of Congress in partnership with the Society of American Archivists.

Moreover, automatic classification and search services are also provided in MANENT. The WordNet Domains Ontology is exploited for automatically classifying heterogeneous documents owned by hundreds of Digital Libraries, based on the schema mining of the metadata associated with them. Once the metadata harvester retrieves the metadata describing information objects, a service for automatic domain detection of such contents is run in order to provide a classification of both local and remote information based on their topic. A mechanism for extracting the most relevant keywords from metadata annotations (either local or remote) and for tagging them with domain labels has been developed and will provide new structured data for classifying contents, extending existing schemas and ontologies or simply indexing and searching resources based on them.

3.1 The MANENT Architecture

The MANENT architecture includes the following components:

- **User Interface Portal** for archive visualisation, browsing, searching, content management and administration tasks.
- **Digital Library Content Manager**, the core content management component in charge of the basic functionalities such as managing contents and requests for visualisation and browsing, managing and updating collections, adding new documents, and so on.
- **Collection Template Composer** for designing and structuring EAD compliant archives.
- **Collections Object Manager**, consisting in all the operations needed to access the knowledge base. Every operation is essentially a SPARQL [42] wrapper.
- **Knowledge Base Component**, the repository with all its Collections and CollectionSets data.

- **Metadata Integration Service**, offering services for integrating different metadata format. This component contains the metadata mapper interface for easy configuration of metadata matching tasks.
- **Metadata Harvester Service**, working as a web service interface for OAI-PMH metadata harvesting.
- **Metadata Classification Service**, providing automatic classification by topics of the repositories metadata records harvested; the service is treated in a dedicated Section (namely Section 3.2) due to its higher relevance with respect to the scope of the chapter.

Besides exploiting the WordNet Domains ontology for offering classification services, MANENT relies on two OWL ontologies that model the hierarchical structure of MANENT archives ("Archives Structure Ontology"), and the basic elements for translating metadata formats into OWL ("Metadata Structure Ontology") respectively. We may consider these two ontologies as the MANENT "meta-model".

The "Archives Structure Ontology"

Inside this ontology the following classes are defined:

- *Collections*: represent a set of documents ordered and preserved together; collections are created under CollectionSets; collections may be the result of a single process or of a specific activity and are described through a metadata structure derived from EAD and designed by means of the Collection Template Composer. At the time being MANENT digital objects are maintained in the filesystem while the knowledge base keeps a reference to their path through the dao[5] tag. The solution is temporary and the creation of a Digital Object Management System based on OWL is foreseen.
- *CollectionSets*: high level containers for aggregating related or linked Collections. Each CollectionSet may contain either Collections or CollectionSets that, in this case, are called CollectionSubSets. The two elements represent a set of collections created or collected by a single user or institution during their activities. As for a Collection each CollectionSet is also described through a metadata structure derived from EAD, with the difference that, in this case, their structure is fixed a-priori with a set of tags that is not changeable.

The "Metadata Structure Ontology"

Inside the "Metadata Structure Ontology" the class Element has been defined to represent the set of all XSD elements of the EAD XML Schema [14]. Each of them is translated into an individual of that class. The result is the EAD ontology that is stored in the knowledge base. This procedure is extended to XML Schemas of any

[5] Encoded Archival Description Tag Library, Version 2002, EAD Elements, <dao> Digital Archival Object, `http://www.loc.gov/ead/tglib/elements/dao.html`.

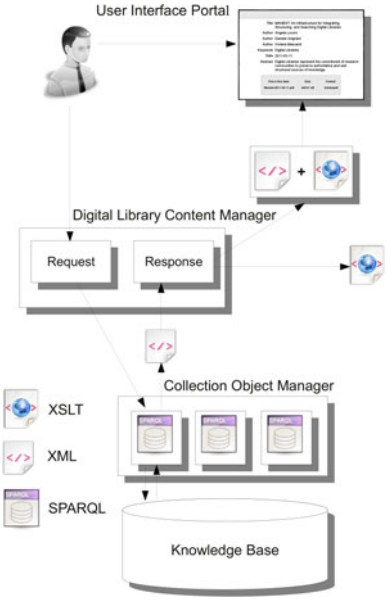

Fig. 1 A Collection browsing and visualisation workflow in MANENT

other metadata formats (i.e. in this way the MARCXML [29] XML Schema may become MARCXML ontology).

The device responsible for the translation, namely XSD2OWL, is able to parse all the XSD elements and convert them in individuals of the class Element following the "Metadata Structure Ontology". For each XSD element, attributes and relationships are detected and created. At the end of the process the generated ontology is stored in the knowledge base.

Browsing Contents in MANENT

Figure 1 depicts a typical MANENT browsing operation. The Digital Library Content Manager receives a request from the user, keeps in memory its type and invokes the proper methods of the Collections Object Manager while waiting for results. When data are ready they are organised based on the resource type description (CollectionSet or Collection) and visualised.

Collection Management

CollectionSets and Collections are created with a request for the creation of a new element sent to the Digital Library Content Manager. The component queries the Collections Object Manager to obtain the proper schema and sends it to the interface

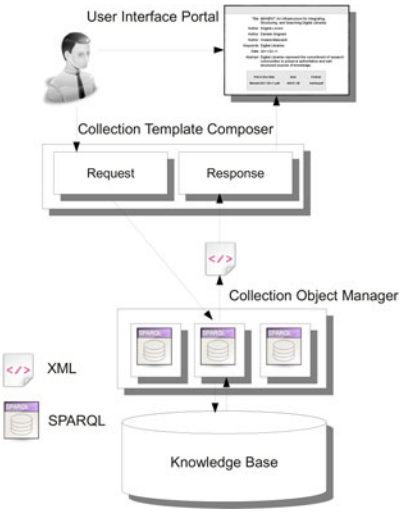

Fig. 2 An example of Collection creation workflow in MANENT

for user data insertion. Once completed, the two components communicate again in order to store the new data in the knowledge base and visualise it to the user.

