Integrated Access to Legal Literature through Automated Semantic Classification

E. Francesconi, G. Peruginelli
Institute of Legal Theory and Techniques, Italian National Research Council (ITTIG-CNR), Florence, Italy

Abstract. Access to legal information and, in particular, to legal literature is examined for the creation of a search and retrieval system for Italian legal literature. The design and implementation of services such as integrated access to a wide range of resources are described, with a particular focus on the importance of exploiting metadata assigned to disparate legal material. The integration of structured repositories and Web documents is the main purpose of the system: it is constructed on the basis of a federation system with service provider functions, aiming at creating a centralized index of legal resources. The index is based on a uniform metadata view created for structured data by means of the OAI approach and for Web documents by a machine learning approach, which, in this paper, has been assessed as regards document classification. Semantic searching is a major requirement for legal literature users and a solution based on the exploitation of Dublin Core metadata, as well as the use of legal ontologies and related terms prepared for accessing indexed articles have been implemented.

Keywords: Semantic Web, Legal information retrieval, Machine learning, Document classification.

1. Introduction

Access to legal information is a fundamental democratic right to be guaranteed to citizens (*ignorantia legis non excusat*) in its constituent parts.
Legal literature[1] in particular is of primary importance in legal research, since its specific function is to enable the interpretation and distribution of legislation and jurisprudence. It consists of an abundant, high quality output of printed material and a certain amount of electronic contributions.
The majority of legal literature is still in traditional format and mainly scholars and legal professionals rely on commercial and academic publishers for their outputs. However electronic resources are now starting to be produced: the current scenario of access to electronic legal literature, with the increased provision on the net of disparate legal Web

[1] Legal literature in Civil law systems consists in legal intellectual outputs published in monographs, journal articles, manuals, grey literature, proceedings, etc.

resources, presents new opportunities and challenges, as well as problems. There is widespread availability and quick access to specialised databases, bibliographic catalogues, Web sites and individual contributions, but there are also difficulties, mainly due to differences in user interfaces for accessing this type of material. The extreme variety in classification systems and an ever-growing amount of uncontrolled electronic resources are additional issues to cope with.
In Italy this situation has given rise to a national project called NormeInRete[2], the objective of which is the consistent retrieval of national legislation, jurisprudence and literature. At present this project's focus is on legislative resources and until now there are no plans to extend it to case-law texts and literature. In this context, a feasibility study for a project focusing on Italian legal literature has been launched by ITTIG-CNR, the Institute of Legal Information Theory and Techniques. The project attempts to offer a unified point of access to multiple legal literature resources, by exploiting metadata and by providing tools for the discovery, selection and use of relevant legal materials.
This paper shows a possible approach for implementing a vertical portal, endowed by a semantic search engine, which has been called CALLIOPE (ClAssified Legal LIterature OPen access Engine), concentrating on legal literature within which users are not only referred to relevant information sources, but are also provided with services and ready-made solutions which meet users' needs.

This paper is organized as follows: in Section 2 an overview of the purposes of the project is discussed; in Section 3 the peculiarities of legal users' needs in searching legal information and the services provided are considered. In Section 4 data sources to build up the CALLIOPE system are briefly introduced, as well as in Section 5 the metadata approach to federate them is described. In Section 6 the architecture of CALLIOPE is illustrated, while from Sections 7 to 12 the details of each component of such architecture are shown, along with experimental results on automatic document indexing. Finally, in Section 13 some conclusions are discussed.

2. A project for accessing Italian legal literature

The creation of a unified access system to legal literature is conceived as a way of exploiting well-established institutional tools and services, as well as other academic and commercial projects currently aimed at the collection, analysis and distribution of legal resources in Italy.

[2] Legislation on the Net `http://www.normeinrete.it`

Such an objective is to be pursued by a thorough analysis of legal user needs, a careful selection of resources and reliance on rich and consistent metadata. CALLIOPE will give access to different level of legal literature information (bibliographic references, TOCs, abstracts, full text).
The first obstacle to face is the retrieval of legal literature. This process is a long, hard one encompassing multiple types of material. There is in fact no sole information provider that users can point to during their search; different sources have to be identified and this requires services which will seek out and locate them.
In general, the main requirements for effectively accessing legal literature are similar to those governing other types of material. These are:

a) coverage, i.e. exhaustiveness, which nevertheless requires proper selection criteria;

b) currency as regards the production and updating of the legal material concerned;

c) high quality of indexing and retrieval services.

Quality here refers to richness of the semantics as precision in cataloguing. Relevance and precision are the main requirements in analysing legal literature in order to achieve consistency between the indexing language and the language used in the production of laws and case-law reports. In such direction particular attention has been addressed to semantic search services, as well as to the implementation of specific facilities to support legal users in semantic querying the system trying to better meet their information needs. This is a crucial issue to be considered in developing an information retrieval system; it has to be particularly taken care of in legal domain, since legal users, being specialists in the domain, usually have high level expectations from the services of an information retrieval system.
Hereby some difficulties in accessing legal literature, together with possible solutions, are pointed out.
Availability of documents
Integration between access to information and availability of documents is today an ever-increasing demand from users.
Different user interfaces
One crucial issue is the variety of search interfaces offered by specialised databases, catalogues, Web sites and the like. Very often differences in the way options are presented to users and differences in the terminology used by retrieval systems cause serious problems of disorientation

to legal users.

Identification of legal resources on the net

References to on-line legal literature are often intermixed in current Web sites with different legal sources, such as legislation and case-law reports. That is not a limit in itself, but it is in fact a drawback when such references are presented in a confused way.

