Eunomos, a legal document management system based on legislative XML and ontologies
(position paper)

Guido Boella
University of Torino
and Nomotika s.r.l.
Italy
guido@di.unito.it

Llio Humphreys
University of Torino
Italy
humphreys@di.unito.it

Marco Martin
University of Torino
Italy
notmart@gmail.com

Piercarlo Rossi
Università del Piemonte Orientale
Italy
piercarlo.rossi@unipmn.it

Leon van der Torre
University of Luxembourg
Luxembourg
leon.vandertorre@uni.lu

ABSTRACT
In this position paper, we illustrate the ongoing work and future developments on the Eunomos software, an advanced legal document management system to classify norms, based on legislative XML representation of laws which are retrieved automatically from institutional legislative portals, which complements the tool for building legal ontologies called Legal Taxonomy Syllabus.

1. INTRODUCTION
To operate efficiently, a law firm needs to regularly create and update legal documents, classify them according to the different domains it operates into, access reliable information on the state of the law and keep track of changes in legislation, regulations and contracts.

Currently much of this work is done by hand. Law firms employ personnel who assiduously travel through various sources to find relevant legislation and influential cases. Law management in law firms, in-house legal offices and law departments is today more complicated than ever due to the number of laws that have to be considered from various sources at many different levels - international, European, national, regional, or even internal regulations and standards. The complexity of researching the state of the law on a particular topic from various sources is difficult not only for businesses in countries such as Italy, which is well known for legislative over-production, but for any business that needs to operate in an international context and deal with multiple legislations.

Another problem is managing regulations, contracts and other documents. Most law firms today do not use dedicated document management systems even for their own legal documents. Law firms typically use folder trees as repositories and folder names as classification tags. They use master contracts to help formulate actual contracts for clients, but no links are made between elements in master contracts and derived contract instances. Different versions of contracts are often maintained using basic versioning features of standard editors for word processing.

Legal document management is more complicated than general document management, particularly regarding the requirement to continually review documents in the light of regulatory changes. This requirement means that documents need to keep track of the laws referred to by the various parts. Such links may be implicit or made explicit in the text using a legal reference. In the first case, the document must be annotated with references by a knowledge engineer. Once all legal references are made explicit, it is possible to automatically identify documents that need to be revised because the legal text referred to has been changed. Most commercial legal document management systems fail to address these issues. As far as we know, no commercial legal document management systems offer an integrated workspace for classifying relevant laws, drafting legal documents and monitoring changes in laws, norms and concepts.

Lately, articles have begun to appear in specialist and even mainstream press about an increased interest in bespoke ITC solutions, and in particular, human language technologies, for legal domains.

There are several factors that enter into play at this point in time:

2 http://business.timesonline.co.uk/tol/business/law/article7005373.ece
The research question of this paper is thus: How to extend knowledge management systems to deal with specific needs of lawyers and law scholars?

The methodology we use is to take inspiration from the technologies developed in the related fields of legislative drafting for parliaments, so called legislative XML, and legal ontologies, and export them in the context of applications for lawyers and law scholars.

In this paper we illustrate for the first time how this methodology resulted in the Eunomos software being developed in the context of the project ICT4LAW\(^3\) to address these needs, and compare the product with other systems in the field. Eunomos is an advanced legal document management system based on legislative XML representation of laws which are retrieved automatically from institutional legislative portals, and extending a tool for building legal ontologies called Legal Taxonomy Syllabus\(^2\)\(^,\)\(^1\).

In the next section we describe the background technologies from which Eunomos emerged: legislative XML and legal ontologies. In Section 3.1 we describe the main functionalities of the software and in Section 3.2 we describe the semi-automated classification mechanism of Eunomos. Related and future work and conclusions end the paper.

2. BACKGROUND

2.1 Legislative XML

In many regions in Europe and beyond, there are now online portals making laws and decrees available to citizens. These portals are updated on a regular, often daily basis. Some initiatives, such as legislation.gov.uk by The National Archives in the UK aim to go beyond being a legislative portal, providing a co-operative editorial tool, thereby giving others a stake and an incentive to work with them to create and maintain, open, free to use, up to date revised legislation. Every document published on their website will be available in machine readable XML format, as well as PDF.