The Collection Template Composer is dedicated to the creation of a Collection and may be activated only inside the CollectionSet. It consists of an interface whose purpose is to let the user customise the Collection structure by selecting different combinations of the tags set based on the EAD ontology; as the Collections Template Composer is constrained by the EAD ontology, the insertion of different tags from those expected is not allowed.

The creation procedure for a Collection is depicted in Figure 2 where the user selects a CollectionSet, sends a request for creating a new collection, creates it, and the results are stored in the knowledge base and visualised.

The Metadata Integration Service

As already discussed, MANENT represents metadata in an internal format derived from EAD and formalised in the EAD ontology. A matching service is responsible for the mapping of different metadata formats with the EAD ontology provided that they are previously translated in OWL ontologies by means of the XSD2OWL procedure. In order to integrate information expressed in a format *F* different from EAD, a manual conversion from *F*'s metadata format and the EAD ontology is required. An easy-to-use interface drives the user in the definition of "*F* to EAD" mappings. Of course, this conversion is needed only once: when mappings from *F* elements to the EAD ontology have been defined, they are stored into the MANENT knowledge base and will become part of the available representation formats of MANENT collections.

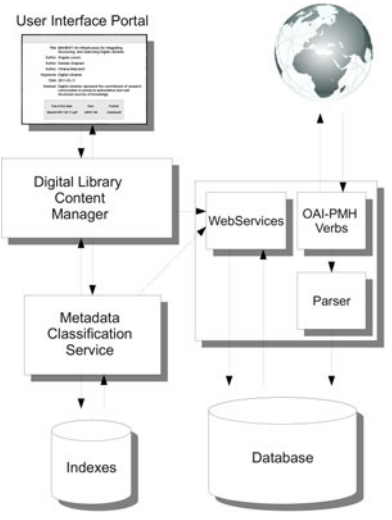

Fig. 3 The MANENT Metadata Harvester and Metadata Classification Services

The Metadata Harvester Service

This service is responsible for metadata harvesting and relies on OAI-PMH protocol release 2.0. The Web Services implemented are integrated with the MANENT interface that accesses data retrieved by means of the harvester itself. Figure 3 exemplifies the service.

The harvesting mechanism is built upon URLs belonging to data providers[6], which are under the administration of the Open Archives Initiative[7]. These are periodically read and stored locally.

The service is modelled according to Open Archives Initiative guidelines [37], while the download procedures are those envisioned by the OAI-PMH standard[8]. The service reads one after the other the URLs listed in the data providers page. Two kinds of harvesting are supported:

- Complete, where the repositories description, the list of metadata formats, the repositories structure and, for each metadata format, the list of records are downloaded.
- Incremental, where only those records added in the period between the last download and the current time instant, are downloaded.

[6] "Data Providers [are] systems that support the OAI-PMH as a means of exposing metadata"; The Open Archives Initiative Protocol for Metadata Harvesting, Document Version 2008-12-07T20:42:00Z, `http://www.openarchives.org/OAI/openarchivesprotocol.html`.

[7] `http://www.openarchives.org/Register/ListFriends`.

[8] `http://www.openarchives.org/OAI/openarchivesprotocol.html`.

3.2 Metadata Classification According to the WordNet Domains Ontology

On the documents' metadata retrieved and stored following the approach described in the previous section, we run our classification algorithm that automatically assigns a set of WordNet domains to each document, based on its metadata. Since concepts in the WordNet Domain Ontology (see Section 2) correspond to WordNet Domains, this activity amounts to linking each document to one or more concepts in the WordNet Domains ontology, hence classifying it according to the WordNet Domains ontology structure.

Our approach integrates statistical and semantic natural language processing techniques. Stemming from an assignment of domain labels to each relevant noun found in document metadata and computing their occurrencies, we extend such domain labels with super-domains ones by reasoning on the WordNet Domains hierarchy. We then propagate frequencies of super-domain labels by summing subdomains frequencies up to the topmost domain nodes. The list of the most relevant domains is output by a scoring function that weights domain labels according to the metadata type they appear in (i.e., we weight the *dc_subject* field in the Dublin Core [12] metadata schema more than the *dc_description* one, since our goal is to classify documents by topic and we expect that *dc_subject* represents the document topic better than other fields).

A detailed description of the main steps of our procedure is introduced in the following paragraphs, where we depict the pre-processing steps conducted on metadata contents, the tagging stage, and the score computation. To better exemplify each passage we apply the automatic classification methodology to the record shown in Figure 4, described following the Dublin Core metadata schema.

```
dc:title A weather forecast study
dc:subject global warming
dc:description This paper deals with a study on weather
based on forecasts of the last 30 years where the weather
has changed due to climate conditions.
```

Fig. 4 Example of document metadata

For each record we execute the following steps.

Tokenisation and POS tagging. By exploiting the GATE [9] Natural Language Processing tool we tokenise the text in the record, tag contents with a POS (part-of-speech) tagger and retain only noun words.

Lemmatisation. We lemmatise each noun in order to gain the canonical word forms from derivationally formed ones. To acquire the lemma from one noun we use the WordNet dictionary.

[9] `http://gate.ac.uk`.

Lemmas occurrencies count. We count the occurrencies of each lemma in each metadata content slot. The formula we apply is

$$\forall lemma_i \in metadata_j = F_{metadata_j}(lemma_i) = \mid lemma_i \mid$$

and results in the total number of times the lemma occurs into a metadata slot.

Filtering. In case the lemmas occurrencies count results in a huge amount of nouns with low occurrence, which do not characterise the domain of discourse (which may happen with the *dc:description* field in particular), we filter out the least relevant words on this metadata field, by normalising the occurrencies of each word and retaining only those words whose frequencies sum amounts to 50% of the total frequencies counts. The frequency computation is as follows:

$$p(w_i) = \frac{freq(w_i)}{\sum_{i=1}^{W} freq(w_i)}$$

where w is a word lemma, $freq$ is the number of occurrencies of that lemma, W is the total number of occurrencies of the whole set of words in the description field. The sum of all $p(w_i)$ amounts to 1.

