Using XML as a means to access legislative documents: Italian and foreign experiences

Andrea Marchetti **IAT-CNR**

Fabrizio Megale **Camera dei Deputati**

Enrico Seta **Camera dei Deputati**

Fabio Vitali **Università di Bologna e IAT-CNR**[1]

Abstract

In this paper we describe the goals and the organisation of the ongoing project "Norme in Rete" (NIR – http://www.normeinrete.it), which involves several important Italian institutions and organisations. This project aims at the production of tools for the access to Italian normative documents, and data formats for the standardisation of the text of laws and rules both national and local. One of its many goals is the conversion of the national law corpus into XML.

Within the context of this project, our effort has concentrated on the development of an XML DTD already, and of an XML Schema very soon, to describe Italian national and local laws. We illustrate in this paper the overall structure of the DTDs. They are organised in a stricter, normative set of rules, with normative power, for new law drafts, and of a looser, descriptive set of rules for existing documents over which no rules can be imposed. In this paper we examine both types of DTD (strict and loose), their global organisation, the modules for legal elements, for textual and tabular tags (resembling HTML), and for modular, generic elements, that allow easy extendibility to the DTD. Also the treatment of meta-information is examined in this paper.

We produce a short account of several analogous experiences in Northern Europe, carried out by both public institutions and private legal publishers. Mention is also made of the European Union's similar projects.

The project "NormeInRete"

An important and widespread debate has started in recent years (in Italy as well as in most countries) about the chances for the new Information Technologies to simplify the relationship of the citizens with the norms and laws that rule their lives. Huge document bases accessible through Internet, as well as clear and usable drafting rules to simplify the technical jargon and the frequent structural flaws, have been created for the simplification of the access and understanding of norms.

The project "Norme in Rete" (translation: *Norms on the Net*), or NIR, started in 1999 with the leadership of the Italian Ministry of Justice, and it gathers several Italian public institutions and research organisations. It is financed and managed by AIPA, the Italian Authority for the Information Technology in the Public Administration.

The main aims of the NIR project are twofold:

- on the one hand the creation of a single Internet portal that provides free access to all the documents with legal relevance produced in our country and available on various institutional sites: overcoming the centralised software architectures so far employed by the legal databases, the NIR project aims at providing a central access point to a plurality of institutions producing normative documents having validity in Italy;
- on the other hand, the development of data formats and markup vocabularies that those institutions should formally adopt for the generation of their normative documents, in order to

[1] Gli autori sono contattabili all'indirizzo e-mail fabio@cs.unibo.it

facilitate the development of document collections, search tools, annotation tools and any other kind of application that may increase the ease of use of our massive system of norms.

Information about the NIR project can be found on the site "Norme in Rete" itself (http://www.normeinrete.it) and on the main AIPA site (http://www.aipa.it/attivita[2/progettiintersettoriali[10/norms[8/index.asp). Furthermore the Government has placed it among the main activities of the Italian Plan for the E-government (http://www.palazzochigi.it/fsi/).

The problems and difficulties connected to legal drafting (the context in which the NIR project is positioned) are largely known by both professionals of the field and the general public. Large debates have started among the experts, within the general framework of the relationships between the citizens and their laws. It should be noted that these problems and difficulties concern not only Italy, but most developed countries, as we will briefly explain at the end of this paper.

The production of norms in Italy presents quantitative problems, due to the proliferation of laws and other acts, and qualitative ones, due to intrinsic faults in the produced documents. In addition, several other problems arise. On the one hand, there is an increasing number of institutions that can create norms as well as the multiplication of the types of document produced. An example is the proliferation of Authorities (such as the Telecommunication Authority) and the ambiguous classification of their documents. On the other hand, external legislation bodies (such as the European Union or other international bodies) have an increasing importance in our legal system.

Among the solutions being undertaken, two are relevant to our discussion.

