XML documents within a legal domain: standards and tools for the Italian legislative environment

C. Biagioli, E. Francesconi, P. Spinosa, M. Taddei

ITTIG - Institute of Legal Information Theory and Technologies
Via Panciatichi 56/16- 50127 Florence - Italy
{biagioli, francesconi, spinosa, m.taddei}@ittig.cnr.it
WWW home page: http://www.ittig.cnr.it

Abstract. The *Norme in rete* (NIR) [Legislation on the Net] national project aims at making easier the retrieval and the navigation between legal documents in a distributed environment and to encourage the development of systems with characteristics of interoperability and effective of use. In order to obtain this, two standards have been defined: a URN standard, to identify these materials through uniform names, and XML-DTDs to describe legislative documents within the NIR domain. In this paper the definition of such standards and the developments of tools aimed at making easier their adoption are illustrated. Particularly this paper presents a specific law drafting environment, *NIREditor*, able to produce legal documents and to handle legacy legislative documents according to the NIR standards.

1 Introduction

Access to legal information for citizens is one of the main democracy objectives. Users and legal experts increasingly feel the need to retrieve legal documents from the Web and the links between them in order to learn about the law and fully understand legal texts. In order implement these services and to eliminate information historical fragmentation in legislative environment, in Italy the "Norme in Rete" (NIR) project ("Legislation on the Net") has been proposed by the CNIPA [Italian National Center for Information Technology in the Public Administration] in conjunction with the Italian Ministry of Justice. The project aims at creating a unique access point on the Web with search and retrieval services of legal documents, as well as a mechanism of stable cross-references able to guide users towards relevant sites of public authorities participating in the project. To achieve these purposes, the NIR project proposed the adoption of XML as a standard for representing law documents. Particularly the project proposed a description of law texts by three DTDs with increasing degree of depth: they aim at representing a legal text with respect to its structural or formal profile, and using particular meta-information to its semantic or functional profile.

Moreover a uniform cross-referencing system, based on URN standards [1], able to provide a stable system of cross-referencing has been established. In order to make easier the adoption of such standards, some tools have been developed within the NIR project. In particular, in this paper, the *NIREditor* authoring tool is presented, which includes facilities and modules aiming at managing new or legacy law documents according to the established standards.

In Section 2 the standards established within the project are illustrated. In Section 3 the main features of *NIREditor* are presented: particularly in Sections 3.1, 3.2, 3.3 different working situations are described. Finally, in Section 4, some conclusions are discussed.

2 The NIR standards

The feasibility study of the NIR project proposed the adoption of XML as a standard for representing legal documents. This study aimed at representing a legal text with respect to its formal structure, using also additional meta-information and a uniform cross-referencing system providing documents with characteristics of interoperability and effective of use. This preliminary study, carried on by two specific national work groups produced two main official standards:

1. a standard for cross-referencing legal documents has been defined in accordance with the uniform name (URN) technique: an unambiguous identifier, that allows the references to be expressed in a stable way, independently of their physical location;
2. a standard for legal document description has been formulated by defining XML-DTDs (NIR-DTDs) of increasing degree of depth in text hierarchy description for different kind of legal documents (similar initiative is the MetaLex project [2]). As well as including the NIR-URN standard for cross-references, the NIR-DTDs provides:
   - a structural description of text, establishing constraints in the hierarchy of the formal elements of a legislative text (collections of articles);
   - a specification of the metadata which can be applied to a legislative document or to parts of it.

2.1 The URN standard

Within the NIR project, documents are identified through a uniform name. Uniform Resource Names (URNs) were conceived by the Internet community for providing unambiguous and lasting identifiers, independent of physical location, of network resources. In legal documents, references to other legislative measures are very frequent and extremely important. The hypertext links of the Web meet this need, but do not appear to be suitable for wide-scale use in the law: reference to the resource referred to is, in fact, based on its physical location expressed in a uniform mode through its Uniform Resource Location (URL), which presents the following well-known problems:

- difficulty in knowing the location of the cited resource;
- the loss of validity over time of the locations (URL) in the references;
- the impossibility of referring to resources that have not been published yet;

which, therefore, make the network of links between documents extremely limited with respect to their potential and to their increasing unreliability over time.