Assignment of the right WordNet domain to each lemma. We look into the "WordNet synsets - WordNet Domains" mapping files (WnToWnD)[10] and assign to each word lemma its domain labels as they result from such file. The result of the mapping operation on our example record is shown in Figure 5.

```
dc:title weather [meteorology] forecast [meteorology] study
[school]
dc:subject warming [meteorology]
dc:description paper [factotum] study [school] weather
[meteorology] forecast [meteorology] year [time_period]
weather [meteorology] climate [meteorology] condition
[factotum]
```

Fig. 5 Document metadata after text processing: domain labels are in squared brackets

Extension of WordNet domains to super-domains following the WordNet Domains hierarchy. Besides the "direct tagging" with the WordNet domain associated with the given lemma (if any), we also tag lemmas with the super-domain labels of the domain labels just assigned. By looking at the WordNet Domains ontology and, hence, at the whole domain hierarchy space, we add all the superclasses, from the

[10] The WordNet Domain version we used is 3.1. For conversion from WordNet 2.0 to WordNet 3.0 synsets we use mappings files available at `http://www.lsi.upc.es/~nlp/tools/mapping.html`.

direct superclass up to the root domain label. What we obtain in the end is a set of domain labels associated with each lemma, that we represent as

$$Dom(l_i) = \{d \mid d \in D\}$$

where l stands for the lemma and d stands for a domain label belonging to the WordNet Domains set of labels D.

Lemmas that were tagged with no domain label are eliminated.

Score computation. We associate the number of occurrencies of each word with the domain labels by means of the following formula

$$\forall d \in Dom(l_i), Dom(l_i) \in metadata_j = F_{metadata_j}(Dom(l_i)) = \mid l_i \mid$$

and we apply a weight to this number, according to the type of metadata field to which the domain labels belong to. In this case we weighted domain labels in *dc:subject* 1, domain labels in *dc:title* 0.5, and domain labels in *dc:description* 0.25. For each domain label we then compute the following score function:

$$s(d_i) = \sum_{j=1}^{3} F_{metadata_j}(d_i) * w_j =$$

$$F_{dc:subject}(d_i) * 1 + F_{dc:title}(d_i) * 0.5 + F_{dc:description}(d_i) * 0.25$$

Ranking of domain labels. We rank the domain labels according to the resulting score. Each relevant root domain together with its sub-domains, if present, will appear in the ranking. An example of the final results of our procedure on our sample document is depicted in Figure 6.

```
[pure_science, earth, meteorology] (3)
[social_science, pedagogy, school] (0.75)
[factotum] (0.5)
[time_period] (0.25)
```

Fig. 6 Document domain classification according to WordNet Domains hierarchy after score computation and ranking

The *Factotum* domain is filtered out. If we want to visit the sub-domains and take the best of them with highest score we can see if the most specific domain has been ranked.

Querying and Matching Metadata Contents to Discover Semantic Similarities

Starting from a natural language query expressed by a user, we extract the WordNet domains associated with them by using the approach discussed above, and we exploit them to retrieve those records whose WordNet domains better match those

in the user's query. The matching criterion we adopt is to look for records whose WordNet Domains have the highest overlapping with the user's ones. For refining the query, we use relevant keywords extracted from the metadata and match them with the keywords extracted from the user's query. Moreover, we may use semantic text similarity to further refine queries, when a set of records annotated with the WordNet domains and keywords matching those extracted from the user's query have been found and more fine-grained search must be performed.

4 Experiments and Results

We conducted our experiments on a dataset composed by 10 repositories, chosen among 138 randomly picked up from the 1.342 all over the world repositories that expose OAI-PMH services. We downloaded only records within a temporal range (January 2008 to October 2010), able to capture the more recent entries of each harvested provider, for a total of 475.000 records; to begin with, we selected 10 repositories, for a total of 1000 records automatically classified, showing different features in terms of both their content and their structure, and we asked to domain experts in the records' disciplines to manually verify the correctness and completeness of the automatic classification over 100 of them (10 for each repository).

The harvesting procedure is the one depicted in the "complete harvesting" paragraph except for minor details that do not change its behaviour in a substantial way. Since the only format always available for all the repositories is the Dublin Core (*dc* from now on), we harvested records in this format.

The selection of the 10 repositories out of 138 as well as of the 100 records belonging to them to be manually inspected has been based on different qualitative aspects. A discriminating factor was the completeness of the metadata available for each record (i.e., most records of the 10 repositories have at least one *dt:title*, one *dc:subject* and one *dc:description* field). Another relevant factor was the domain of the repository from which the 100 manually evaluated records were drawn from: we considered mono-thematic repositories as well as miscellaneous ones. Notwithstanding an ideal choice would have been to capture at least one repository for each of the WordNet top domains, this choice turned out to penalise completeness of the metadata subset, as most of the 138 repositories lack some of the metadata we choose to exploit for our analysis.

A good compromise was then to select as many metadata complete repositories as possible. For our experiments we performed a further selection on the dataset language, by automatically detecting only metadata records written in English. To operate such selection we used the Java Text Categorisation Library (JTCL)[11] [11], a tool for guessing the language of sentences. A description of the 10 repositories is outlined in Table 2. For each of them we report the institution that owns the repository and the total number of records downloaded. In the rest of this section we will identify such repositories by their number, from r1 to r10. In the sequel, we refer to the 100 records that were manually inspected as "the benchmark".

[11] `http://textcat.sourceforge.net`.