- At the beginning of the legislative process, it is worth mentioning the reform of the so-called drafting rules, i.e., the technical formulation of the legal documents. With the *circolare* (circular letter) of 20th April 2001, the Prime Minister Cabinet, the Chamber of Deputies and the Senate have updated and standardised the formal rules for the drafting of norms issued by our central State, regardless of their denomination. On the other hand, in former times regions and other local institutions have established independent sets of drafting rules, among which the most important are those elaborated in 1991 by the "Osservatorio legislativo interregionale".
- At the end of the legislative process, account should be taken of the recent transformations in the society: the shift towards a decentralised, bottom-up, internationalised approach to the political life, fostered by the ongoing diffusion of net-related technologies, gives a new importance to the access rights of all citizens to the text of the laws. And the European Community and other supra-national bodies make this issue felt not only for the national laws, but for all the norms that affect us.

We dare say that markup languages, and in particular XML, can provide interesting results at both ends of the legislative process: at the drafting stage, enforcing some or all the drafting rules defined for our norms; at the accessibility stage, fostering easy and sophisticated searching and rendering tools for the public at large. Furthermore, XML may constitute a great influence on several other aspects of the legislative process, providing support for the consolidation of laws, rationalising the legislative process, improving the referencing and connections among the norms, etc.

To the aforementioned aims, besides the issues connected to the portal itself, three working groups have been formed within the NIR project for the development of standards for the representation of norms and legal documents. The three working groups, composed of government officials, computer scientists, lawyers, researchers and documentalists, are responsible for building a set of DTDs for law documents, determining the relevant meta-data to accompany the documents, and of devising a global addressing mechanism based on URNs (Uniform Resource Name) for the easy access to documents.

In this paper we will present the results of the first and second working group and in particular we will illustrate the characteristics of the DTDs of the legal documents that have been elaborated. The output of the URN working group has been made available as a draft on the NIR web site. At this time, the recommended set of XML tools that will be used by the organisations involved in the normative process is not defined yet. Of course, the aim of the three working groups is to define documents and formalisms that can be used freely with the widest range of commercial and open-source tools.

XML and normative documents

XML (Extensible Markup Language) is the most recent and promising among the markup languages. This syntax for structuring text documents derives from SGML, Standard Generalized Markup Language (standard ISO 8879 since 1988), and from HTML. Although it is a not proprietary standard (it was proposed by W3C, the international committee developing the languages and the protocols for the World Wide Web), it is promoted and supported by the widest collection of commercial and research software producers. The main aspect of XML relevant to our aims is the possibility, like SGML, of specifying not quite the rendering characteristics of a document (such as its font name, size, and style, or its margins and alignments), but rather the structures typical of the class to which the document pertains (such as titles, sections, subsections, paragraphs, and so on). In other words, with XML it is possible to provide descriptions and enforce constraints relevant to any arbitrary class of documents, by listing the constituent elements and the structural rules.

The formally correct drafting of a normative document contributes significantly to the comprehensibility, usability, effectiveness, and economy of the norms it contains. The rules for a formally correct legal drafting are collected and explained in rules manuals, and they provide rules regarding the spelling, the lexicon, the syntax, the writing style and the structure to be employed for the legislative documents. Of course, even before these rules were explicitly written down, there already existed a drafting praxis, more or less detailed depending on the normative institution, that has manage to produce in the past norms showing a remarkable homogeneity in style and structure. It is worth mentioning that in most cases these norms are still in force, and are part of the Italian norms system.

Normative documents (such as laws, decrees, regulations, and so on) are ideal candidates for an XML representation;

1. they have a clear, systematic and predefined structure: the normative content is usually contained in clauses organised in a hierarchical structure of several levels ("libri" being divided into "parti", divided into "titoli", then "capi", "sezioni", "articoli" and finally "commi", that contain the actual text of the norm. The hierarchical structure (called "articolato") is usually preceded and followed by formulaic texts, and may also present preambles and annexes.

2. they contain required and optional elements (e.g.: the upper levels of the hierarchy are always numbered, while the lower level, "commi", are numbered only occasionally - actually, only in recent texts).

3. there exist containment constraints: i.e. it would be incorrect (in XML terms, it would be invalid) to create a legal document where the "commi" contain "articoli", or where the title of the "articolo" (the "rubrica") is divided into "commi" itself. A document with such a structure must undoubtedly be rejected because it does not follow the fundamental structural rules of normative documents.

The DTD (Document Type Definition) is the optional part of an XML document where the correctness rules for a class of document are detailed. The DTD contains the list of the elements allowed within the class of documents and some rules about the composition of these elements. A special class of XML engines use the DTDs to check the correctness of a document against the rules expressed there, helping in discovering and correcting structural errors.