Quality of electronic resources

Apart from a few online, peer-reviewed journals, quality control of the rest of resources is poor and their instability causes serious problems in accessing them. Another difficulty is due to the scarce availability and inconsistency of bibliographic descriptions and metadata, which hamper resource discovery and retrieval.

Delimitation of legal domain

Access services often point to material not strictly pertinent to legal matters, but concerning disciplines such as economics, sociology etc., without clear evidence of what is law-pertinent. Moreover, there is an uneconomic and needless overlapping of projects regarding the same branches of law, and these very rarely provide exhaustiveness, obliging users to search across different systems.

In order to overcome such difficulties, the careful selection and clear presentation of the various information resources available are pinpointed as necessary measures for the construction of the system. These resources include primary sources such as printed and electronic documents and Web sites, as well as secondary reference sources such as OPACs[3], and indexing databases.

For this purpose a specialised retrieval system is being developed, designed in such a way as to be open to the contribution of authors and publishers in delivering and structuring their output, adopting descriptive and communication standards that allow interoperability.

3. Search and retrieval of legal literature

Users of legal literature share the characteristics, attitudes and needs of other users in seeking information, but they also have some peculiarities due to the sophisticated nature of legal information. In particular, they belong to different categories, professional and non-professional, with different skills in using bibliographic and indexing tools. They have different interests according to their profession and specialisation, which bring them to make various uses of legal literature materials. In particular academics make quite an extensive use of legal literature documents

[3] On-line Public Access Catalogues

for their teaching and research activities, followed by students preparing their thesis and by law professionals. Lawyers, judges, administrators and ordinary citizens usually start from statutes and case-law reports as their primary sources, and later search for specific legal literature items, showing a special interest in reference mechanisms allowing access to legislation and jurisprudence.
Results of a recent survey on legal users of the DoGi Database[4] have helped in understanding their behavior in seeking information and their purpose for searching. The aim of the on-line questionnaire was to identify user profile in order to plan services meeting their requirements. The results of the questionnaire show what users would like to expect from a legal literature access service. They essentially demand access to bibliographic references to both printed and online legal literature by using a single point of access. Desired services include accurate assistance during their search session, as well as access to documents and effective use of the selected material.
Furthermore legal users are mostly interested in subject access facilities. Legal concepts are generally expressed both in natural and technical language and for this reason they must be provided with adequate semantic tools helping them to contextualize information, and enhance searching performances.
Legal users should be provided with semantic tools (controlled vocabulary and automated indexing tools) enabling adjustment and reformulation of their information needs, helping them to better identify their requirements and consequently expanding their queries.

4. The data sources

As discussed in Section 3, what is needed for legal users is an integrated access to individual contributions hosted in different repositories, to articles appearing both in printed and online journals (analysed using different subject systems), to monograph bibliographic references, as well as to specialised Web sites which need high quality indexing in order to allow users to select and retrieve such information. The need for a uniform metadata format has led to the choice of the Dublin Core (DC) metadata set, in its XML version, as the target bridging format.

[4] The DoGi database (`http://nir.ittig.cnr.it/dogiswish/Index.htm`), is, in the Italian legal landscape, one of the most precious sources for legal literature research. It is a database created in 1970, offering abstracts of articles published in the most important legal periodicals (more than 250). Its main goal is to provide law scholars and professionals with exhaustive and updated information as found in Italian law reviews.

As regards for example journal articles, we relied on the work done on DCMI Cite (`http://www.dublincore.org/groups/citation/`) (Apps, 2003) as the target format from the native DoGi Database.
The CALLIOPE retrieval system under development is intended to deal with a variety of data, that can be divided into two main classes:

1. structured data coming from bibliographic repositories in libraries, which use a specific metadata format;
2. Web documents, namely HTML semi-structured documents, that, in most cases, do not follow any particular metadata scheme, nor any reliable or uniform HTML meta-tags, which could help the qualification of material of interest; such documents usually abound in plain text.

In the prototype of CALLIOPE, structured data are expected to be gathered from three data sources:

- DoGi: a metadata repository of articles related to legal literature, maintained by our Institute;
- University library OPACs;
- A publisher catalogue of bibliographic records, using a proprietary metadata format[5].

As far as the collection of Web documents is concerned, exploration of the Web began with a subset of sites of interest, selected by a group of legal experts, researchers and information professionals in legal literature. These resources can be used for two purposes:

- to train the software modules able to select and classify Web documents (Sections 8 and 9);
- to perform a selective exploration of the Web (Section 8), starting from such documents and following the hyperlinks with high probability of pointing to other relevant documents.

5. Metadata approach

In order to provide integration of different data sources, while offering a uniform view on them to the users by adopting the DC metadata set, two different approaches have been used according to the different nature of the data sources:

[5] Currently a study of a publisher metadata format is under analysis, therefore the related DC mapping is not described in this paper.

– for structured data, the metadata schemes supported by the selected repositories (so far based on UNIMARC format[6]) have been mapped to the DC metadata set. In particular an accurate analysis of the DCMI Cite has been carried out to describe DoGi records;

– for Web documents, a specific module generating a meaningful subset of DC metadata has been developed.