Table 2 The 10 repositories from which we selected our benchmark

Institution owning the repository	♯ rec
r1. Centro de information y gestion tecnologica Matanza, Cuba	154
r2. University of Stirling, Scotland	2.204
r3. Queensland Dept. of Primary Industries and Fisheries, Australia	1.502
r4. Nara University of Education Academic Repository, Japan	1.962
r5. University of Bayreuth, Germany	250
r6. University of Hohenheim, Germany	304
r7. University of Fukui, Japan	1.639
r8. University of Regensburg, Germany	386
r9. Stellenbosch University, South Africa	3.393
r10. Indiana University, USA	831

As *dc* standard allows the multiple cardinality of each metadata (i.e. each document may be described by more than one *dc:subject* field and so on), our algorithm processes them in order to obtain a unique entry for each metadata type by concatenating all contents of a *dc* field as they appear in the document metadata record description.

Once the benchmark was set up through such preliminary normalisation steps, we proceeded by applying the classification procedure depicted in the previous section to each record in the benchmark. For each record we obtain a list of relevant keywords as well as the prevailing domain labels arranged hierarchically.

The topic detection results are reported in Table 3 where, for each repository and each domain label whose ranking score was among the first three ones, the number of documents automatically classified according to such domain is shown.

In particular the ranking mechanism adopted for each record works as follows:

- for the top level WordNet Domains (namely, the 11 first level classes that include applied science, doctrines, factotum, free time, metrology, etc; see Section 2) ranked according to the first two higher scores, search their direct sub-domains (the 42 mid level classes that include medicine, economy, sport, and so on) at the same ranking level down to the third ranking score;
- do the same for their leaf domains (classes at the lowest level: psychiatry, banking, athletics, ...), if they exist and if the scoring function has ranked them among the first three ranking scores.

For the final classification shown in Table 3 we used the entire set of domain labels ranked at record level as explained above and we aggregated each domain label by summing up the number of records tagged with them. For conciseness, we only show the first two top domains with higher rank and, for each of them, the first three higher sub-domains.

To evaluate the correctness and completeness of our classification algorithm we asked domain experts to manually check our benchmark.

Table 3 Final classification of the benchmark, an excerpt

domain label	r1	r2	r3	r4	r5	r6	r7	r8	r9	r10
social_science	66	70	35	74	32	49	44	63	83	46
commerce					5	11				
economy	28	18	16		7	27		12		11
industry	13						10			
law	29									
pedagogy		20		51	2		11	16	11	11
politics		24	12			7	9		65	11
publishing									11	
religion								9		
sociology				29						
telecommunication			6	14						
applied_science	43									
architecture	29									
computer_science	12									
engineering	12									
pure_science		46	86	39	79	63	81	51	76	83
biology		24	63	19	24	29	35	16	10	37
chemistry		11	25	11	40	23	41	12	7	46
earth		14	27	14		17		16	69	
physics					26		16			21

The evaluation results are summarised in Table 4 where, for each group of records, we report the presence of the relevant *dc* metadata with their number, the number of correctly detected top level domains (*top lev.* column), sub-domains (*sub-dom.* column) and, if present and properly assigned, the number of leaf domains (*leaf dom.* column). The total number of domain labels assigned (♯ *dom. lab.* column) together with the average number of labels assigned to each record (*avg lab.* column), that we recall are resulting from the ranking mechanism selection above explained, are also shown. A sum of each column value is reported in the *Tot* row and the percentage value (see row *% on Tot*) has been obtained by dividing the total of correctly assigned domain labels for the number of records correctly tagged, while the average number of total labels assigned to each record is reported in the *avg* row and has been obtained by dividing the total number of labels assigned to the benchmark (518) by 10. Moreover, an indication of relevant top level domains and direct sub-domains that are missing, meaning that our procedure was not able to guess them, is reported in the discussion below together with an overview of the results.

The overall results of the evaluation show that about 89% of the records were tagged with a correct top domain label whereas 72% have being assigned a correct sub-domain label, which are very encouraging results. At the leaf level, only 20% of the leaf domain labels were correctly set. Following the preliminary results of [26] we confirm our claim: the top level domains and their direct sub-domains are

Table 4 Manual evaluation over the benchmark carried out by domain experts

rep.	title	subj.	descr.	top lev.	sub-dom.	leaf dom.	♯ dom. lab.	avg lab.
r1	8	4	10	8	5	0	61	6.1
r2	5	14	10	8	5	0	56	5.6
r3	7	9	10	10	8	0	42	4.2
r4	6	10	8	9	9	4	66	6.6
r5	7	6	10	9	9	1	42	4.2
r6	10	4	10	9	6	2	40	4
r7	6	12	10	7	5	1	55	5.5
r8	7	28	12	10	8	2	45	4.5
r9	7	16	10	9	5	0	48	4.8
r10	7	8	10	9	9	0	63	6.3
Tot	70	111	100	87	69	10	518	
% on Tot				89%	72%	20%		
avg								5.18

those able to better classify documents, while the leaf domains are still difficult to be reached and correctly detected. In support of this evidence we may observe that 50% of the records examined do not even have a leaf domain in the first three ranking scores and only 10% of the records have been tagged with a sound leaf domain.

For the limited significance of leaf domain labels we exclude them from the following discussion, that provides a qualitative evaluation of the results obtained in each individual repository.

The main lesson that we learned from this evaluation, is that records about medicine, genetics and biology are among the most difficult ones to correctly classify in an automatic way. This is quite surprising, since, instead, they are among the easiest ones to classify for a human being, even without specific skills in the field. Computer science is again almost hard to recognize, probably because of the use of very technical terms.

These results, once confirmed by experiments on larger benchmarks, might provide the basis for suggesting an extension to the WordNet Domains classification in order to keep track of specific terms that strongly characterise a domain, but that are not considered yet.