Thanks to the particular adaptability of normative documents to XML, the NIR Working Group built several DTDs to express the structural rules of the Italian normative documents. Two main classes of documents have been taken into consideration: those normative documents that are to be created according to the recent set of drafting rules, and those that have already been created in the past, possibly following the drafting rules and possibly making exceptions.

The DTDs are useful to enforce structural homogeneity among the documents of one kind: True, DTDs allow rules only for elements structures, and not content structure. For instance, we can specify that a "message" document necessarily contains one "sender", one "receiver", and one "date" element, but we cannot require that the "date" element actually contains a well-formed and existing date. To provide this kind of rules, a new standard has been created, called XML-Schema, that permits the specification of both structural and content rules.

The DTD working group is working on a Schema version of the drafting rules, but at the moment it seems that, syntax aside, the new features of Schema will affect these rules in only a very limited way. Thus we believe that presenting the DTD in this paper is still rather representative of the ongoing work.

The NormeInRete DTD

Strict DTD and Loose DTD

The most evident aspect of the NormeInRete DTD is the parallel support of the documents that follow the drafting rules expressed in the "circolare" of 20th April 2001, and of the documents that, having been written earlier or by institutions that are not bound to it, may present some differences.

The drafting rules contained in the above-mentioned "circolare" express constraints about the order and names of the hierarchical parts, as well as regarding their titling and numbering.

Norms have always followed most of these constraints, but some of them have been formalised recently, and not all documents are compliant to them. In order to describe both kinds of documents, the working group produced two different DTDs, called "Strict" and "Loose". The main characteristic of these DTDs is their reciprocal compatibility: the Strict DTD does not describe structures different from the Loose DTD, but only adds more constraints to the acceptable structures. Consequently, all the documents that are valid according to the Strict DTD are also valid according to the Loose DTD; besides, all the documents that were drafted according to the rules of the "circolare" but were created earlier can still use the Strict DTD.

The Three Classes of Documents

Apart from a few specific differences, the Strict and Loose DTD identically describe all normative documents, and divide them into three main categories: those employing a formalised hierarchical structure preceded by a preamble ("articolato con preambolo"), those employing a formalised hierarchical structure but without any preamble ("articolato senza preambolo"), and those where the hierarchical structure is either not present or differently structured ("semi-articolato").

In the intentions of the working group, the two "articolati" can be used to describe all the documents that are created according to the "circolare" or according to the informal drafting rules that have always existed. These rules should cover most of the national legislation, an in particular national laws and many sorts of national decrees. The "semi-articolato" is meant for the (hopefully very rare) exceptions, and for all the documents that do not follow the structure or normative documents, either by their very nature or because they were created by institutions that are bound to neither the formalized nor the traditional drafting rules.

There are 12 classes of norms currently covered by the DTDs: "legge ordinaria", "legge costituzionale", "legge regionale", "decreto-legge", "decreto legislativo", "decreto ministeriale", "decreto del Presidente della Repubblica", "decreto del Presidente del Consiglio dei Ministri", "atto di authority", "decreto ministeriale non numerato", "decreto del Presidente della Repubblica non numerato" and "decreto del Presidente del Consiglio dei Ministri non numerato". There are further four generic document classes that can be used for all he documents that are not explicitly contained in the previous list. With time, the list of covered document classes will increase, but the document types will always be kept to manage those documents that do not deserve a class on their own.

The structure of the DTD

The NormeInRete DTD is organized in six main documents and in seven secondary ones. The secondary documents contain a large number of character entities used to manage special or dangerous characters (such as accented letters that can be coded differently on different operating systems).

The six main documents describe the structures of the DTD as follows:

1. Specific definitions of the Strict DTD in the file strict.dtd
2. Specific definitions of the Loose DTD in the file loose.dtd
3. Global DTD definitions in the file global.dtd
4. Special norm structures in the file norm.dtd
5. Textual, tabular or modular structures in the file text.dtd
6. Structures for the management of the meta-data in the file meta.dtd

The files global.dtd, norme.dtd, text.dtd and meta.dtd (and of course the character entity definitions), are used identically by both the Strict and the Loose DTD. The only differences between them are thus contained in the two files strict.dtd and loose.dtd, that simply include the other files.