Over the last few years, several XML specifications for legal documents have arisen in Europe with a view to make laws accessible to citizens and suitable for processing by specialist applications. Examples of legislative XML in use are FORMEX for the EU Publications Office and the NormeInRete (NIR) project defining several DTDs for Italian legislation and identifiers through URNs (Uniform Resource Names). In Denmark the government is working on LexDania and the Swiss and Austrians are also busy trying to provide better access to their legal sources with the use of XML. Boer and Winkels\(^4\) argue that there is a need for an interchange XML standard for describing legal documents. Such a standard should be language and jurisdiction independent, but law specific. It should enable external knowledge models about (the content of) legal documents to link to text from the original sources at the right level of granularity - i.e. legally relevant subparts. The XML standard they developed, CEN Metalex, is an interchange format which defines standards for naming conventions and cross referencing for information exchange and interoperability. It is not intended to replace jurisdiction-specific standards and vendor-specific formats.

The NormeInRete (NIR) standard is a well-established legislative XML used in Italy. It specifies the structure of legal documents in terms of XML tags for metadata, articles, paragraphs, etc. and that such components of legislation should be identified through URNs (Uniform Resource Names). URNs are designed specifically for the Internet community to provide unique identifiers, unambiguous and persistent network resources, regardless of their physical location. Assigning a uniform name for each legal document aims to assign a unique identifier, in a standardized format, which depends only on the characteristics of the document itself and is independent of availability in the network, physical location and means of access. This identifier is used as a tool to represent references - and more generally any kind of relationship - between acts. It facilitates the construction of a global hypertext among legal documents in a network environment with computer resources distributed among several publishers. It also allows the construction of knowledge bases containing the relationships between these documents.

An URN for a document constructed according to the NormeInRete standard will have the following components:

1. An ID for the original document, comprising the authority responsible for publishing the law (e.g., Ministry, Region, City, Court), the type of measure (e.g., law, decree, order, decision, etc.), the date and number and IDs for any annexes.

2. A version identifier, including the date of issue.

3. The ID of the press publishing the law.

4. An identifier of the fragment of the resource itself the URN refers to (e.g., article, paragraph, etc.). The URN can be used in a HTML (<META name="nir.urn" content="urn:nir:stato:legge:1996-12-31:675">) or XML (<urn valore="urn:nir:stato:legge:1996-12-31:675">) file.


\(^2\)Legal Taxonomy Syllabus

\(^1\)An URN for a document constructed according to the NormeInRete standard will have the following components:

\(^4\)Boer and Winkels
2.2 Ontology

The main assumptions of the Legal Taxonomy Syllabus ontology on top of which Eunomos is built come from studies in comparative law [17] and ontologies engineering [12].

- **Terms**—lexical entries for legal information—are distinguished; for this purpose we use lightweight ontologies [8], i.e. simple taxonomic structures of primitive or composite terms together with associated definitions.

- We distinguish the ontology implicitly defined by EU Directives (EUD), the EU level, from various national ontologies. Each national legislation refers to a distinct national legal ontology. We do not assume that the transposition of an EUD automatically introduces in a national ontology the same concepts that are present at the EU level.

- Corresponding concepts at the EU level and at the national level can be denoted by different terms in the same national language.

A standard way to properly manage large multilingual ontology is to make a clear distinction between terms and their interlingual acceptions (or axes) [19, 13]. The basic idea in our system is that the conceptual backbone consists in a taxonomy of concepts (ontology) to which the terms can refer to express their meaning. We do not assume the existence of a single taxonomy covering all languages. In fact, the different national systems may organize the concepts in different ways. For instance, the term contract corresponds to different concepts in common law and civil law, where it has the meaning of bargain and agreement, respectively [18].

In most complex instances, there is no correspondence between terms-concepts such as frutto civile (legal fruit) and income, but respectively civil law and common law systems can achieve functionally the same operational rules thanks to the functioning of the entire taxonomy of national legal concepts [9]. Consequently, the Legal Taxonomy Syllabus includes different ontologies, one for each involved national language plus one for the language of EU documents. Each language-specific ontology is related via a set of association links to the EU concepts.