Regarding structured repositories, a preliminary effort is required by data providers in order to expose metadata, making repositories compliant to the metadata scheme of the system and ready to be harvested for the creation of a centralised index (Section 7).
In particular, as regards the mapping of DoGi records, the DCMI Cite is used for encoding bibliographic citations of the articles. The application adopted three distinct hierarchical levels: the journal, the journal issue and the individual article level. Some peculiarities of DoGi records were accommodated in a DC record, especially elements such as *dc:source*, *dc:relation* and *dc:type*. Regarding this latter tag, we used a list of specific DoGi document types, describing independent contributions, book reviews, seminar and workshop reports, comments on European or national case law and on laws or administrative acts.
With reference to Web documents, we decided to provide them with a uniform metadata scheme and chose a subset of the DC one. However, unlike the structured repositories, no specific effort is required from data providers to apply DC metadata to Web documents, since documents are collected with no preliminary agreement between data and service providers. Therefore, in our project, the metadata application to Web documents relies on the service provider, whose work is based on an automatic metadata generator which acts according to specific criteria, as described in Section 9.

6. The architecture of the CALLIOPE federation system

CALLIOPE is the result of a federation system with service provider functionalities, the architecture of which consists of five main modules (Fig. 1):

1. a metadata harvester, aimed at selecting and collecting metadata from structured data providers;

[6] Mapping Dublin Core/UNIMARC is based on tables prepared by ICCU, Rome: `http://www.iccu.sbn.it/Edubluni.htm`

2. a focused crawler, selecting and collecting semi-structured data, namely documents of interest from the Web;

3. an automatic metadata generator, supplying metadata to the selected Web documents;

4. an indexer of the selected data;

5. a user interface supporting a controlled vocabulary of legal terms and categories, able to help users in submitting a query meeting their information needs.

Hereinafter these five modules, their implementations, as well as the experimental results obtained for the implementation of the CALLIOPE prototype are described.

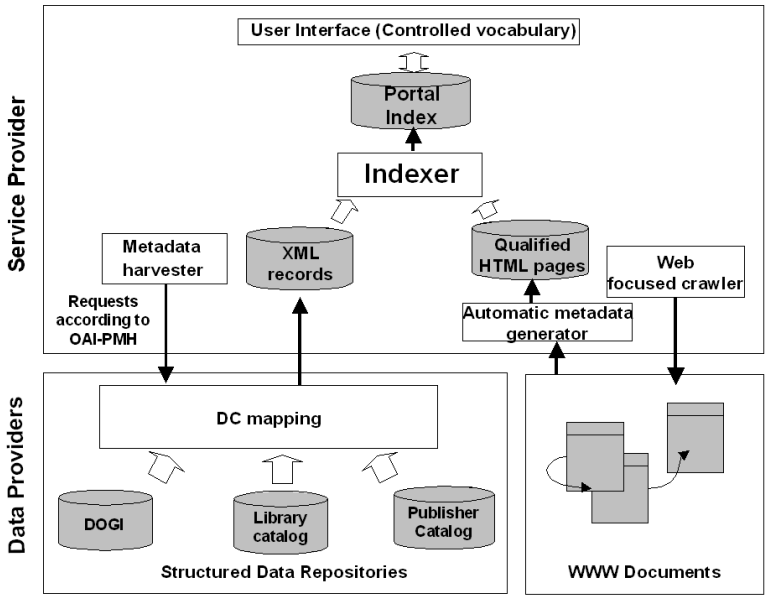

Figure 1. The architecture of the CALLIOPE federation system

7. Metadata harvester for structured resources

In order to select resources of interest from structured repositories, we considered the specific classification scheme of each data source.
Bibliographic metadata from structured repositories usually provide a classification scheme in accordance with the UDC (Universal Decimal Classification) and DDC (Dewey Decimal Classification) systems, as

well as proprietary classification schemes (as the DoGi one). Since the DoGi database contains only records on legal literature, no particular selection was necessary, therefore all DoGi records have been selected. Library records contain DDC entries. This means that the selection of the material is carried out using the 34 class and other additional classes from this code, as defined by the Italian DDC version. Only specific sections of these additional classes (reported in Tab. I) are considered in particular.

Table I. Selected classes out of 34 code

Dewey Codes	**Dewey Description**
262	Ecclesiology
306.1	Sociology of law
320	Political science
350	Public administration
364	Criminology
365	Penal & related institutions
614.1	Forensic medicine

Problems in selecting legal resources on the basis of DDC classification codes are mainly due to:

- Overlapping of heterogeneous legal sources.
  Traditional materials such as codes, commentaries and collections of law cases are difficult to isolate from legal literature.

- Multi-disciplinarity of legal literature.
  Legal literature can be separated into a number of classes and divisions which are not easily identifiable for the purposes of selection services. What is needed is an interpretation that can be based only on accurate work of intellectual selection.

For the purposes of CALLIOPE, a first effort has been made to start from the list of law faculty subjects, relating them to DDC classes.
In order to harvest data from structured repositories, we decided to use the OAI[7] approach. The OAI approach, mainly used in librarian information systems, establishes the entities (data and service providers), their roles and the format of the exchanged messages. The service provider is entitled to query data providers according to the requests

[7] Open Archives Initiative (`http://www.openarchives.org/OAI/openarchivesprotocol.htm`)

established by the OAI Protocol of Metadata Harvesting (OAI-PMH) (see Fig. 2 as a sketch of the OAI architecture) and metadata are harvested according to the Dublin Core (DC) scheme. Data providers are entitled to translate such requests according to their own query modalities and to translate the results according to the DC scheme in its XML version.