Norm structures

The structure of a legislative document, as described in the file norme.dtd, is composed by a header, an optional preamble, a hierarchical structure (called "articolato"), a conclusion and a variable number of annexes. These are complemented by the initial and final formulas of the Italian law texts.

Each of these elements has its own characteristics and internal structure. Thus for instance the "articolato" is composed of a hierarchy of elements such as "libro", "parte", "titolo", "capo", "sezione", "articolo" and "comma". The structure "paragrafo" is not allowed by the "circolare", but is present in older texts. Thus it has to be present in both the Strict and Loose DTD, but it is not available in the Strict DTD.

Figure 1 and 2 show the symmetric parameter entities, located in the files strict.dtd and loose.dtd, for the element "articolato" and its subelements. Figure 3 shows the actual definition, shared by both the Strict and Loose DTD, placed in the file norme.dtd.

```
<!ENTITY % CMcompleto   "(libro+ | parte+ | titolo+ | capo+ | articolo+) ">
<!ENTITY % CMlibro      "(inlinemeta?, num, rubrica?, (parte+|titolo+|capo+|articolo+))">
<!ENTITY % CMparte      "(inlinemeta?, num, rubrica?,        (titolo+|capo+|articolo+))">
<!ENTITY % CMtitolo     "(inlinemeta?, num, rubrica?,                 (capo+|articolo+))">
<!ENTITY % CMcapo       "(inlinemeta?, num, rubrica?,             (sezione+|articolo+))">
<!ENTITY % CMsezione    "(inlinemeta?, num, rubrica?,                       (articolo+))">
<!ENTITY % CMparagrafo                                                            "EMPTY">

<!ENTITY % CMarticolo   "(inlinemeta?, num, rubrica?, decorazione?, (comma+))">
<!ENTITY % CMcomma      "(inlinemeta?, num, ((corpo | (alinea, el+, coda?)),
decorazione?))">
```

fig. 1 The parameter entities of the strict version of the "articolato"

```
<!ENTITY % CMcompleto  "(libro|parte|titolo|capo|sezione|paragrafo|articolo)*">
<!ENTITY % CMlibro     "(inlinemeta?, num?, rubrica?,
(parte|titolo|capo|sezione|paragrafo|articolo)*)">
<!ENTITY % CMparte     "(inlinemeta?, num?, rubrica?,
(libro|titolo|capo|sezione|paragrafo|articolo)*)">
<!ENTITY % CMtitolo    "(inlinemeta?, num?, rubrica?,
(libro|parte|capo|sezione|paragrafo|articolo)*)">
<!ENTITY % CMcapo      "(inlinemeta?, num?, rubrica?,
(libro|parte|titolo|sezione|paragrafo|articolo)*)">
<!ENTITY % CMsezione   "(inlinemeta?, num?, rubrica?,
(libro|parte|titolo|capo|paragrafo|articolo)*)">
<!ENTITY % CMparagrafo "(inlinemeta?, num?, rubrica?,
(libro|parte|titolo|capo|sezione|articolo)*)">

<!ENTITY % CMarticolo  "(inlinemeta?, num?, rubrica?, decorazione?, comma*)">
<!ENTITY % CMcomma     "(inlinemeta?, num?, ((corpo | (alinea?, el*, coda?)),decorazione?))">
```

fig. 2 The parameter entities of the loose version of the "articolato"

```
<!ELEMENT articolato           %CMcompleto; >
<!ELEMENT libro                %CMlibro; >
<!ELEMENT parte                %CMparte; >
<!ELEMENT titolo               %CMtitolo; >
<!ELEMENT capo                 %CMcapo; >
<!ELEMENT sezione              %CMsezione; >
<!ELEMENT paragrafo            %CMparagrafo; >
<!ELEMENT articolo             %CMarticolo; >
```

fig. 3 The shared definition of the element "articolato" and of its subelements

The element "comma" (the clause) contains the actual text of the norm, either as a textual body or as a list of text elements.

The annexes can be hierarchies, blocks of text, tables or also whole documents. According to the DTD it is possible to add a list of the annexes, and for each annex to specify a linkage formula to the hosting document (e.g., "Annex 1") and any pre-annex free text information. The annexes can be placed freely either within the hosting document, or as external documents themselves. This is extremely appropriate for annexes that are whole documents themselves.