- In Repository 1, 5 records out of the 10 in the benchmark deal with computer_science but were not correctly tagged with this sub-domain. On the other hand, the 5 remaining records were correctly tagged with top and sub-domain applied_science → computer_science or engineering, which result both correct, and the only one about sociology was correctly labelled with the social_science top domain even if the automatically detected sub-domain was not sociology, but economy and commerce.
- In Repository 2, 9 records out of 10 were assigned the correct top level domain, while only 50% of the records were tagged with a correct sub-domain. The

only record whose top domain classification failed was about applied_science → computer_science. Other relevant missing sub-domains were sociology (for 2 records), biology (1 record) and again computer_science (1 record) despite their top level domains were correctly found.
- For Repository 3, all the top level domains were exactly detected and 8 records out of 10 were tagged with their correct sub-domain. The two records with wrong sub-domains dealt with agricultural systems from the point of view of soil fertilisation and of agricultural models. The correct tagging should have been applied_science → agriculture.
- In Repository 4, 9 records out of 10 were tagged with correct top domains and sub-domains. The only misclassified record should have been tagged with pure_science → mathematics.
- Repository 5 shows a situation similar to Repository 4, with 9 out of 10 records correctly classified at top level as well as sub-domains level. The misclassified record was about medicine and genetics.
- For Repository 6, the 4 misclassified records (including the one with wrong top level) are about biology.
- In Repository 7, only 7 records out of 10 have been classified with a correct top level domain. As a consequence only 50% records were assigned a correct sub-domain. The records with wrong classifications lacked pure_science and biology (1 record) as well as applied_science (2 records), sociology (1 record) and medicine (3 records).
- Repository 8 has only 2 records out of 10 with misclassification at sub-domain level, caused by missing pedagogy and sociology labels respectively.
- For Repository 9, 9 records out of 10 were correctly classified at the top domain level (the misclassified one should have been tagged with social_science → publishing) whereas only 5 records were assigned the right sub-domain. Among these ones, 4 records miss the pedagogy and psychology sub-domains. One record represents a borderline case, since it contains only 2 words and the human expert herself could not assign a discriminating sub-domain.
- In Repository 10 only 1 record was misclassified and its top and sub-domains should have been applied_science and medicine respectively.

From the above results we may conclude that the top level domains classification gives very positive results. As shown in Table 4, our procedure tags each record with a limited amount of domains (about 5 on average) in order to limit noise and to provide very synthetic results.

Considering the top level domain classification, in 20% of the repositories (r3, r8), 100% records were correctly classified; in 50% of the repositories (r4, r5, r6, r9, r10), the correct classification applied to 90% of the records; in 20% of the repositories (r1, r2), to 80% of the records. The worst case is represented by Repository 7 where only 70% of the records found a correct top level domain classification.

In 50% of the repositories, sub-domain classification succeeds for more than 80% of the records (r3, r4, r5, r8, r10). In the remaining repositories, the percentage of records assigned a correct sub-domain is between 50% and 60%.

Thanks to these experiments carried out on a large set of records, we observed that many *dc:subject* entries do not affect the classification performance. This confirms our impression that the manual creation of metadata is often error-prone and a service as the one MANENT provides may prove useful in real world scenarios where data provided by users are incomplete.

Despite completeness of our classification procedure (namely, how many domain labels that should have tagged the record, were actually discovered by it) cannot reach impressive values because of the limited number of domain labels we assign to each record, correctness (namely, how many domain labels discovered by the classification procedure, are indeed correct) is very good. Refining techniques such as metadata content relevant keywords, extracted and tagged with domain labels as well as sentence similarity methodologies may be further applied to improve the completeness while keeping the correctness.

5 Related Work

In this Section we outline some state of the art Digital Libraries infrastructures as well as approaches that exploit WordNet Domains.

5.1 Related Work on Digital Libraries Infrastructures

The consistent amount of investments towards European projects dealing with digital libraries infrastructures for cultural heritage and preservation[12] witnesses the swift growth of a research field becoming crucial for the management of information commitment foreseen in the near future.

Studies on the state of the art of semantic digital libraries architectures [6], [25] emphasise some conclusive aspects for the sustainability of new generation infrastructures. They are:

- easy to use information discovery that may strongly rely on natural language facets;
- design of digital libraries infrastructures and services that should more and more rely on trusted reference models and well grounded standard resource description formats enriched with semantics;
- basic services such as indexing and classification augmented with user-centered annotations for systems customisation and evolution;
- interoperability at different levels of granularity: from document descriptions, through systems architectures and their tight integration.

Projects that have contributed to the fulfillment of this vision towards the integration, structuring and searching of Digital Libraries are, to cite only a few, the TELplus

[12] `http://cordis.europa.eu/fp7/ict/telearn-digicult/`, following the link to “DigiCult”.

project[13] and the DELOS project[14]. The main outcomes are the definition of a Reference Model [10] for Digital Libraries systems design and a Digital Library Management System [2]. Formal models for semantic annotations [3] have been studied. A set theoretic model [17] for managing hierarchical structures such as those of the OAI-PMH metadata harvester at resource level and those of the EAD at archival description level has been defined. A search and retrieval component [16] focusing on annotations similarity measures based on boolean operators and hypertext relational features is also provided.

The BRICKS [25] infrastructure joins the flexibility of a service oriented distributed architecture able to orchestrate interoperability among digital library nodes. A metadata manager component that relies on RDF representation of different standard schemas is in charge of providing advanced query search based on SPARQL syntax.

In JeromeDL [25] several ontologies have been integrated under the MarcOnt[15] ontology umbrella, whose design roots from MARC 21 bibliographic standard. MarcOnt is expressed in OWL and encompasses FOAF[18], Dublin Core, BibTeX[16] and S3B Tagging [17] ontologies, each of them representing one aspect of a semantic digital library system, namely people, resources, metadata and users annotations respectively. Based on that some services are built to provide semantic query in form of regular expressions templates from which the users may choose. Moreover search operations are based on full-text and bibliographic search as well as query expansion based on keywords suggested by the users that are saved in their preferences profile for later use and results ranking. To the best of our knowledge topical information in JeromeDL, which can be based on several classification standard such as DDC[18], UDC[19] and LCC[20] has to be manually added by librarians or resources owners and no automatic classification service is foreseen.

