DCPAPERS

Access to Italian legal literature: Integration between Structured Repositories and Web Documents

E. Francesconi, G. Peruginelli
ITTIG – Institute of Legal Information Theory and Technologies
Italian National Research Council , Italy
{francesconi, peruginelli}@ittig.cnr.it

Abstract

The problems of accessing legal information and, in particular, legal literature are examined in conjunction with the creation of a portal to Italian legal doctrine. The design and implementation of services such as integrated access to a wide range of resources are described, with a focus on the importance of exploiting metadata assigned to disparate legal material. On the basis of the results of a survey of legal users' requirements, the main features of the planned system are presented: accurate selection of resources, user profiling and assistance in the search process, reliance on rich and consistent metadata, navigation facilities for legal literature, legislation and case-law sources.

The strategies devised have been experimented, such as the mapping of both the UNIMARC format and a proprietary citation format for journal articles to the Dublin Core unqualified metadata set and to DCMI cite. These formats are extracted respectively from OPACs and from a well-established legal bibliographic database, DoGi. Specific semantic problems of legal literature indexing are tackled, mainly concerning classification systems.

Similar issues are present in legal doctrine available on the web. In particular web documents usually lack metadata and, where present, they do not usually conform to well-established metadata patterns.

A research study was therefore carried out, followed by the creation of a prototype and testing with the aim of automatically providing web documents with DC metadata, thereby supporting the intellectual activity of a service provider in organizing qualified access services to web documents.

The integration of structured repositories and web documents is the main purpose of the portal: it is constructed on the basis of a federation system with service provider functions, aiming at creating a centralized index of such 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.

Keywords*: legal literature, integration of resources, Dublin Core, DCMI Cite, OAI-PMH, automatic generation of metadata, federation architecture.*

1. Introduction

Access to legal information is a fundamental democratic right to be guaranteed to citizens (*ignorantia legis non excusat*). Legal information consists of laws and rules, case law and legal literature.

Legal literature in particular is of primary importance in legal research, since its specific function is to enable the interpretation and distribution of legislation and jurisprudence.

Italian legal doctrine, like that of other countries, consists of an abundant, high quality output of printed material and a certain amount of electronic contributions, where traditional material is mainly provided by commercial and long-established publishers, as well as by scholars and professionals. Secondary sources of printed literature are, in general, offered free of charge and include different types of information: a) bibliographic references to collections and holdings in Italian libraries; b) TOCs prepared by indexing services (free or for a fee); c) abstracts provided by documentation centres, such as those produced by the Istituto di Teoria e Tecniche per l'Informazione Giuridica del C.N.R - ITTIG (Institute of Legal Information Theory and Technologies of the National Research Council) which produces the DoGi – Dottrina Giuridica database.

Electronic resources are now starting to be produced, while high-quality literature is still in print. The current scenario of access to electronic legal doctrine, with the increased provision on the net of disparate legal web 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 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 NIR - Norme in Rete, the objective of which is the consistent retrieval of national legislation, jurisprudence and literature. At present this project is concentrating on legislative resources and as of now there are no plans to extend it to case-law texts and literature. In this context, a project focusing on Italian legal literature has been launched by the Institute of Legal Information Theory and Technologies. The project attempts to offer a unified point of access to multiple legal doctrine

https://doi.org/10.23106/dcmi.952107381

resources, by exploiting metadata and by providing tools for the discovery, selection and use of relevant legal materials.

2. A project for accessing Italian legal literature

The objective of the project is to implement a so-called vertical portal 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 their specific needs. The creation of a unified access system 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 sources in Italy. What the project wants to achieve is integrated access to a wide range of high quality resources and services. In our opinion, such an objective must be pursued by means of a thorough analysis of legal user needs, a careful selection of resources and reliance on rich and consistent metadata.

The retrieval of legal doctrine is a long, hard process encompassing multiple types of material. There is in fact no sole information provider that the user can point to during his search; different sources have to be identified and this requires services which will seek out and locate them.

In general, the main requirements for 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 its production and development;
c) the high quality of indexing and retrieval services.