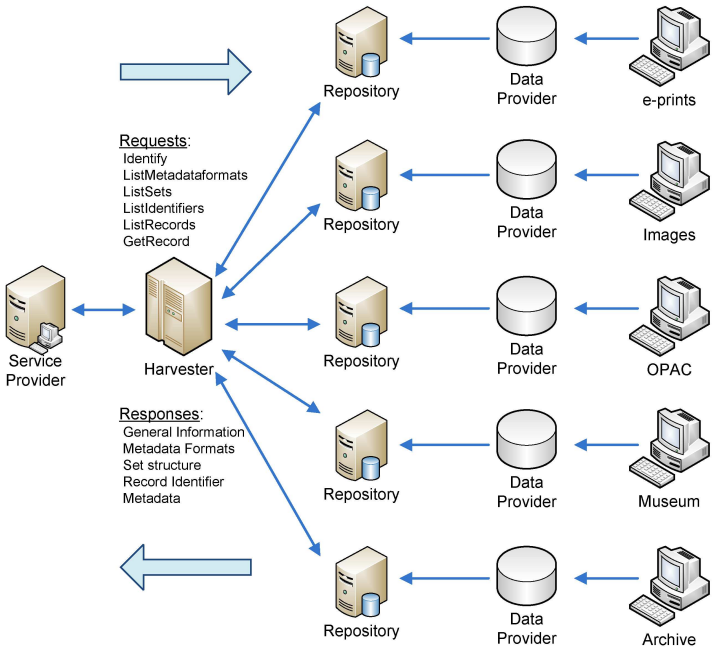

Figure 2. The OAI approach

Most of the projects using OAI protocols aims at constructing portals as federative architectures of structured data repositories (mainly bibliographic records or e-print resources).
The TEL project[8], for instance, aims to construct a co-operative model of architecture able to access the foremost national bibliographic repositories in Europe. CYCLADES[9] provides a collaborative, multidisciplinary, virtual archive service supporting scholarly communities in their work. The TORII portal[10], which also offers a cross-referencing service with different documents, is dedicated to those working on research into high-energy physics.

[8] TEL – The European Library (`http://www.europeanlibrary.org`)
[9] CYCLADES – An Open Collaborative Virtual Archive Environment (`http://www.ercim.org/cyclades/`)
[10] TORII - The Digital Research Community (`http://library.cern.ch/HEPLW/4/papers/4/`)

Table II. Mapping between DCMI Cite and DoGi.

DC metadata	**DoGi metadata**
dc:title	Titolo articolo
dc:creator	Autore
dc:contributor	Curatore
dc:subject (scheme=DoGi70-99)	Classificazione 70-99
dc:subject (scheme=DoGi)	Classificazione
dc:description	Sommario
dc:description	Riassunto
dc:source	Fonte del contributo *(Trib. Trieste 23 luglio 1999, Convegno internazionale su. . . , etc.)*
dc:relation	Fonti normative, giurisprudenziali, nazionali, internazionali, comunitarie, straniere, storiche e canoniche
dcterms:issued (scheme=W3CDTF)	Anno pubblicazione fascicolo
dc:type(scheme=MARCGenre)	Tipo documento *(Contributo indipendente, relazione, intervento a convegno, nota a sentenza, etc.)*
dc:publisher	Editore
Dcterms:citation (scheme=DCMICite)	
JournalTitle	Titolo rivista
JournalIssueNumber	Fascicolo e/o Supplemento rivista
JournalIssueDate (scheme Dogi)	Data fascicolo
Pagination	Pagine contributo
dc:language(scheme=RFC1766)	Lingua contributo originale
dcterms:isPartOf(scheme=URI)	Urn:ISSN: codice ISSN
dc:rights	ITTIG-CNR (DoGi database) Firenze Italia

Our project uses OAI-PMH to collect data from structured repositories. These are made to comply with the adopted protocol requests and the CALLIOPE metadata specifications (summed-up in Tab. II).

In the CALLIOPE prototype we have implemented an OAI data provider for the DoGi repository in its XML version, using a Java package[11] available on the Open Archives Initiative Web site[12]. The same Java package has been used to develop our module which implements the OAI-PMH. For the CALLIOPE prototype, metadata coming from the DoGi archive have been harvested at the service provider level in DC format.

[11] ARC developed by Digital Library Group, Old Dominion University
[12] `http://www.openarchives.org`

8. Focused crawler for Web documents

Legal literature documents available on the Web cannot be treated in the same way since, in most cases, no classification metadata is provided.
Selecting documents of interest on the Web represents a key issue in populating a domain-specific retrieval system. Contrary to general-purpose agents, which usually explore the hyperlinks of the Web with the aim of finding as many different documents as possible, in this project we are interested in selecting domain-specific documents and, therefore, only in following the hyperlinks which point to documents of interest for the system. In order to perform such a function we use the approach described in (McCallum et al., 2000), based on a policy aimed at following the links on the path which provide the closest and highest reward. Such a reward is obtained in terms of the probability of a link leading to a pertinent document.
In the implemented system, documents will be harvested using a crawler that selects documents by following the hyperlinks with a high probability of leading to documents of interest; such a probability is obtained by means of a procedure of text categorization (Sebastiani, 2002) on a set of words in the vicinity of the hyperlink. In this approach the problem of focused crawling is equivalent to a problem of document classification where the class of a target document is predicted by the context surrounding a corresponding hyperlink in a parent document (McCallum et al., 2000). Crawling is therefore reduced to a binary classification problem, where the intelligent spider should follow hyperlinks leading to documents of a topic area (in our case "legal literature"), avoiding to follow all the others.