In order to avoid these problems, a system of references based on assigning a uniform name to each legal resource and on resolution methods (RDS: Resolver Discovery Service) able to retrieve the corresponding object has been chosen. These tools are in conformity with those defined within IETF (Internet Engineering Task Force) by the special working group (URN Working Group) and described in various documents - from the official standards (RFC: Request For Comments) to the drafts - to which alignment is guaranteed even in the future.

Assigning a uniform name to every legal document has the scope of associating every legal document with an unambiguous identifier, in standardized format, that only depends on the characteristics of the document itself and is, therefore, independent of on-line availability, of physical location and of access mode. This identifier is used as a tool for representing the references - and more generally every type of relation - between the legal acts. In an on-line environment with distributed resources between different Web publishers, its use facilitates the construction of a global hypertext between legal documents and a knowledge base storing the relations interconnecting them. The association of the uniform name to the document occurs through meta-information, that may be:

– inserted in the document itself: it is the solution that can be adopted in HTML files (through the META tag) and also in XML files (through a suitable tag);
– external but strictly related to the document: by traditional techniques as a specific attribute in a database, or using growing methods as adopting RDF technology.

In any case, the software tools used must be able to implement and update the (distributed or centralized) catalogues which are functional for resolution and, therefore, to give access to the document through the uniform name. Other meta-information (for example, details, title, subject-matter, relations, whether in force, etc.) which enrich the system response, can be present in these catalogues that store the uniform name and location for each document. The uniform names system of the domain of interest must include:

- a schema for assigning names capable of representing unambiguously any legal measure, issued by any authority at any time (past, present and future);
- a resolution mechanism - in a distributed way - from uniform name to on-line location of the corresponding resources.

Uniform names in the law, as proposed by a special NIR working group has been adopted as a technical regulation by Italian legislative system. In conformity with RFC 2141 *URN Syntax* [3], which defines the general syntax of a uniform name, for legal documents a name-space identified by "nir" (this space identifies the context in which the names are valid and significant) has been defined and, therefore, the relative URN have the following format:

```
<URN> ::= "urn:nir:" <NSS-nir>
```

The specific name ¡NSS-nir¿ must contain information appropriate for unambiguously identifying the document. In the legal domain they are essentially four data: the enacting authority (or the authority referred to), the type of measure, the details and any annex. For legislation, it is also necessary to distinguish between any later versions of the document, following amendments that have been made over a period of time. In this case, the identifiers of the legislative act remain the same, but information is added regarding the version under consideration. Therefore, the more general structure of the specific name appears as follows:

```
<NSS-nir> ::= <document> ["@" <version>]
```

A structure for identifying the document is defined, composed of the four fundamental elements mentioned above, clearly distinguished one from another in accordance with an order identifying increasingly narrow domains and competence:

```
<document> ::= <authority> ":" <measure> ":" <details> [":"
                                <annex>]
```

The main elements of the uniform name are generally divided into several elementary components, each having established rules of representation (criteria, modes, syntax and order). Such a syntax allows the automatic construction of the URN, starting from the text of the citation. The complete syntax specification of the uniform names belonging to the "nir" name-space can be seen in [1], whilst some important examples of uniform names of legal documents are:

Act 24 November 1999, No. 468

`urn:nir:stato:legge:1999-11-24;468`

Decree of Ministry of Finance of 20.12.99

`urn:nir:ministero.finanze:decreto:1999-12-20;nir-3`

AIPA circular of 21 June 2001, No. 31

`urn:nir:autorita.informatica.pubblica.amministrazione:circolare:2001-06-21;31`

Decision of the Italian Constitutional Court No.7 of 23 January 1995

`urn:nir:corte.costituzionale:sentenza:1995-01-23;7`

To each uniform name, the system of resolution has the task of associating the respective network locations. It is based, within a distributed architecture, on two basic components: a chain of information in DNS (Domain Name System)

and a series of resolution services from URNs to URLs, each competent within a specific domain of the name space. Particular attention has been paid to the resolution system in order to provide an answer to the user, even in case of uncompleted or uncorrected uniform names, derived from uncorrected citations (for example the resolution service gives back the list of the documents whose URNs partially match the provided URN, or it attempts to correct automatically the URN itself).