Quality here means the richness of the semantics as well as the consistency of cataloguing. Relevance and precision are the main requirements in analysing legal doctrine in order to achieve consistency between the indexing language and the language used in the production of laws and case-law acts.

2.1. Problems and difficulties to overcome and possible solutions

Availability of documents

Users today demand access to information and direct, as well as immediate, availability of documents.

Different user interfaces

One crucial factor 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 search systems cause legal users serious problems of disorientation.

Identification of legal resources on the net

References to on-line legal doctrine are often intermixed in 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 on-line, 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 through different systems.

In order to overcome such difficulties, the careful selection and consistent organisation of the various information resources available were pinpointed as necessary measures. These resources include primary sources such as printed and electronic documents and web sites, as well as secondary reference sources such as OPACs, and indexing databases. Users need a unified point of access, a uniform interface for accessing heterogeneous resources, together with document delivery services. For this purpose a specialised portal is being developed, designed in such a way as to be open to the contributions of authors and publishers in delivering and structuring their output, adopting descriptive and communication standards that allow interoperability.

2.2. Survey of legal user needs

In implementing access services, the availability of standardised indexing and consistent metadata are of prime importance in order to allow for precision in searching and customisation according to user profiles. This is particularly crucial to users of legal information, who are far from homogeneous: they have quite different professional backgrounds, disparate needs and, consequently, extremely varied search approaches. They belong to various categories, such as lawyers and judges, students and scholars, employees of public administration offices, reference librarians and information professionals, politicians, managers, journalists, ordinary citizens, etc. For this reason it soon became clear that an accurate analysis of legal user needs was essential for the design of the portal. Each category of legal users has different requirements, knowledge and skills, and several user profiles have to be designed and implemented accordingly.

A survey of user needs was carried out using a

https://doi.org/10.23106/dcmi.952107381

questionnaire[1] to gain insight into their needs, requirements and habits in seeking for legal information.

The most important questions concerned their work, the type of material most used and desirable, the level of resources (bibliographic references, abstracts, full text, etc.), search parameters used (key words, author/editor, title, classification, temporal details, publisher, legislation and case law as treated in legal literature), desired services (document delivery, professional support of qualified personnel by e-mail, user-interactive services, personalised user interfaces, directories and lists of specialised Web sites, forums, newsgroups and newsletters), readiness to pay for services.

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 on-line 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. One specific requirement emerging from the questionnaire is the possibility to navigate through legal literature, legislation and case-law sources.

At this stage in the project no evaluation of the service had been carried out and efforts had been concentrated on the production of the prototype based on the above-mentioned requirements, as well as on the study analysis of comparisons between Dewey Decimal Classification (present in OPACs) and DoGi – Dottrina Giuridica database classification to allow cross-searching and semantically consistent retrieval of data.

2.3. Functionality and services

The portal will give access to different types of legal literature (bibliographic references, TOCs, abstracts, reviews, full text, web sites). For each of these sources the following functions are planned: a) resource discovery; b) identification of resources on the basis of their semantics (bibliographic and conceptual elements); c) localisation of documents through links to law libraries, serials' catalogues, publishers' lists, specialised web sites, etc; e) specific services like document delivery, downloading of digital texts, copy reproduction; f) control over the period of electronic resources' stability.

Based on a survey of European thematic vertical portals giving access to bibliographic references, the following additional services which could prove useful to legal users have been identified:

- lists of links to relevant material, organised on the basis of different branches of law;
- platforms for cooperation projects between users and authors for delivering contributions (such as articles, reports, reviews, opinions on sentences, theses, etc.);
- on-line legal advice and consultation;
- availability of an interactive legal dictionary offering legal notions and proposing free association of legal concepts;
- "Your Desk": an on-line personalised desk provided to users, where personal and current awareness files are stored.