In the HarvANA system [23] institutional metadata and users annotations are both harvested and aggregated. An OAI-PMH interface to remote user-annotation servers has been developed for storage and retrieval whereas a mapping procedure from RDF converted annotations to Dublin Core schema is performed. In this way the system is able to index both metadata records and user defined annotation records. The search operation on resources can be then performed on both kind of

[13] `http://www.theeuropeanlibrary.org/telplus`.
[14] `http://www.delos.info`.
[15] `http://www.marcont.org/ontology/2.0`.
[16] `http://purl.org/net/nknouf/ns/bibtex`.
[17] `http://s3b.corrib.org/tagging`.
[18] Dewey's Decimal Classification.
[19] Universal Decimal Classification, `http://www.udcc.org`.
[20] Library of Congress Classification
`http://www.loc.gov/aba/cataloging/classification`.

semantic information. As far as we know the system works mainly with images repositories and annotations are suggested to the users via a controlled vocabularies interface, hence no automatic tagging service is provided.

Starting from the mid nineties, the use of ontologies for the integration of large heterogeneous information sources has been the subject of a very lively research activity [43].

Among the recent systems that exploit ontologies for annotating and classifying documents in knowledge sources, we mention SOBA [9] which processes structured information, text and image captions to extract information and integrate it into a knowledge base whose coherence is ensured by a reference ontology, built at system design time.

The approach discussed in [8] is even closest to ours, since it presents a framework for integrating digital library knowledge sources as well as facts extracted from the content under consideration by means of an ontology-based digital library system. Documents in the knowledge sources are annotated and classified according to the PROTON upper ontology[21] using natural processing techniques, in the same way as we annotate and classify records according to the WordNet Domain Ontology. Bloehdorn et al. allow users to express queries in natural language, as we do. The main differences between our work and Bloehdorn et al.'s one is that in their work, topics extracted from metadata and unstructured documents are instances of the `Topic` concept in PROTON and can be automatically organised in a `subTopic` hierarchy, thus allowing the PROTON ontology to grow during time. We use the WordNet Domains ontology instead and with a bottom-up approach such that an expansion of the domain labels assigned to relevant keywords extracted from metadata contents, and hence a subset of WordNet synsets, is applied by attaching super-domain labels to the lemmas considered. Moreover, WordNet has been translated into different languages thus providing multilingual facilities that can be easily integrated with our domains tagging procedure. Also, we use MANENT to harvest metadata records retrieved from more than 1,300 digital libraries spread all over the world. To the best of our understanding, Bloehdorn et al.'s infrastructure was used for annotating documents in the Digital Library of British Telecommunications only.

The PIRATES framework [6] is part of a semantic layer, included in a service-oriented architecture, providing primitive services to the applications built on top of the digital library which communicates with. The framework provides primitive services to automatically classify, annotate and recommend specific content using techniques based on natural language processing. A controlled, ontology-based vocabulary, is used to classify documents as result of the automatic tagging process. The PIRATES framework is still a theoretical model, although a prototype version has been already developed. We are aware neither of a massive use of PIRATES on a large number of digital libraries, nor of an experimental evaluation as the one we performed.

[21] `http://proton.semanticweb.org`.

5.2 Related Work on WordNet Domains

Gliozzo and Strapparava [1, 19] have built a Domain Model on top of WordNet Domains and have exploited it in several Natural Language Processing tasks in order to provide a topic similarity measure to documents. Their WordNet Domains Model (WND_{DM}) is formed by the set of all WordNet Domains (WND) and the set of WordNet synsets in WordNet ($Syns_c$). In this model a function applied to each element of $Syns_c$ results in a subset of WND (that we can call WND_c) associated with that element whereas a domain relevance function is able to return a real value for each element of WND_c, which represents the relevance (*rel* from now on) of each domain label in WND_c. In this model all the senses for a word w can be viewed as a subset of $Syns_c$ and hence as a union of all those WND_c associated with each element of this subset (and we can call this union $Syns_{cw}$). The domain relevance function for each w in WND_{DM} is computed as the averaged sum of all the *rel* calculated on $Syns_{cw}$.

The Domain Model explained above has been instantiated for defining the "Domain Driven Disambiguation" (DDD) methodology [28, 19]. This method has the peculiarity to apply disambiguation level to a domain and hence it is interesting in all those tasks where the results do not need to be as fine grained as a word level disambiguation task requires.

This methodology is based on Domain Vectors (DVs) that represent the domain relevance of an object with respect to a Domain Space, which is a vectorial space. Each value of the vector is an estimate of domain (a dimension) in the Domain Space.

Given a target word w to be disambiguated, DDD is computed on every single term of a context window CW surrounding the target term w and gives a score to each possible sense s of the term w. The building blocks of the score computation function are the DV for w, represented by the relevance values computed by the functions of the WND_{DM} and the DV for CW, which are computed according to the Domain Relevance Estimation technique in [20]. Moreover, a prior probability function, relative to a specific distribution of sense s in a reference corpus, is used as a smoothing parameter. The final result undergoes a fixed threshold in order to filter out not relevant outcomes.

A work similar to ours is that of [15] where the DOMINUS framework is described and an approach for tagging documents with WordNet and WordNet Domains is carried out in order to provide text categorisation and extraction of relevant keywords. Stemming from the document structured elements, such as title, abstract, body and bibliography, the authors exploit some natural language processing steps and extract WordNet synsets and the domain labels associated with them. With the aid of a density function based on Naïve Bayes they assign weights to synsets, varying upon the structured element just considered, and hence propagate them to domain labels in order to obtain both a classification of the document by topic and a raked list of the keywords that best represent the document content. The main differences with our work from the point of view of the local classification procedure are that we use metadata contents instead of document contents and we exploit the

WordNet Domains as an ontology in order to augment the domain label associated with WordNed synsets with super domains label, hence automatically structuring a hierarchical classification service. The main differences from the point of view of the global procedure are that we operate on documents metadata of different existing digital libraries all over the world obtained through our metadata harvesting service.