One such case is the formal ratification of an international treaty. In this case, the body of the norm only contains a small and hardly interesting ratification text, and the actual treaty is placed as an annex to it. Since most users will be interested in the treaty, and not in the ratification, it makes sense to put the annex in an autonomous file, and refer to it within the real norm.

This module also contains the definitions of special inline elements having semantic and structural relevance. They include references, dates, places, documents, institutions and official subjects that may be worth identify in the document. The purpose of these elements is to normalise the details of the references, the values of the dates, the names of places, documents, bodies and subjects, without interfering with the text of the law or limiting the freedom of the legislator on the wording.

Textual, Tabular and Form Elements

Textual, tabular and form elements are defined in the file text.dtd. They provide two functions. First, they are used to describe special structures such as tables and forms, inserted both as annexes to the legal documents, and within the documents themselves. Second, they define structures having a typographic more than semantic value, such as paragraphs, bolds, italics, etc.

The first function is universally applicable, because tables and modules are frequently included in norm documents, whatever their nature. On the other hand the second function is very important for those types of documents (e.g., documents by authorities) or those parts of documents (e.g., preambles) that do not have a strict and formalised structure, and that can only be vaguely and generically described in terms of their typographical aspect.

Furthermore, it can happen that the original text presents particular typographical styles for some parts of text without an identifiable semantic reason. In these cases we propose the use of *ad hoc* elements, which reflect the typographical choice adopted and ignore any semantic interpretation. For these categories of elements the working group decided to use HTML elements, so that previous experiences and tools could be reused. The DTD uses elements such as "b", "p", "i", "table", etc. that everyone with even a cursory experience with HTML can already employ.

It should be specified that these are not really HTML elements: the list of available elements is limited, the number of usable attributes is much narrower and their use is more constrained than in HTML. Finally it must be noted that is possible to assign a specific typographical style to any element of the document, including the norm element, by associating it to a CSS class (Cascading Style Sheet, the stylesheet language used by the World Wide Web).

Meta-information

The NormeInRete DTD allows meta-data to be associated both to the whole document, and to any structural part of it. The file meta.dtd contains the meta-data elements that can be associated to the documents and document parts. These are of course only an initial set of meta-data elements, and further stages of the project will undoubtedly bring forth new and important meta-data requirements.

The meta-data elements are divided into five categories:

1. **Descriptors**: these are fundamental meta-data that are used to describe the document. These include for instance the formal publication of the document, its date, the associated URN, the possible aliases (other names by which the document is known), the relationships with other documents, the time frame in which the document is or has been in force, and several types of keywords for the description of the document.

2. **Preliminary works**: free text to include information and documents relevant with the drafting and approval stages of the norm.

3. **Proprietary**: an unstructured element where it is possible to add any kind of element and text, relevant to the specific application for which the document is being created. This may include elements that are relevant only to the specific database in which the document is maintained. They are defined by the maintainer of the database and are placed in an autonomous file which is included in the meta.dtd module.

4. **Editorial data**: the editorial staff may want to add any kind of information about the text, including notes, comments and other references.
5. **Dispositions**: these are a special type of meta-data providing an early semantic analysis of the norms of a text. They may be used to identify the clauses that contain a prohibition, an obligation, a sanction, etc. Some forty types of dispositions have so far been identified.

The DTD provides two places for the storage of the meta-data: either in the "meta" element at the beginning of the text, or in the "metainline" element anywhere in the structure. This does not mean that the working group suggests that all the meta-data are stored with the text of the norm. On the contrary, it only suggest the form in which these meta-data should be structured, but it gives no constraint on where to put them, or on how many sources of meta-data can be generated for each normative document. The definition of the main elements of the meta-data module is shown in figure 4.

```
<!ELEMENT  meta         (descrittori,
                         lavoripreparatori?,
                         redazionale?,
                         proprietario*,
                         disposizioni?)>

<!ELEMENT inlinemeta    (redazionale?,
                         proprietario*,
                         disposizioni?)>

<!ELEMENT descrittori   (pubblicazione, urn+, alias*, vigenza+, relazioni?,keywords*) >
<!ELEMENT lavoripreparatori   %blocchi; >
<!ELEMENT redazionale   (nota | avvertenza | altro | %Rproprietario;)+ >
<!ELEMENT proprietario   %ProprietarioMeta; >
<!ELEMENT disposizioni  (caratterizzanti?, analitiche?) >
```