On the basis of these functions the logic architecture of the portal has been designed, as summarised below:

- Interface
- User profiles
- Technologies for access
- Administrative and control services
- Publishing and distribution services for users
- Storage and data management services
- Production and harvesting of contents

3. Metadata approach

The essential conditions for setting up services capable of integrating a wide range of data and systems are:

a) the adoption of standards for the cataloguing (both descriptive and subject) and coding of resources;
b) the implementation of search tools and context-sensitive linking mechanisms in line with user requirements.

What is needed for legal users is integrated access to individual contributions hosted in different servers, to articles appearing both in printed and on-line journals (analysed using different subject systems), and to specialised web sites which need careful cataloguing in order to allow users to retrieve and select them. The need for a uniform metadata format has led to the choice of the Dublin Core metadata set, in its XML version, as the target bridging format.

As regards journal articles, we relied on the work done on DCMI Cite (http://www.dublincore.org/groups/citation/) as the target format from the native DoGi (database of analytics of Italian legal journals) records [1].

In our project we deal with a variety of data that can be divided into two main classes:

1. structured data coming from bibliographic repositories in libraries, which provide a specific metadata scheme of description;
2. web documents, namely HTML semi-structured documents, that, in most cases, do not contain 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 the portal, the structured data collected were selected from three data sources:

- DoGi: a metadata repository of articles related to

[1] The questionnaire (English text) can be found on url: http://www.ittig.cnr.it/QuestionnaireOnLegalLiterature.htm

https://doi.org/10.23106/dcmi.952107381

legal literature, maintained by our Institute;
- University library OPACs: UNIMARC-based library catalogues;
- A publisher catalogue of bibliographic records, using a proprietary metadata format[2].

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 to select and classify web documents (Sections 5.2 and 5.3);
- to perform a selective exploration of the web (Section 5.2), starting from such documents and following the hyperlinks with a high probability of pointing to other relevant documents.

4. Mapping of structured and semi-structured data

In order to provide integration with different data sources, while providing a uniform view for users in accord with the Dublin Core metadata set, two different approaches were adopted, relative to the different nature of the data sources.
- For the structured data, the metadata schemes supported by the selected repositories (so far based on UNIMARC format[3]) were mapped to the DC metadata set. In particular an accurate analysis of the DCMI Cite was conducted in order to have the DoGi records mapped against it (Tab. 1).
- For web documents, a specific module generating a meaningful subset of DC metadata was developed (Tab. 2).

As regards structured data a preliminary effort is required by data providers in order to expose metadata, making repositories compliant to the metadata scheme of our portal and ready to be harvested for the creation of a centralised index (Section 5.1).

In particular, as regards the mapping of DoGi records, the DCMI Cite is used for encoding the bibliographic citations of articles from journals. The application adopted the three distinct hierarchical levels: the journal level, the journal issue level 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 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.

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

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

Tab. 1 Mapping between DCMI Cite and DoGi.

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:type(scheme DCMIType)	Testo
dc:publisher	Editore
Dcterms:citation(scheme=DCMI Cite)	
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

Tab. 2 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 5.3)
dc:description	The content of the html metatag “description”, if any, othewise the content of the html tag <body>.
dc:type	“Web document”
dc:publisher	Web domain name

With reference to web documents, we decided to provide them with a uniform metadata scheme and chose a subset of DC metadata. However, unlike the structured repositories, at this stage of the project no specific effort is required from data providers to apply DC metadata to web documents, since documents are collected with no

https://doi.org/10.23106/dcmi.952107381

prior accord 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 summed up in Tab. 2 and described in Section 5.3.

5. The architecture of the federation system

The portal is the result of a federation system with service provider functions, the architecture of which consists of four main modules (Fig. 1):

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

2) a focused crawler, selecting and collecting semi-structured data, namely documents of interest from web sites;

3) an automatic metadata generator, supplying metadata to the selected web documents;

4) an indexer of the selected data.