6 Conclusions and Future Work

In this chapter we described MANENT, an infrastructure for integrating, structuring and searching Digital Libraries. The technologies and standards enabling the realisation of infrastructures of this kind have been reviewed, as well as the related work.

The experimental results we obtained by evaluating the reliability of MANENT automatic classification of 100 records drawn from libraries spread all over the world are extremely encouraging.

Besides completing and testing the implementation of all the MANENT services, in order to release a working prototype in mid 2011, there are some improvements that will drive our medium-term future work:

- **Integration of different metadata formats**: as anticipated in Section 3, we provide the user the means for defining her own mappings from any metadata format to our EAD ontology. However, we plan to provide a set of already defined, built-in mappings from the most commonly used metadata formats. We would also like to provide a visualisation service for showing our built-in mappings and allowing users to define their own in a graphical and intuitive way.
- **Integration of a community in MANENT**: in the same way as we already foresee an active involvement of users in the definitions of mappings between metadata schemas, we would like to extend the user involvement to any service provided by MANENT. Becoming more and more similar to a community, MANENT, besides containing documents, should also "contain persons". Communities of practice are characterised by homogeneous patterns of experiences. Digital libraries are the results of the interplay between such implicit knowledge and the explicit layer that they incorporate. In a knowledge society a swift and efficient access to all of the knowledge devices, either people or artifacts, is a pressing requirement. Linking people to collections, enabling communities to annotate digital objects in MANENT and providing a working space for users to collect, organise and annotate such digital objects is the social dimension that, besides the organisational one, should become MANENT's most characterising feature.
- **Integration of digital objects in the knowledge base**: the MANENT prototype stores digital objects in the filesystem. In the near future a mechanism for storing them in the knowledge base should be provided. The idea is to exploit the METS [30] metadata format, after a proper conversion in OWL, to describe such resources in the light of the MANENT rationale, which keeps the collections

descriptions separated from that of single resources. This would also enable the straightforward provision of MANENT contents via OAI-PMH.
– **Classification of multilingual resources**: as an extension of our automatic classification service a study for providing classification of multilingual contents is on its way. Our choice of adopting a WordNet based approach has also been driven by its already available multilingual versions provided by worldwide institutions [49].

References

1. Agirre, E., Edmonds, P.: Word Sense Disambiguation - Algorithms and Applications. Springer, Heidelberg (2007)
2. Agosti, M., Berretti, S., Brettlecker, G., del Bimbo, A., Ferro, N., Fuhr, N., Keim, D., Klas, C.P., Lidy, T., Milano, D., Norrie, M., Ranaldi, P., Rauber, A., Schek, H.J., Schreck, T., Schuldt, H., Signer, B., Springmann, M.: DelosDLMS - the integrated DELOS digital library management system. In: Proceedings of the First International Conference on Digital Libraries: Research and Development, pp. 36–45 (2007)
3. Agosti, M., Ferro, N.: A Formal Model of Annotations of Digital Content. ACM Trans. Inform. Syst., 26(1) (2007)
4. Balasubramanian, N., Allan, J., Croft, W.B.: A comparison of sentence retrieval techniques. In: Proceedings of the Thirtieth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 813–814 (2007)
5. Banerjee, S., Pedersen, T.: An Adapted Lesk Algorithm for Word Sense Disambiguation Using WordNet. In: Gelbukh, A. (ed.) CICLing 2002. LNCS, vol. 2276, pp. 136–145. Springer, Heidelberg (2002)
6. Baruzzo, A., Casoto, P., Challapalli, P., Dattolo, A., Pudota, N., Tasso, C.: Toward Semantic Digital Libraries: Exploiting Web2.0 and Semantic Services in Cultural Heritage. Journal of Digital Information 10(6) (2009)
7. Bentivogli, L., Forner, P., Magnini, B., Pianta, E.: Revising WordNet Domains Hierarchy: Semantics, Coverage, and Balancing. In: Proceedings of the Twenty-First International Conference on Computational Linguistics (COLING 2004),, pp. 101–108 (2004)
8. Bloehdorn, S., Cimiano, P., Duke, A., Haase, P., Heizmann, J., Thurlow, I., Völker, J.: Ontology-based question answering for digital libraries. In: Kovács, L., Fuhr, N., Meghini, C. (eds.) ECDL 2007. LNCS, vol. 4675, pp. 14–25. Springer, Heidelberg (2007)
9. Buitelaar, P., Cimiano, P., Frank, A., Hartung, M., Racioppa, S.: Ontology-based information extraction and integration from heterogeneous data sources. Int. J. Hum.-Comput. Stud., 66(11), 759–788 (2008)
10. 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. ISTI-CNR, PISA (2007)
11. Cavnar, W.B., Trenkle, J.M.: N-Gram-Based Text Categorization. In: Proceedings of the Third Annual Symposium on Document Analysis and Information Retrieval, pp. 161–175 (1994)
12. Dublin Core Metadata Element Set, `http://www.dublincore.org/documents/dces/`
13. EAD: Encoded Archivial Description, `http://www.loc.gov/ead/`