3. DESCRIPTION OF EUNOMOS

3.1 Features of Eunomos

We have developed a sophisticated legal document management system based on ontology and legislation monitoring system called Eunomos with the following features:

- A large database of laws (about 70,000 Italian national laws in the current demo) maintained in XML format in accordance with the NormeInRete (NIR) standard for Italian laws.

4 The software does not depend on the specific NIR DTD, and can be used for other XML standards for other languages.
• Automatic downloads of laws from institutional legal portals via dedicated spiders. Currently the software harvests the Italian national portal http://www.normattiva.it including over 50,000 laws, the portal Arianna of Regione Piemonte http://arianna.consiglioregionale.piemonte.it/ and a portal of regulations from the Italian Ministry of Economy.

• The conversion of laws into NIR XML if they are in pure textual format.

• Automated parsing of legal references using the URN format of NIR. This enables legal references to be transformed into hypertext links to the relevant legislation, thus facilitating automated linking and reasoning and user navigation.

• Semi-automated classification of laws at the level of paragraphs or articles according to domains specified by the expert user.

• An alert messaging system, using URN references and semantic similarity tools, that informs users of new laws downloaded into the database and suggests which existing laws could be affected by the new legislation.

5 The Arianna portal already exports documents to NIR XML format. The conversion in the current version of the software is done using the XMLeges Marker tool developed by Istituto di Teoria e Tecniche dell’Informazione Giuridica (ITTIG) of Florence (http://www.xmleges.org).

6 This is done using the XML Leges Linker tool developed by ITTIG.

• Enabling concepts from the Legal Taxonomy Syllabus ontology to be linked via URN to legal definitions within relevant legislation.

Figure 1 shows the legislation search page. The user can search legislation via name, year, or URN. The research results are displayed below in the table below the search box. The first column contains the name of the law, and a link to the full text of the relevant legislation. The second column contains a summary of the law. If coordinated versions of the norms are available, they are shown besides the original ones. The navigation on the right hand side enables the expert user to view paragraphs and articles relevant to a particular domain, view similar pieces of legislation, analyse usage of terms within the legislation, and make links between terms within legislation and concepts in the ontology.

Figure 2 shows a concept with its place in the taxonomy and a link to relevant legislation (with a link expressed as a URN to the shown article). The ontology can be created contextually to a piece of legislation, thus facilitating the creation of the link and of the description. Here we find the concept of vehicle, and sub-categories such as trolleybus, motorcycle etc. By clicking on the plus/minus signs, the user can view definitions and references for each concept displayed in a table below.

Ontology and legislation document management is designed to be an online service provided by Eunomos to several clients, information and costs are shared. Another advantage of having several clients using the model is that with more people using the system, the higher higher the like-
Figure 3: The law classification form.

lihood that errors are quickly detected and corrected. Pu-

Table 3: The law classification form.

<table>
<thead>
<tr>
<th>Article 1, comma 1</th>
<th>nuove normative aggiunte</th>
</tr>
</thead>
<tbody>
<tr>
<td>Art. 5.</td>
<td></td>
</tr>
<tr>
<td>Modalità di pagamento della tassa</td>
<td></td>
</tr>
<tr>
<td>(Type reference: Tassuro v.)</td>
<td>Registro legge del 26 luglio 1939, n. 121, art. 2 (8 comma),</td>
</tr>
<tr>
<td>(Type reference: Tassuro v.)</td>
<td>Decreto legislativo del 7 maggio 1949, n. 109, art. 7 e 8.</td>
</tr>
</tbody>
</table>

La tassa di circolazione a stabilimento in ragione di anno solare.

<table>
<thead>
<tr>
<th>Art. 6</th>
<th>(Type reference: Tassuro v.)</th>
<th>Decreto legislativo del 7 maggio 1949, n. 109, art. 7 e 8.</th>
</tr>
</thead>
</table>

1. some laws include various norms for a variety of different and unrelated topic areas;
2. some laws contain norms that implicitly override norms in other laws, but fail to include references to the norms they override.