This module is currently under implementation and testing, while in the CALLIOPE prototype a set of Web documents belonging to different categories of law are collected by legal experts, in order to train both the crawler and the automatic metadata generator, discussed in Section 9. The algorithms for document classification used to implement the automatic metadata generator are expected to be tested for the focused crawler as well, and they are more detailed in Section 9.

9. Automatic metadata generator

The application of metadata to Web documents is an issue deeply investigated in literature for the semantic Web construction. Some works aim in particular at evaluating the reliability of authors' own metadata generation, or of similar collaborative activities between authors and

Table III. The DC metadata assigned to Web documents and the corresponding methods used for their automatic generation.

DC metadata	**Criteria of the DC metadata automatic generation for Web documents**
dc:identifier	The document url
dc:title	The content of html tag <title>
dc:date	The 'last-modified' date of the Web document
dc:subject	Automatic generation using a machine learning approach (see Section 9)
dc:description	The content of the html metatag "description", if any, otherwise the content of the html tag <body>
dc:type	"Web document"
dc:publisher	Web domain name

metadata experts (J. Greenberg, 2002).
Other services (`www.ukoln.ac.uk/metadata/dcdot/`) have proposed a different approach, aimed at integrating a service of automatic DC metadata generation, limited to reliable mapping with the ones originally included in Web documents, combined with the collaboration of the authors, who are requested to complete the metadata scheme.
Some other projects (`www.klarity.com.au`) have been carried out in order to automatically provide Web documents with metadata, on the basis of keywords supplied by the authors, thus providing uniform and consistent metadata for such documents.
On the basis of these experiences[13] and considering the aims of our experimentation, we decided to develop a module of automatic metadata generation providing documents with a subset of DC metadata. This module aims at supporting the intellectual activity of a service provider in organizing qualified access services on the Web. Once documents of interest have been selected from the Web, an automatic metadata generator is applied in order to provide documents with a subset of DC metadata, which have to be validated by human experts. Most of DC metadata are obtained by a simple mapping between particular document tags (as the `<title>` html-tag), properties of the documents (as the URL) or html-metatags (as meta="description"), if any, to DC metadata.

Tab. III summarizes the list of DC metadata applied to the selected Web documents and the criteria used to their generation.

[13] most of which is summarized at `http://www.lub.lu.se/tk/metadata/dctoollist.html`

Particular attention is addressed to document classification.
In order to provide documents with uniform and reliable *dc:subject* metadata, we cannot rely on the information provided by the authors in the html-metatags, unless they belong to a collaborative community of both authors and cataloguers (J. Greenberg, 2002). This, however, is not our case, since Web documents are selected without any preliminary accordance with authors or publishers. Therefore a specific approach, based on automatic criteria of classification, has been used.
The automatic Web document classifier implemented for CALLIOPE mainly consists of a text categorization algorithm which takes as input the plain text of a Web document d and outputs its predicted type (or "class") c choosing from a set of candidate classes $\mathcal{C}$. In order to perform such a function, it relies on a machine learning algorithm which has been trained on a set of training documents $\mathcal{D}$ with known class, and thus learned a model able to make predictions on new unseen documents.
Machine learning approaches have been widely applied in literature to automated text categorization (Sebastiani, 2002). In this work for the prototype of the system we have considered two different machine learning approaches to classify Web documents: a *Naïve Bayes* classifier, which has been shown to be effective for text categorization (Joachims, 1997), and compared it with a multiclass extension of the *Support Vector Machines*. Such algorithms have been recently applied in legal domain to classify legislative documents (Francesconi and Passerini, 2007) (Biagioli et al., 2005) and for legal text summarization (Hachey and Grover, 2005).
For the experiments we have used a standard implementation of the Naïve Bayes approach as described in (Sebastiani, 2002). As regards the Support Vector Machines (SVM) methodology, in literature different variants of multiclass extension of the one for binary classification tasks (Vapnik, 1998) (Cortes and Vapnik, 1995) (Burges, 1998) exist: as either combinations of binary classifiers, or by directly implementing a multiclass version of the SVM learning algorithm (MSVM) (Hsu and Lin, 2002). For our experiments we used a direct implementation of the multiclass version of the SVM learning algorithm (MSVM) as developed independently by Vapnik (Vapnik, 1998) and Crammer and Singer (Crammer and Singer, 2002).

As discussed, these methods have been tested to provide the selected Web documents with *dc:subject* metadata. Along with the methods reported in Tab. III the automatic metadata generator is able to qualify Web documents with a meaningful subset of DC metadata, to be validated afterwards by human experts.

10. The experiments on automatic metadata generator

For the CALLIOPE prototype an evaluation of the automatic metadata generator performaces has been carried on as regards the module able to provide automatic semantic classification to Web documents. In this section the steps needed to its implementation and testing are described.

10.1. Document representation

A number of alternatives are possible in order to represent a document in a format to be managed by an automatic classifier. As in (Francesconi and Passerini, 2007), we faced two main problems: the meaningful textual units, representing the atomic terms of the document, and the level of structure to be maintained when considering the combination of such terms. As regards the first problem the simplest possibility is that of representing words as terms (Buckley and Salton, 1988), (Biagioli et al., 2005). Concerning the second problem, we used an approach, usually followed in literature (Apté et al., 1994; Dumais et al., 1998), aiming at ignoring the sequential order of the terms within a given document and at representing a document as an unordered bag of terms.
A number of preprocessing operations have been further tested on pure words in order to increase their statistical qualities and reduce the computational complexity of the problem:

- digit characters can be represented using a special character;
- non alphanumeric characters can be represented using a special character as well.