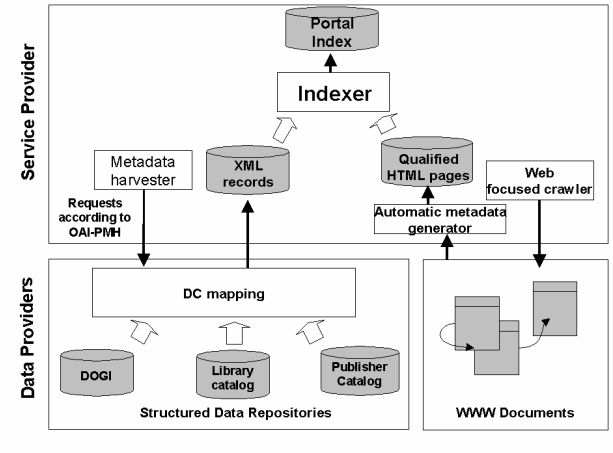

Fig. 1 The architecture of the federation system

5.1. Selection and harvesting of structured resources

In order to select resources of interest from structured repositories, we considered the specific classification metadata description within each data source.

The data from structured bibliographic databases are usually described by metadata providing a classification in accord with the UDC (Universal Decimal Classification) and DDC (Dewey Decimal Classification) systems and proprietary schemes (such as DoGi classification). Since the DoGi database contains only records related to legal literature, no particular selection was necessary, so all DoGi records were taken into consideration.

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. 3) are considered in particular.

Tab. 3 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 (civil, criminal, etc.), commentaries and collections of law cases are difficult to isolate from legal literature.

Multi-disciplinarity of legal literature:

Legal doctrine 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 the portal, a first effort was 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 (Open Archive Initiative) approach [2].

Most of the projects using OAI protocols aim at constructing portals as federative architectures of structured data repositories (mainly bibliographic records or e-print resources).

The TEL project [3], for instance, aims to construct a co-operative model of architecture able to access the foremost national bibliographic repositories in Europe. PHYSDOC [4] also offers a centralised access service of resources for the medical community, while CYCLADES [5] provides a collaborative, multidisciplinary, virtual archive service supporting scholarly communities in their work. The TORII portal [6], which also offers a cross-referencing service with different documents, is dedicated to those working on research into high-energy physics.

Our project also uses the OAI protocol of metadata harvesting (OAI-PMH) to collect data from structured repositories. These are made to comply with the adopted protocol requests and the portal metadata specifications as described in Section 4.

5.2. Selection and harvesting of web documents

The data from legal literature 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 portal. Usually general-purpose agents (often called spiders or crawlers) explore the hyperlinks of the web with the aim

https://doi.org/10.23106/dcmi.952107381

of finding as many different documents as possible. On the contrary, we are interested in selecting domain-specific documents and, therefore, only follow the hyperlinks which point to documents of interest for our portal.

In order to perform such a function we use the approach described in [8], 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.

Documents are 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 naive Bayes classifier [14] on a set of words in the vicinity of the hyperlink.

5.3. The automatic metadata generator

The application of metadata to web documents is an issue investigated in depth in literature within the field of semantic web construction. Some works aim in particular at evaluating the reliability of the authors' own metadata generation, or of similar collaborative activities between authors and metadata experts [9].

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 the web documents, combined with the collaboration of the authors, who are requested to complete the metadata.

Some other projects (www.klarity.com.au) have been carried out in order to provide web documents with metadata automatically, on the basis of keywords supplied by the authors, thus providing uniform and consistent metadata for such documents.

On the basis of this experience (most of which is summarized at www.lub.lu.se/tk/metadata/dctoollist) 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 to web documents.

Once the 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. Most of them are extracted 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") to the DC metadata.

Tab. 2 summarises the list of DC metadata applied to the selected web documents and the criteria used to generate them.

Particular attention has to be 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 [9]. This, however, is not our case, since web documents for our portal are selected with no prior accord between us and the authors or publishers. Therefore a special approach, based on automatic criteria of classification, has been used.

Firstly, being D the set of the selected documents, each document $d_j \in D$ has been described by a feature vector. Considering that the documents we deal with are usually rich in text, we have considered words as features.

Other possible choices, such as considering phrases, rather than individual words, as features to describe documents have been discarded *a priori*, in fact some experiments with this approach ([10], [11]) did not produce significantly greater effectiveness.