2.2 The NIR-DTDs standard

As well as the NIR-URN standard, the NIR project has defined a standard based on XML, aimed at describing the content of legislative documents. For this purpose three DTDs with increasing degree of depth have been established:

- the "DTD flessibile" (niloose.dtd) contains about 180 elements: it does not establish any mandatory rules (unless in a very small quantity) and it is used for legacy legislative documents not following drafting rules;
- the "DTD base" (nirlight.dtd) contains about 100 elements: it represents a subset of the "DTD completo": it is useful to train users in adopting the DTD standards;
- the "DTD completo" (nirstrict.dtd) contains about 180 elements: it follows legislative drafting rules and it is used to write new legal documents.

The "DTD flessibile" and "DTD completo" are composed by four common files:

1. global.dtd: containing general definitions;
2. norme.dtd: containing definitions of the division structures;
3. text.dtd: for text, table and form structure definitions;
4. meta.dtd: containing metadata schemes definitions.

Differences are present in the main files nirstrict.dtd and nirloose.dtd. The nirstrict.dtd establishes an order to the partitions of a law text. Collections of articles are still considered the basic elements of the norm (their numbering is independent from the hierarchical organization of the other elements). Numbering of the divisions is mandatory. Titles of the divisions are not provided, while they are optional for the other elements. The nirloose.dtd establishes only few constraints and it is used for legacy legislative documents which usually do not follow particular legislative drafting rules. The NIR-DTDs basically describe a legislative text under two profiles:

– the *formal profile* which considers a legislative text as made up of divisions;
– the *functional profile* which considers a legislative text as composed by elementary components called provisions (fragment of a regulation) [6].

In other words, the fragments of text inserted have a formal and a functional appearance. They are, at the same time, partitions and provisions, according to

whether they are seen from a formal or functional view-point. The two points of view can be alternated as required during the definition of the text.

In particular the functional profile can also be considered as composed by two sub-profiles: the *regulative profile* and the *thematic profile*. The first one reflects the lawmaker directions, the second one the peculiarities of the regulated field. On the NIR-DTDs point of view, the regulative profile is identified by particular metadata called analytical provisions, the thematic profile are partly illustrated in the so-called subjects of the provisions.

3 The NIREditor

The NIR-DTDs identify a wide and complex subset of documents: basically law texts and regulative acts. The production of new documents, as well as the transformation of legacy contents according to the NIR standards, can be a hard problem to face without an editing system guiding and supporting the user.

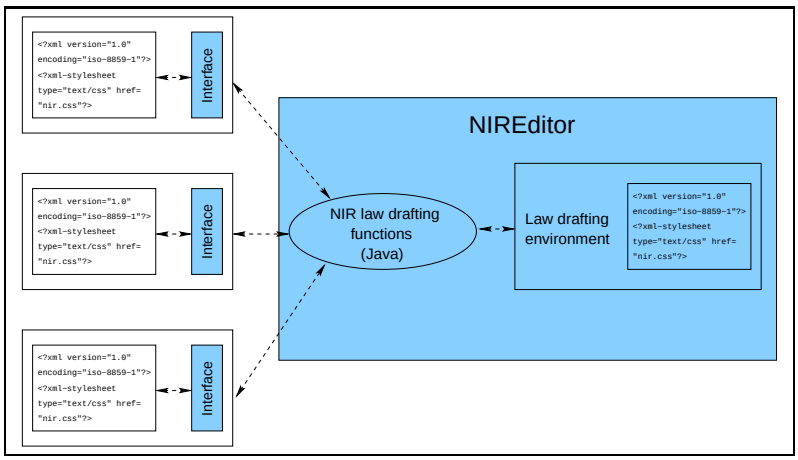

Fig. 1. The NIREditor and its connections to general-purpose XML editors

Even though programs for drafting texts in XML already exist, we have decided to develop a specific environment to handle NIR-XML documents. The limits of present XML editors in fact, whether used for a specific class of documents, concern the generality and inadequacy of their editing functions, in particular as regards functions implementing the NIR-DTDs constraints.

Therefore, as well as for producing HTML documents according to the HTML-DTD, specialized editors exist, similarly to help law texts drafting according to NIR-DTDs standard, a specialized visual editor (*NIREditor*) has been developed [7] [2]: it consists of a law drafting environment supporting specific Italian

legislative technique functions. The software architecture of *NIREditor* is represented by a kernel of Java specific functions library, fully integrated within the law drafting environment; they can also be integrated to the main XML general purpose editors supporting a Java API (Fig. 1).

The *NIREditor* operates within the URN and DTD NIR framework and it is designed to assist the drafting of new texts, as well as to process legacy law texts. Two working situations are thus catered for: the processing of an existing text or the processing of new texts, with its different situations: composition and organization of new texts. In Figure 2 the *NIREditor* drafting environment is shown.