Other preprocessing operations as stemming or the use of word stoplists (stopwords), in this phase, have been considered.

Once basic terms have been defined, a vocabulary of terms $\mathcal{T}$ can be created from the set of training documents $\mathcal{D}$, containing all the terms which occur at least once in the set. A single document d will be represented as a vector of weights $[w_1, \ldots, w_{|\mathcal{T}|}]$, where the weight w_i represents the amount of information which the i^{th} term of the vocabulary carries out with respect to the semantics of d. As in different types of weights, with increasing degree of complexity:

- a *binary* weight $\delta(w, d)$ indicating the presence/absence of the term within the document;
- a *term-frequency* weight $tf(w, d)$ indicating the number of times the term occurs within the document, which should be a measure of its representativeness of the document content;

– a *tfidf* weight which indicates the degree of specificity of the term with respect to the document ($tfidf(w,d) = tf(w,d) * \log(|D_w|^{-1})$ where $\mathcal{D}_w$ is the fraction of training documents containing at least once the term w (Sebastiani, 2002)).

Statistics computed for extremely rare terms will be far less reliable, thus possibly leading to *overfitting* phenomena. In order to address such a problem, *feature selection* techniques can be applied to reduce the number of terms to be considered, thus actually restricting the vocabulary to be employed (Sebastiani, 2002) (Yang and Pedersen, 1997). We tested two methods:

– an unsupervised *min frequency* threshold over the number of times a term has been found in the entire training set, aiming at eliminating terms with poor statistics;

– a supervised threshold over the *Information Gain* (Quinlan, 1986) of terms, which measures how much a term discriminates between documents belonging to different classes (Biagioli et al., 2005).

10.2. Automatic metadata generator implementation and testing

The two classification algorithms have been tested on a set of 2478 documents, belonging to the 11 classes shown in Tab. IV.[14]

Table IV. Classes and number of documents for each class in the experiments

Class labels	**Classes of the data set**	**Number of documents**
c_0	Environmental law	75
c_1	Administrative law	605
c_2	Constitutional law	132
c_3	Ecclesiastic law	34
c_4	European law	117
c_5	Computer Science law	221
c_6	International law	147
c_7	Labour law	256
c_8	Criminal law	298
c_9	Private law	430
c_{10}	Taxation law	163

The documents have been selected from a set of Web sites of interest by a group of ITTIG-CNR legal experts. This data set has been used both to train and to test the classifiers. First of all a pre-processing step has been carried out aiming at removing html tags and javascript code within the documents. Then a number of combinations of the document representation and feature selection strategies have been tried. The parameters used for document representation and feature selection, which gave the best results for the two classification methods on our dataset are reported in Tab. V. The first two rows represent possible pre-processing operations. The third row indicates the term weighting scheme employed. The two following rows are for feature selection strategies: the unsupervised minimum frequency and the number of terms to keep, after being ordered by Information Gain.
After having trained the Naïve Bayes classifier using the data set of Tab. IV, experiments have been carried out in order to validate such a learning procedure by evaluating the classification capability on the training set (*train accuracy*). The experiments produced the best results using the parameters of Tab. V (column NB), obtaining a train accuracy of 82.5%.

Table V. Parameters for document representation and feature selection which produced the best results with Naïve Bayes (NB) and MSVM classifiers.

Document representation and feature selection parameter	**NB**	**MSVM**
Replace digits	yes	no
Replace not alphanum. characters	yes	no
Term weighting scheme	tf	binary
Minimum frequency selection	2	0
Number of words selected with max Information Gain	all	1500

Then, using the same data set a MSVM classifier[15] has been trained and tested. The best results have been produced using the parameters of Tab. V (column MSVM), obtaining a train accuracy of 85.1%.
A comparison of the results of the two classifiers are shown in Fig. 3.

[14] Such classes, organized in a single-tier set only, have been chosen to test the approach. Possible extensions or hierarchical organization of the classes can be approached respectively by re-training the classifiers according to the new set of classes or using a set of classifiers hierarchically organized as classes are organised.

[15] We used the MSVM implementation at `http://www.csie.ntu.edu.tw/~cjlin/bsvm/index.html`

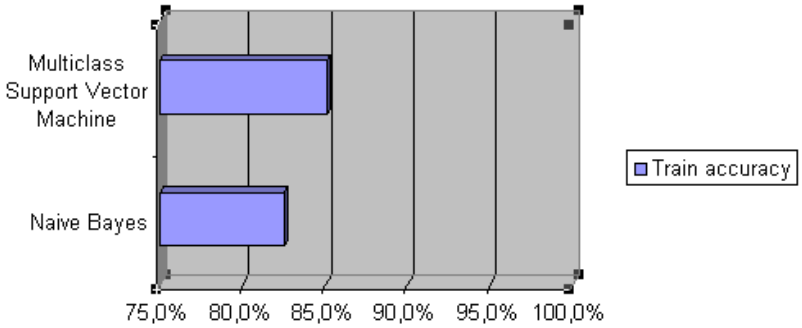

Figure 3. Naive Bayes and MSVM classifiers train accuracy.

The experiments showed that the MSVM classifier outperforms the Naïve Bayes one as regards the train accuracy (Fig. 3).
Chosen the MSVM classifier, its generalization capability has been tested using a *Leave One Out* (LOO) strategy.