Moreover, according to [12], even if the choice of using phrases can be supported by the fact that they usually have superior semantic qualities characterizing the class of the documents in which they appear, their statistical qualities are usually inferior.

Each document d_j is therefore described by a vector of term weights $\vec{d}_j = [w_{1j},...,w_{|T|j}]$, where T is the set of words occurring at least once in at least one document; $0 \le w_{kj} \le 1$ is calculated according to the standard *tfidf* function [13], [14], which considers the weight w_{kj} as a function of the number of times the k^{th} word occurs in d_j.

The "dc:subject" metadata generator has been constructed in terms of an automatic document classifier.

According to [14], being D a set of documents and $C = \{c_0, c_1,..., c_{|C|}\}$ a set of categories, our "dc:subject" metadata generator consists in the construction of a ranking classifier that for a given document d_j returns the scores for the different categories. The score for the i^{th} class is defined in terms of the function $CSV_i : D \rightarrow [0,1]$ that, given a document $d_j \in D$, returns a *categorization status value* for document d_j in relation to class c_i. Such a score represents the evidence for a given document to belong to class c_i.

The automatic classifier was constructed using the naive Bayes method of automatic document classification [14], [8], where the $CSV_i(d_j)$ is obtained in terms of $P(c_i \mid \vec{d}_j)$, namely the probability that a document represented by a vector $\vec{d}_j$ belongs to class c_i. A naïve

https://doi.org/10.23106/dcmi.952107381

Bayes classifier computes this probability by the application of Bayes's theorem as follows:

$$P(c_i \mid \vec{d}_j) = \frac{P(c_i)P(\vec{d}_j \mid c_i)}{P(\vec{d}_j)}$$

where $P(\vec{d}_j)$ is the probability that a randomly picked document is represented by the vector $\vec{d}_j$, and $P(c_i)$ the probability that a randomly picked document belongs to c_i. $P(\vec{d}_j \mid c_i)$ represents the probability that a document, belonging to class c_i, is represented by the vector $\vec{d}_j$: the estimation of this term is a problematic point because the number of possible vectors $\vec{d}_j$ is too high (the same problem holds for $P(\vec{d}_j)$, but in the computation of the terms $P(c_i \mid \vec{d}_j), \forall i$, it can be omitted, since it is the same in each of them). Therefore in order to estimate the term $P(\vec{d}_j \mid c_i)$, the naïve Bayes assumption considers that the words in a document occur independently of each other given the class; such independence assumption simplifies the estimation of $P(\vec{d}_j \mid c_i)$ as follows:

$$P(\vec{d}_j \mid c_i) = \prod_{k=1}^{|T|} P(w_{kj} \mid c_i)$$

According to the previous remarks, the computation of $P(c_i \mid \vec{d}_j)$ to be used as $CSV_i(d_j)$ becomes:

$$P(c_i \mid \vec{d}_j) \propto P(c_i) \cdot \prod_{k=1}^{|T|} P(w_{kj} \mid c_i)$$

The naïve Bayes approach to document classification has been tested to provide the selected web documents with the "dc:subject" metadata. The experiment results are reported in Section 6.

As a result of the module of automatic metadata generation, the html meta-tags corresponding to the DC metadata reported in Tab. 2, and their contents, automatically generated by such a module according to the policies summed up in Tab. 2, have been applied to each selected document.

5.4. Index to selected resources

After having collected metadata from structured data repositories using OAI-PMH (Section 5.1) and having applied metadata to the web documents (Section 5.3), we obtain two archives containing data in different formats (XML records and HTML documents), sharing the same metadata description scheme.

At this stage an indexing procedure has been implemented, aimed at providing a uniform view and integrated access to the data of our portal.

In our experiments we have used an indexer [7] 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 the "dc: identifier"; 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 the record "dc:identifier".

The indexer works separately on the two archives 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 [7], representing the index of our portal.

In the search phase, data coming from structured repositories or 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.

6. The experiment

At this stage in the project we tested our approach on both data from structured repositories and web documents.