14. EAD XML Metaschema, http://www.loc.gov/ead/ead.xsd
15. Ferilli, S., Biba, M., Basile, T., Esposito, F.: Combining Qualitative and Quantitative Keyword Extraction Methods with Document Layout Analysis. In: Proceedings of the Fifth Italian Research Conference on Digital Libraries (IRCDL 2009). DELOS: an Association for Digital Libraries (2009)
16. Ferro, N.: Annotation search: The FAST way. In: Agosti, M., Borbinha, J., Kapidakis, S., Papatheodorou, C., Tsakonas, G. (eds.) ECDL 2009. LNCS, vol. 5714, pp. 15–26. Springer, Heidelberg (2009)
17. Ferro, N., Silvello, G.: The NESTOR framework: How to handle hierarchical data structures. In: Agosti, M., Borbinha, J., Kapidakis, S., Papatheodorou, C., Tsakonas, G. (eds.) ECDL 2009. LNCS, vol. 5714, pp. 215–226. Springer, Heidelberg (2009)
18. FOAF: Friend of a Friend ontology, http://www.foaf-project.org/
19. Gliozzo, A., Strapparava, C.: Semantic Domains in Computational Linguistics. Springer, Heidelberg (2009)
20. Gliozzo, A., Strapparava, C., Dagan, I.: Unsupervised and supervised exploitation of semantic domains in lexical disambiguation. Computer Speech & Language 18(3), 255–299 (2004)
21. Gruber, T.: Definition of Ontology. In: Liu, L., Özsu, M.T. (eds.) Encyclopedia of Database Systems, Springer, Heidelberg (2009)
22. Hargittai, E., Fullerton, F., Menchen-Trevino, E., Thomas, K.: Trust Online: Young Adults' Evaluation of Web Content. International Journal of Communication 4, 468–494 (2010)
23. Hunter, J., Khan, I., Gerber, A.: Harvana: harvesting community tags to enrich collection metadata. In: Proceedings of the Eighth ACM/IEEE-CS Joint Conference on Digital Libraries, pp. 147–156 (2008)
24. Itzcovich, O.: L'uso del calcolatore in storiografia, Milano (1993)
25. Kruk, S.R., McDaniel, B.: Semantic Digital Libraries. Springer, Heidelberg (2009)
26. Locoro, A.: Tagging Domain Ontologies with WordNet Domains: An Approach for Fostering Ontology Classification, Engineering and Matching. Technical Report DISI-TR-10-10, CS Dept. of Genova University (2010), http://www.disi.unige.it/person/LocoroA/download/DISI-TR-10-10.pdf
27. Magnini, B., Cavagliá, G.: Integrating Subject Field Codes into WordNet. In: Proceedings of the Second International Conference on Language Resources and Evaluation (LREC 2000), pp. 1413–1414 (2000)
28. Magnini, B., Strapparava, C., Pezzulo, G., Gliozzo, A.: The role of domain information in Word Sense Disambiguation. Natural Language Engineering 8, 359–373 (2002)
29. MARCXML, http://www.loc.gov/standards/marcxml/
30. METS: Metadata encoding and Transmission Standard, http://www.loc.gov/standards/mets/
31. Metzler, D., Dumais, S.T., Meek, C.: Similarity measures for short segments of text. In: Amati, G., Carpineto, C., Romano, G. (eds.) ECiR 2007. LNCS, vol. 4425, pp. 16–27. Springer, Heidelberg (2007)
32. Mihalcea, R., Corley, C., Strappavara, C.: Corpus-based and Knowledge-based Measures of Text Semantic Similarity. In: Proceedings of the Twenty-First National Conference on Artificial Intelligence and Eighteenth Innovative Applications of Artificial Intelligence Conference. AAAI Press, Menlo Park (2006)
33. Miller, G.A.: WordNet: A Lexical Database for English. Communications of the ACM 38(11), 39–41 (1995)

34. OAI-PMH: Open Archives Initiative Protocol for Metadata Harvesting, http://www.openarchives.org/OAI/openarchivesprotocol.html
35. Ortoleva, P.: Persi nella rete? Circolazione del sapere storico. In: Soldani, S., Tomassini, L. (eds.) Storia & Computer, alla ricerca del passato con l'informatica, Milano (1996)
36. The Open Archives Initiative Protocol for Metadata Harvesting: Metadata Prefix and Metadata Schema, http://www.openarchives.org/OAI/openarchivesprotocol.html#MetadataNamespaces
37. The Open Archives Initiative Protocol for Metadata Harvesting: Guidelines for Repository Implementers, http://www.openarchives.org/OAI/2.0/guidelines-repository.htm
38. The Protégé Ontology Editor, http://protege.stanford.edu/
39. Rocchio, J.: Relevance feedback in information retrieval. In: Salton, G. (ed.) The SMART retrieval system: Experiments in Automatic Document Processing. Prentice-Hall, Englewood Cliffs (1971)
40. Rowland, R.: L'informatica e il mestiere dello storico. In: Quaderni Storici, pp. 26–78 (1991)
41. Salton, G., Lesk, M.: Computer evaluation of indexing and text processing. Journal of the ACM (JACM) 15(1), 8–36 (1968)
42. SPARQL Query Language for RDF, http://www.w3.org/TR/rdf-sparql-query/
43. Wache, H., Vögele, T., Visser, U., Stuckenschmidt, H., Schuster, G., Neumann, H., Hübner, S.: Ontology-based integration of information - a survey of existing approaches. In: Proceedings of the Twelfth International Joint Conference on Artificial Intelligence (IJCAI 2001) Workshop on Ontologies and Information Sharing, pp. 108–117 (2001)
44. Wenger, E.: Communities of practice, learning, meaning and identity, Cambridge (1998)
45. W3C . OWL Web Ontology Language Overview – W3C Recommendation (February 10, 2004)
46. W3C . RDF Vocabulary Description Language 1.0: RDF Schema – W3C Recommendation (February 10, 2004)
47. W3C . RDF/XML Syntax Specification (Revised) – W3C Recommendation (February 10, 2004)
48. W3C . Extensible Markup Language (XML) 1.0 (Fifth Edition) – W3C Recommendation (November 26, 2008)
49. Wordnets in the world, http://www.globalwordnet.org/gwa/wordnet_table.htm