Table VI. LOO results of the MSVM classifier.

	c_0	c_1	c_2	c_3	c_4	c_5	c_6	c_7	c_8	c_9	c_{10}
c_0	25	30	0	0	1	2	0	2	8	7	0
c_1	6	501	3	1	6	11	1	18	21	30	7
c_2	1	40	47	0	1	2	5	8	11	15	2
c_3	0	1	0	23	0	1	1	0	4	4	0
c_4	0	10	0	0	94	4	2	3	1	2	1
c_5	0	10	0	0	2	186	2	1	8	11	1
c_6	0	2	1	0	8	5	114	0	3	11	3
c_7	0	22	3	0	1	3	2	205	8	12	0
c_8	1	14	1	0	0	9	5	6	250	10	2
c_9	4	42	1	2	2	32	11	17	32	280	7
c_{10}	0	9	1	0	2	2	5	2	1	14	127

For a data set of n documents, the LOO strategy consists in n experiments where all the n examples of the data set are used to train the classifier, except for one different example at each run used to test. *LOO accuracy* measures the number of documents correctly classified with respect of all the entire number of tests. This test strategy is a possible solution to the problem of the availability of a meaningful set of documents both for training and testing the system.
These experiments gave a LOO accuracy of 74.7% for the MSVM classifier. Tab. VI shows the MSVM classification details for individual classes: each entry (c_i,c_j) represents the number of documents of class c_i classified in class c_j.
Such results are encouraging for the aim of providing CALLIOPE with

an automatic tool to support the human activities of organising and classifying documents.

11. Indexer of selected resources

After having harvested metadata from structured repositories using OAI-PMH (Section 7) and selected Web documents as well as qualified them with metadata through the support of the automatic metadata generator (Section 8 and 9), two collections are obtained, containing data in different formats (XML records and HTML documents), sharing the same metadata scheme.
At this stage an indexing procedure has been implemented, aimed at providing a uniform view and integrated access to the data of the retrieval system. In our experiments we have used an indexer[16] providing the possibility of indexing both HTML documents, according to their meta-tags, and XML documents according to their metadata.
The indexer works on files as a unit of indexing. HTML documents are stored as files named using their URL, which represents *dc:identifier* metadata; therefore they are ready to be indexed. On the other hand, the stream of XML records, coming from structured repositories, has been organised in files, so each XML record is stored in a file named according to *dc:identifier*.
The indexer works separately on the two collections of files. At the end of the indexing phase we obtain two indexes following the same metadata scheme. The two indexes are then merged in a single index, using the utilities of the chosen indexer; such single index represents the CALLIOPE index.
In the search phase, data coming from structured repositories and Web documents are requested from the index according to the same DC metadata scheme; in the retrieval phase data are identified by the content of the *dc:identifier*, and retrieved accordingly.

12. User interface supporting a legal controlled vocabulary

The CALLIOPE user interface is provided with facilities able to help users in submitting queries expressing the semantics of their information needs. In the system architecture, users access a DC metadata index (Fig. 1) as well as a full text collection (Moens, 2005) with two query modalities (Francesconi and Peruginelli, 2004):

[16] Swish-e, Simple Web Indexing System for Humans – Enhanced (`http://swish-e.org`)

1) advanced search: *metadata-based document querying* (MBDQ);

2) simple search: *keyword-based document querying* (KBDQ).

Case 1): in this case users submit a query using DC metadata fields. Query terms are required to match DC metadata contents.

Case 2): in this case users submit a query, using an unqualified text box. Query terms are required to match metadata contents or the full text.

Facilities on user side are also desirable to expand a query so that users are supported in expressing their information needs. Such facilities are provided using the DoGi legal specialised controlled vocabulary which allows to guide users in formulating a query. This Italian legal classification system is in fact the one adopted for indexing DoGi documents[17]. The indexing language is a controlled one and it is based on the areas of law as structured in the Italian law faculty scheme. Such classification is a valid tool not only for retrieving legal literature items in the DoGi database, but also for an in-depth understanding of the structure of Italian law. There are 24 areas of law considered, each designated by a code. The classification scheme is hierarchically structured (up to three levels) and it is composed by alphanumeric codes expressing specific concepts. Codes are associated with descriptors (6600 at the moment). An authority list of descriptors is maintained and updated on the basis of indexers' suggestions, as well as of statistic analysis of searches made by users.

Using the KBDQ query modality for example[18], the user interacts with the DoGi vocabulary choosing terms to compose a query, thus identifying the legal categories such terms are related to. Moreover, to maintain the selectivity of the query the user is asked to choose one of such legal categories by the (*dc:subject*) field, thus contextualising the query itself. Documents matching both terms and legal category are selected by the system.

For example, let us consider a query searching for documents of the legal category "criminal law" where the term "recklessness" is included. If relevant documents for the query do not contain the term "recklessness" a 'no hits' response is given back. In case no documents match the chosen terms, the system tries to expand the query pointing to narrower or broader terms, according to the DoGi controlled vocabulary. Such terms, combined with the legal category, are used to query again the CALLIOPE index.

This procedure allows:

[17] see `http://nir.ittig.cnr.it/dogiswish/consistenze/class2000Eng.htm`

[18] but similar arguments can be provided for the MBDQ modality

– to retrieve relevant documents for the user information needs, even if such documents do not contain the terms firstly chosen by the user;

– to retrieve documents containing query terms and belonging to a specific legal category.

This part of the system is at a preliminary stage: in particular an additional evaluation of such query strategies using typical precision/recall measures on a predefine set of queries is expected to be carried out in future works.