For the structured data we implemented an OAI data provider on the DoGi repository, using a package[4] available on the Open Archives Initiative web site. The data provider[5] we implemented can be tested using the related service[6] accessible from the OAI web site. We tested the OAI-PMH using a package[7] of metadata harvesting available on the OAI web site.

On the contrary, the experiments on web documents were carried out especially, at this stage of the project, on the module of the automatic metadata generator, particularly in testing the naive Bayes approach to document classification, aimed at applying the "dc:subject" meta-tag to such documents.

We collected 1220 documents from web sites of interest (see Section 3), belonging to 10 pre-established classes (Tab. 4); these documents represent the data set used for our experiments.

Tab. 4 The data set for our experiments grouped in classes

Class	Classes of the data set	Number of documents

[4] PHP-OAI Data Provider, University of Oldenburg, http://physnet.uni-oldenburg.de/oai/
[5] base url: http://xseries.ittig.cnr.it/portale/oai/dogi.php
[6] http://oai.dlib.vt.edu/cgi-bin/Explorer/oai2.0/testoai
[7] OAI/ODL Harvester, Digital Library Research Laboratory, Virginia Technology http://oai.dlib.vt.edu/odl/software/harvest/

https://doi.org/10.23106/dcmi.952107381

labels		
c_0	Environmental law	20
c_1	Administrative law	380
c_2	Civil law	177
c_3	International law	25
c_4	Constitutional law	28
c_5	European law	124
c_6	Computer Science law	134
c_7	Labour law	82
c_8	Criminal law	143
c_9	Taxation law	107

They were used both for training the naive Bayes classifier and for testing the reliability of the approach.

The data set was first used to train our naïve Bayes classifier. For these preliminary experiments, the training phase was carried out without considering any particular word stoplists or stemming.

Moreover, we described each document d_j by a feature vector $\vec{d}_j = [w_{1j},...,w_{|T|j}]$, where $0 \le w_{kj} \le 1$ is calculated according to a special case of the standard *tfidf* function, which considers w_{kj} as the number of times the k^{th} term occurs in the document d_j, normalized by cosine normalization [14]. This way of calculating w_{kj} gave the best results in our experiments.

Given a document, for each class i^{th} the CSV_i is calculated and the document is assigned to the class reporting the highest CSV.

The test of the classification capability of the naive Bayes module on the training data set itself produced an accuracy of classification of 87.2%.

The details of the classification results on the training set distributed among the classes, are reported in Tab. 5. The entry of the element (c_i,c_j) represents the number of documents of class c_i classified in class c_j.

Tab. 5 Test of the classifier on the training set.

	c_0	c_1	c_2	c_3	c_4	c_5	c_6	c_7	c_8	c_9
c_0	12	6	0	0	0	0	1	1	0	0
c_1	0	345	1	0	0	0	16	4	3	11
c_2	0	9	148	0	0	0	15	2	1	2
c_3	0	0	0	14	0	2	4	0	4	1
c_4	0	10	1	0	11	0	3	0	2	1
c_5	0	9	2	1	0	89	22	0	0	1
c_6	0	1	1	1	0	0	131	0	0	0
c_7	0	2	0	0	0	0	3	75	2	0
c_8	0	1	0	0	0	0	6	0	136	0
c_9	0	2	0	1	0	0	1	0	0	103

Then, a further experiment, aimed at evaluating the generalization capability of the classifier was carried out using the data set according to the "leave-one-out" testing strategy. In this testing strategy all the collected examples are used for training the classifier module, except one which is not included in the training set but is used to test the classification capability of the module. This is repeated, leaving one different example, at each step, out of the training set, till all the examples are used to test the classifier. The results of all the tests are combined, obtaining an evaluation of the reliability of the classifier on data from the training set.

This experiment produced an accuracy of correct classification of 75.4 %. The details of the classification results of the "leave-one-out" test, distributed among the classes, are reported in Tab. 6.