fig. 4   The definition of the main elements of meta.dtd

Some experiences in Europe and world-wide

Numerous but diverse are the existing experiences on such matters in Europe and in the world. As at mid-2001, law databases in XML and SGML format exist, in different stages of development, in several countries of Europe. Some are managed by public institutions (Ministries of Justice being prominent), and a large number of private publishers also use these formats.

Among the public institutions, SGML is used in Finland for the Raske project of Parliament, as well as for the Finlex database of the Ministry of Justice, in the UK for the Statute Law Database of the Lord Chancellor Department, in Denmark for the Retsinformation database of the Ministry of Justice. Some minor databases in other countries are also in SGML. In different ways, all of these experiences are being studied for a conversion to XML, although in many cases the projects are still at a very early stage.

In the Netherlands the legislation is also in SGML, but the transition to XML is fairly advanced. The Ministry for Internal Affairs has recently contracted out to a private publisher the Basis Wetten Bestand project (Basic Legislation File). This publisher shall publish in a central database the complete legislation of the country in XML format. The project is forecasted to last five years.

In Sweden XML has been adopted for the Rixlex data base, managed by the Swedish Parliament with draft laws and meeting minutes (Uris project).

Similarly advanced is the situation in France: the contract has been recently awarded, and it contains, among several other activities, the conversion to XML of the main public databases Légifrance and Jurifrance within a couple of years.

Outside of Europe, fairly advanced is Canada, with the LIMS project, that aims at managing with XML the drafting, printing and Web publishing of the consolidated legislation of the country. The system is in advanced testing phase and will become operative at the end of 2004. The Ontario and Québec governments are developing similar projects for the regional legislation. The United States are working to produce in XML both Bills and Resolutions, and the DTDs are freely accessable at the site http://xml.house.gov.

The small state of Tasmania, in Australia, has all of its legislation in consolidated text in SGML. The system, called EnAct, manages drafting, management, consolidation and publishing of the legislation. An analogous project is underway in New South Wales, another state of Australia.

Finally, we wish to mention that the European Community is starting to work in this direction. The Eulegis project of the EU Commission is meant to provide a single interface for all the legislative information available in the European Economic Space. At the moment it is still in the design phase.

Among legal private publishers, many are already working with XML, especially in smaller countries such as Sweden, Denmark and Austria. It is interesting to note that the approaches to XML vary considerably: a Danish publisher developed more than 80 DTDs to cater for the whole national legislation, one for each type of normative document. Another publisher, in Sweden, developed just one DTD, and rather simple at that, for the Swedish legislation. The details of these DTDs are obviously protected by copyright.

Acknowledgements

The authors wish to thank all the participants to the DTD working group of the NIR project, as well as all the important feedback of testers, members of the other working groups of the NIR project, and the early implementers of systems based on this DTD.

References

PRESIDENZA DEL CONSIGLIO, *Circolare del 20 aprile 2001 recante Regole e raccomandazioni per la formulazione tecnica dei testi legislativi,* Gazzetta Ufficiale n. 97 del 27 aprile 2001. http://www.senato.it/funz/draf/home.htm

Regole e suggerimenti per la redazione dei testi normativi (also known as "Manuale Rescigno", 1991), in CAMERA DEI DEPUTATI, *Le direttive di tecnica legislativa in Europa* (R. PAGANO editor), 2 voll., Roma, 1997.

DE GIORGI Rosa Maria, *L'accessibilità alla legge: il progetto Norme in Rete*, "Notiziario di informatica della Camera dei Deputati", n. 2, 2001.

LUPO Caterina, *Norme in rete: stato dell'arte*, "Notiziario di informatica della Camera dei Deputati", n. 2, 2000.

ISTITUTO DI DOCUMENTAZIONE GIURIDICA DEL CNR, *Studio di fattibilità per la realizzazione del progetto "Accesso alle norme in rete"*, special issue of the journal "Informatica e diritto", 2000.