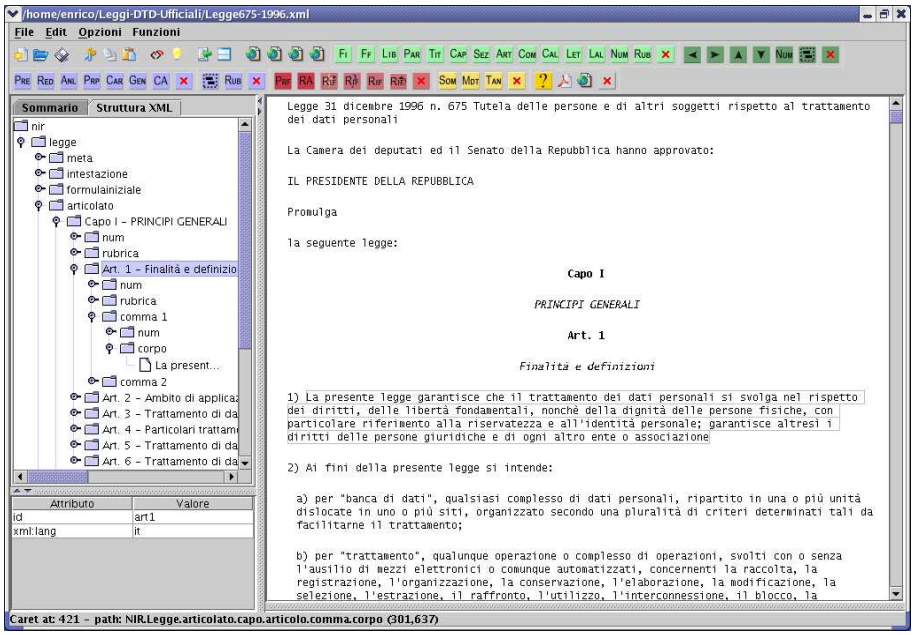

Fig. 2. The NIREditor environment

3.1 Importing texts

In this case, instruments for recognizing the basic aspects of the texts are available, which allow automatic pre-marking of all the parts of the structure recognized in the text analyzed, in accordance with the NIR-DTD, thus recognizing the formal profile of the legislative text. This structure parser is designed to help the XML conversion of documents which otherwise would have to be carried out completely by hand; it includes also a cross-reference parser able to locate cross-references and to assign the related URNs; it is based on a grammar implementing a bottom-up parsing strategy.

Currently the structure parser implements a non-deterministic finite-state automata (NFA), where the states are represented by the elements of the NIR-DTD, and the transitions among the states are associated to formal rules of document parts division. As well the cross-reference parser is constructed as a syntactical parser, on the ground of a cross-reference grammar. In case of parsing errors, the completion, correction and validation of the pre-marking is possible using formal structure management functions, and a text panel where plain text can be handled.

The result of the structure parsing function is the formal profile of the text which is established by the structural elements of the NIR-DTDs.

A further way of marking a pre-existing text is represented by the application of the analytical metadata to a law text, therefore the recognition of the functional profile of a legislative text, whose schema is established by the NIR-DTDs. Such metadata are intended to qualify the provisions of a text law. Examples of provisons are *duty*, *right*, *delegation*, *competence*, *power*.

As the marking of the formal structure, the insertion of analytical metadata for provision classification can be manually carried out, however this function can be particularly time consuming. Therefore, within *NIREditor* a module supporting the user in provision classification, based on machine learning techniques for text classification has been developed: it extracts automatically from the text of the provisions their relevant meanings according to the NIR analytical metadata schemes. Particularly a naïve Bayes approach of text classification [14] has been used.

Class labels	Classes of the data set	Number of documents
c_0	Repeal	70
c_1	Definition	10
c_2	Delegation	39
c_3	Delegification	4
c_4	Duty	13
c_5	Reservation	18
c_6	Inserting	121
c_7	Prohibition	59
c_8	Permission	15
c_9	Penalty	122
c_{10}	Substitution	111

Table 1. Classes (provisions) and number of documents for each class in the experiment

The reliability of the classifier has been tested considering a data set of 582 provisions distributed among 11 classes (Tab. 1) representing as many types of provisions. The collected data set has been used both to train the naïve Bayes classifier and to test the reliability of the approach. In order to reduce the com-

plexity of the problem, a phase of feature selection, in our case words, has been performed. From the vocabulary related to the data set, we selected a number of words with the highest *information gain*, as defined in [15], representing the discriminative power of a word with respect to the classes.