13. Conclusions

In this paper a possible solution for the creation of CALLIOPE, a retrieval system for Italian legal literature, based on protocols of metadata harvesting and AI techniques for document classification and indexing, has been presented. The software architecture aims at integrating structured resources and Web documents pertaining to legal literature into a unified point of access and a uniform view on data. The creation of an architecture able to harmonize different metadata scheme is the most effective way to bridge the gap between different data formats. The selection of relevant material and metadata production are burdensome activities, particularly labour intensive when collecting legal resources. For structured resources a thorough analysis and comparison of different classification systems have been carried out. Web documents, on the other hand, are not usually supplied with metadata following particular schemes and, where available, they are generally not reliable. In order to supply Web documents with metadata, an automatic metadata generator module based on a machine learning approach has been developed. This module aims at supporting the intellectual activity of a service provider in its work of organizing Web documents.
CALLIOPE is, therefore, the result of a federation system which combines the harvesting of structured data using OAI-PMH methodology and the selection and qualification of Web documents through the support of automatic tools.
In particular the current focus of the project is to achieve a high quality retrieval of legal information. CALLIOPE aims at providing a single point of access into disparate repositories where categories of law, as content of *dc:subject* are the essential metadata to point to relevant legal literature documents. Facilities in query formulation are given

to the users through the exploitation of a legal controlled vocabulary, improving retrieval of legal resources.

Acknowledgements

Special thanks go to dr. Anna Archi, senior researcher at ITTIG-CNR, who dedicated her research work to services for retrieving legal literature.

References

Apps, A.: 2003, 'A Journal Article Bibliographic Citation Dublin Core Structured Value'. Retrieved on May 2, 2003 from (http://epub.mimas.ac.uk/DC/citdcsv.html).

Apté, C., F. Damerau, and S. Weiss: 1994, 'Automated learning of decision rules for text categorization'. *ACM Transactions on Information Systems* **12**(3), 233–251.

Biagioli, C., E. Francesconi, A. Passerini, S. Montemagni, and C. Soria: 2005, 'Automatic semantics extraction in law documents'. In: *Proceedings of International Conference on Artificial Intelligence and Law*. pp. 133–139.

Buckley, C. and G. Salton: 1988, 'Term-weighting approaches in automatic text retrieval'. *Information Processing and Management* **24**(5), 513–523.

Burges, C.: 1998, 'A tutorial on support vector machines for pattern recognition'. In: *Data Mining and Knowledge Discovery*. Boston: Kluwer Academic Publishers. (Volume 2).

Cortes, C. and V. Vapnik: 1995, 'Support Vector Networks'. *Machine Learning* **20**, 1–25.

Crammer, K. and Y. Singer: 2002, 'On the algorithmic implementation of multiclass kernel-based vector machines'. *Journal on Machine Learning Research* **2**, 265–292.

Dumais, S., J. Platt, D. Heckerman, and M. Sahami: 1998, 'Inductive learning algorithms and representations for text categorization'. In: *CIKM '98: Proceedings of the seventh international conference on Information and knowledge management*. pp. 148–155.

Francesconi, E. and A. Passerini: 2007, 'Automatic Classification of Provisions in Legislative Texts'. *International Journal on Artificial Intelligence and Law* **15**(1), 1–17.

Francesconi, E. and G. Peruginelli: 2004, 'Opening the Legal Literature Portal to Multilingual Access'. In: *Proceedings of the Dublin Core Conference*. pp. 37–44.

Hachey, B. and C. Grover: 2005, 'Automatic Legal Text Summarisation: Experiments with Summary Structuring'. In: *Proceedings of International Conference on Artificial Intelligence and Law*. pp. 75–84.

Hsu, C.-W. and C.-J. Lin: 2002, 'A comparison of methods for multi-class support vector machines'. *IEEE Transactions on Neural Networks* **13**(2), 415–425.

J. Greenberg, W. R.: 2002, 'Semantic Web Construction: An Inquiry of Authors' Views on Collaborative Metadata Generation'. In: *Proceedings of the International Conference on Dublin Core and Metadata for e-Communities*. pp. 45–52.

Joachims, T.: 1997, 'A Probabilistic Analysis of the Rocchio Algorithm with TFIDF for Text Categorization'. In: *Proceedings of the Fourteenth International Conference on Machine Learning*. pp. 143–151, Morgan Kaufmann Publishers Inc., San Francisco, US.

McCallum, A., K. Nigam, J. Rennie, and K. Seymore: 2000, 'Automating the Construction of Internet Portals with Machine Learning'. In: *Information Retrieval Journal*. pp. 127–163.

Moens, M.-F.: 2005, 'Combining Structured and Unstructured Information in a Retrieval Model for Accessing Legislation'. In: *Proceedings of International Conference on Artificial Intelligence and Law*. pp. 141–145.

Quinlan, J.: 1986, 'Inductive Learning of Decision Trees'. *Machine Learning* **1**, 81–106.

Sebastiani, F.: 2002, 'Machine Learning in Automated Text Categorization'. *ACM Computing Surveys* **34**(1), 1–47.

Vapnik, V.: 1998, *Statistical Learning Theory*. New York: Wiley.

Yang, Y. and J. Pedersen: 1997, 'A Comparative Study on Feature Selection in Text Categorization'. In: *Proceedings of the Fourteenth International Conference on Machine Learning*. pp. 412–420, Morgan Kaufmann Publishers Inc.