Tab. 6 "Leave-one-out" test on the data set.

	c_0	c_1	c_2	c_3	c_4	c_5	c_6	c_7	c_8	c_9
c_0	2	12	1	0	0	2	2	1	0	0
c_1	1	327	5	0	0	1	17	4	6	19
c_2	0	18	128	0	0	1	18	3	5	4
c_3	0	0	0	8	0	6	6	0	4	1
c_4	0	14	4	0	4	0	3	0	2	1
c_5	0	17	9	3	0	65	28	0	1	1
c_6	0	6	4	2	0	1	121	0	0	0
c_7	0	11	5	0	0	1	5	55	5	0
c_8	0	5	3	0	0	1	9	0	122	3
c_9	0	11	3	1	0	2	1	0	1	88

7. Conclusions

The use of Dublin Core for integrating structured resources and web documents is a possible solution for the creation of a portal onto Italian legal literature where a unified point of access and a uniform view of data are offered. The creation of a common form of data architecture is the most effective way to bridge the gap between different types of data. 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 were carried out. Web documents, on the other hand, are not usually supplied with metadata following particular schemes and, where present, they are generally not reliable. In order to supply web documents with metadata, an automatic metadata generator module based on a machine learning approach was developed. This module aims at supporting the intellectual activity of a service provider in its work of organizing web documents.

The project of a portal onto Italian legal literature is, therefore, the result of a federation system which combines the harvesting of structured data using OAI-PMH, whereas the gathering of web documents is selected and qualified automatically.

We are confident that the project will develop on the basis of co-operation between data providers and service providers in supplying common and reliable metadata.

Acknowledgements

Special thanks go to Andrea Passerini, Ph.D. student at the "Dipartimento Sistemi e Informatica" of the University of Florence, Italy, for the software development of the naive Bayes classifier.

References

https://doi.org/10.23106/dcmi.952107381

[1] A. Apps. *A Journal Article Bibliographic Citation Dublin Core Structured Value*, Retrieved on May 2, 2003 from http://epub.mimas.ac.uk/DC/citdcsv.html
[2] OAI – The Open Archives Initative Protocol for Metadata Harvesting, http://www.openarchives.org/OAI/openarchivesprotocol.htm
[3] TEL – The European Library, http://www.europeanlibrary.org
[4] PHYSDOC – Physics Documents Worldwide, http://physnet.uni-oldenburg.de/PhysNet/physdoc.html
[5] CYCLADES, An Open Collaborative Virtual Archive Environment, http://www.ercim.org/cyclades/
[6] TORII – The Digital Research Community, http://torii.sissa.it
[7] Swish-e, Simple Web Indexing System for Humans – Enhanced, http://swish-e.org
[8] A. McCallum, K. Nigam, J. Rennie, and K. Seymore. (2000). Automating the construction of internet portals with machine learning,. *Information Retrieval Journal*, *3*: 127-163.
[9] J. Greenberg, W. D. Robertson. (2002). Semantic web construction: An inquiry of authors' views on collaborative metadata generation. *Proceedings of the International Conference on Dublin Core and Metadata for e-Communities*, 45-52.
[10] H. Schütze, D.A. Hull, J.O. Pedersen. (1995). A comparison of classifiers and documents representation for the routing problem. *Proceedings of SIGIR-95, 18th ACM International Conference on Research and Development in Information Retrieval (Seattle, US, 1995)*, 229-237.
[11] K. Tzeras, S. Hartmann. (1993). Automatic indexing based on Bayesian inference networks. *Proceedings of SIGIR-93, 16th ACM International Conference on Research and Development in Information Retrieval (Pittsburgh, PA, 1993)*, 22–34.
[12] D. Lewis. (1992). Automating the construction of internet portals with machine learning. *Proceedings of ACM International Conference on Research and Development in Information Retrieval*, 37-50.
[13] C. Buckley, G. Salton. (1988). Term-weighting approaches in automatic text retrieval *Information Processing and Management*, *24*(5): 513-523.
[14] F. Sebastiani. (2002). Machine learning in automated text categorization. *ACM Computing Surveys*, *34*(1): 1-47.

https://doi.org/10.23106/dcmi.952107381