Classes	c_0	c_1	c_2	c_3	c_4	c_5	c_6	c_7	c_8	c_9	c_{10}
c_0	70	0	0	0	0	0	0	0	0	0	0
c_1	0	9	0	0	0	0	0	0	0	1	0
c_2	0	0	39	0	0	0	0	0	0	0	0
c_3	4	0	0	0	0	0	0	0	0	0	0
c_4	0	0	0	0	10	1	0	1	0	1	0
c_5	0	0	0	0	0	17	0	0	0	1	0
c_6	1	0	1	0	1	0	118	0	0	0	0
c_7	0	0	0	0	0	2	0	55	0	1	1
c_8	0	0	0	0	0	1	0	1	12	0	1
c_9	0	0	0	0	0	0	0	0	0	120	2
c_{10}	3	0	0	0	0	0	0	0	0	2	106

Table 2. Test of the classifier on the training set.

To train the classifier we have performed a stemming procedure on words to obtain a normalized vocabulary, so that different variants of the same word are considered as occurences of the same normalized form, since they contribute in the same way to the semantic of a text. Moreover we have considered only the n words of the vocabulary with the best *information gain*. The best results of the classifier have been obtained considering $n = 500$. In this configuration, the classification results on the training set obtained an accuracy of 95.5% Being c_i the i^{th} class of the provision, the details of the classification results on the training set are reported in Tab. 2. The entry of the element (c_i, c_j) represents the number of documents of class c_i classified in class c_j.

The generalization capability of the classifier has been tested using the "leave-one-out" strategy: all the collected examples are used to train 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.

The results of the classification capability using the "leave-one-out" strategy obtained an accuracy of 88.6%. The details of the classification results on the training set are reported in Tab. 3.

3.2 The composition of new texts

For the composition of new texts, *NIREditor* is conceived as a visual editor, supporting the user in producing valid documents according to the chosen DTD.

Classes	c_0	c_1	c_2	c_3	c_4	c_5	c_6	c_7	c_8	c_9	c_{10}
c_0	67	0	0	0	0	0	1	0	0	0	2
c_1	0	4	0	0	1	1	0	2	0	1	1
c_2	0	0	39	0	0	0	0	0	0	0	0
c_3	4	0	0	0	0	0	0	0	0	0	0
c_4	0	0	1	0	3	1	0	4	0	2	2
c_5	0	1	0	0	0	11	0	1	2	1	2
c_6	1	0	2	0	1	0	114	1	0	0	2
c_7	0	0	0	0	0	3	1	53	0	1	1
c_8	0	0	1	0	0	7	0	2	3	0	2
c_9	0	0	0	0	0	0	0	0	0	120	2
c_{10}	3	0	2	0	0	1	1	0	0	2	102

Table 3. Test of the classifier according to the "leave-one-out" strategy.

No XML validation function is contained within the editing environment, since the editor allows the user to perform only valid operations. Moreover, it helps the user in composing particular section of a new law document using dialogue windows, and permit the introduction of the metadata provided by the NIR-DTDs in the correspondent part of the document.

The insertion of the XML formal partitions provided by the NIR-DTDs can be obtained by the editor guide which suggests the user the XML elements that can be introduced according to the context of the insertion point.

Particular facilities available within the drafting environment are the automatic numbering of the divisions and the update of internal references in the event of text movements or variations. Automatisms are present as far as the construction of external and internal cross-references are concerned: there is a guided composition window for the formulation of cross-references as well as instruments of automatic recognition and construction of references and the related URNs.

It is possible to construct a new text by determining *a priori* the structure and insert the content of the various parts afterwards, or else passages can be inserted in no particular order, then organized and inserted into a suitable structure at a later time. During the composition, a further valorization of a legislative text is represented by the application of the analytical metadata and their subjects to the divisions. This can be done by hand or using the provision classifier as a support. In the event that metadata have been inserted, which are the result of documentary requirements, it is possible to make use of these notes to help in determining a fine logical structure of the text being processed, as well as for subsequent network information searches.

3.3 The organization of new texts

For the organization *a posteriori* of new texts, two alternative strategies can be followed: the *formal strategy* and *the functional strategy* [6], [16], [**?**].

The formal strategy considers the text according to the formal profile: the text is made up of divisions (collection of articles). Using the formal strategy the partitions of similar rank to be organized are chosen by the draftsman. The editor will create a new part of an immediately higher rank, applying the rules of formal text structuring to the same.

The functional strategy considers the text according to the functional profile: the elementary component of a text is a provision (fragment of a regulation). The draftsman carries out the same operations in an indirect way: the partitions to be organized are chosen according to their content, affinities etc. as well as it is decided where they should be placed in the text, according to the preferences of the drafter and the customary procedure of presentation used in some rules of legislative technique. The attention to the functional profile of a legislative text based on analytical metadata is one of the key points of *NIREditor*; this is the precondition of creating at least a domain-specific semantic portion of the Web.

4 Conclusion

In this paper the standards established for publishing law documents within a distributed architecture, based on DTD-XML and URN for cross-references, have been presented. Such standards has been established within the NIR project promoted by the Italian Ministry of Justice and the Italian National Center for Information Technology in the Public Administration. In order to make easier the adoption of such standards, some tools have been developed. In this paper, in particular, we have presented a visual editing system, *NIREditor*, able to produce new law documents, as well as the transformation of legacy contents according to the NIR standards.

Acknowledgments: Special thanks for the development of *NIREditor* go to Alessio Ceroni, Andrea Passerini and Tommaso Agnoloni, Ph.D. students at DSI - University of Florence, to Gianni Giorgetti, Ph.D. student at the University of Florence, to Lorenzo Sarti, Ph.D. student at DII - University of Siena, to Stefano Pietropaoli, Ph.D. student at the University of Pisa. Thanks also to Andrea Marchetti, researcher at IIT - Italian National Research Council, for many fruitful discussions and suggestions.

References

1. Spinosa, P.: Identification of legal documents through urns (uniform resource names). In: Proceedings of the EuroWeb 2001, The Web in Public Administration. (1997)
2. Boer, A., Hoekstra, R., Winkels, R.: Metalex: Legislation in xml. In: Proceedings of JURIX 2002: Legal Knowledge and Information System. (2000) 127, 163
3. R. Moats, K.R.S.: Urn syntax. Technical Report RFC 2141, Internet Engineering Task Force (IETF) (1997)
4. Daniel, R.: A trivial convention for using http in urn. Technical Report RFC 2169, Internet Engineering Task Force (IETF) (1997)

5. T. Berners Lee, R. Fielding, L.M.: Uniform resource identifiers (uri): Generic syntax. Technical Report RFC 2396, Internet Engineering Task Force (IETF) (1998)
6. Biagioli, C.: Towards a legal rules functional micro-ontology. In: Proceedings of workshop LEGONT '97. (1997)
7. Biagioli, C., Francesconi, E., Spinosa, P., Taddei, M.: The nir project: Standards and tools for legislative drafting and legal document web publication. In: Proceedings of ICAIL Workshop on e-Government: Modelling Norms and Concepts as Key Issues. (2003) 69–78
8. Lesk, M.: Lex - a lexical analyzer generator. Technical Report CSTR 39, Bell Laboritories, Murray Hill, N.J. (1975)
9. Johnson, S.: Yacc - yet another compiler compiler. Technical Report CSTR 32, Bell Laboritories, Murray Hill, N.J. (1975)
10. C. Apté, F.J. Damerau, S.W.: Automated learning of decision rules for text categorization. ACM Transactions on Information Systems **12** (1994) 233–251
11. S.T. Dumais, H.C.: Hierarchical classification of web content. In: Proceedings of ACM International Conference on Research and Development in Information Retrieval. (2000) 256–263
12. Lewis, D.: Automating the construction of internet portals with machine learning. In: Proceedings of ACM International Conference on Research and Development in Information Retrieval. (1992) 37–50
13. G. Salton, C.B.: Term-weighting approaches in automatic text retrieval. Information Processing and Management **24** (1988) 513–523
14. Sebastiani, F.: Machine learning in automated text categorization. ACM Computing Surveys **34** (2002) 1–47
15. Mitchell, T.: Machine Learning. Mc Graw hill (1997)
16. A. Valente, J.B.: A Functional Ontology of Law. C. Ciampi, F. Socci Natali, G. Taddei Elmi (eds), Verso un sistema esperto giuridico integrale, CEDAM (1997)