Eunomos uses natural language technologies to help the expert user with the labour-intensive work of categorisation of norms and retrieval of implicit references. The support is based on two techniques: analysis of outgoing references, and semantic similarity. The Eunomos product provides suitable interfaces for the expert user to create a set of category labels representing domains like taxation, immigration, etc. and to associate each component of a law (identified by a URN) to a particular category. In Figure 3 we can see annotated articles from a piece of legislation. The expert user uses this interface to assign domains to each article and a type (modification, overriding, etc.) for each reference to other legislation. Terms which name concepts in the ontology are highlighted.

Where articles and paragraphs contain references to the articles and paragraphs they talk about or override, this information is used not only to link the relevant legislation via URN, but also to suggest to which category a new piece of legislation belongs. The rationale is that where paragraphs or articles contain references to classified paragraphs or articles in previous legislation, it is more than likely that the new paragraph or article belongs to the same domain. The user can check and deselect the suggested classifications.

For articles and paragraphs that do not contain explicit references, it can be useful to find relevant domains and implicit references by referring to a list of the ten most similar pieces of legislation in the whole database. Eunomos generates this list using Cosine Similarity text classification.

4. RELATED WORK

Our solution has some similarities with Pazienza et al. [4] but has a different aim, since it is not a precompetitive project, and is more wide-ranging in scope. While Pazienza et al. [4] takes XML files as input, Eunomos downloads text-based laws automatically from portals and converts them into XML, generates automatic alerts concerning possible legislative updates, and identifies norms and concepts within
new laws which can be integrated with a sophisticated, multilevel and multilingual ontology tool. The use of ontology in the two systems are also quite different. Pazienza et al. [4] use the Semantic Turkey [10] ontology, where definitions can be taken from any source and arranged in any order. The Eunomos product is more careful, encouraging the expert user to create links to definitions in legislation, judgements and official journals, and to track the evolution of terms in a systematic manner. On the other hand, Eunomos requires considerable maintenance work, as web spiders need to keep up to date with any modifications made to online legal portals, and expert users are required to verify classification and find implicit references. Pazienza et al. [4]’s text similarity tool working at a paragraph level is very interesting, and we intend to add a similar feature in the next development phase of our product.

It is instructive also to refer to de Maat et al. [6]’s comparison of machine learning versus knowledge engineering in classification of legal sentences, since Eunomos uses machine learning and knowledge engineering techniques. de Maat et al. [6] uses knowledge engineering to find standardised patterns suggestive of a particular class, while we use knowledge engineering to find standardised patterns for references to classified norms in previous legislation, which provides a clue as to the classification of new norms. On the machine learning side, de Maat et al. [6] uses Support Vector Machines for text classification, while we use Cosine text similarity to find the most similar pieces of legislation, which provides clues on relevant domains and norms that may be overridden implicitly. The conclusion of de Maat et al. [6]’s research (ibid, page 16) was that ‘a pattern based classifier is considered to be more robust in the categorisation of legal documents at a sentence level.’ However, the classification task is quite different since that research was concerned with classifying sentences as norms, delegations, penalizations, value assignments, application provisions etc, while our classification task is to categorise norms as belonging to domains such as taxation. The author (ibid. page 14) noted that Support Vector Machines were better than knowledge engineering at categorisation where word order was less restricted, and as such may be more suitable for our work.

Concerning text classification techniques, there are a number of different solution to evaluate [14]. They work on the principle of labelling a collection of documents in various categories, training classifiers on the various categories, and using these classifiers to select the most appropriate topic for a new document. Most classifiers (Naive Bayes Classifier, Bernoulli, Vector Space Model) use as features keywords that have high frequency within the topic but not in general. Some implementations remove stop-words. Some give different weights to different keywords in terms of how representative they are of the topic. Mutual information and Chi2 are popular measures for ranking keywords. Compression-based classifiers are usually character-based. Adaptive Prediction by partial matching (PPM) is a lossless compression technique that assigns different code lengths to different letters based on their frequency within a document. The optimum coding will vary for each language, sub-language and topic. A new document can thus be classified by selecting classifiers trained on a collection of related documents that can compress the new document most efficiently [7]. Of the keyword-based classifiers, the Vector Space Model is widely regarded as the most accurate, but is also the most computationally expensive. Bingioli et al. [5] achieved an accuracy of 92% in the task of classifying 582 paragraphs from Italian laws into ten different categories.

5. FUTURE WORK

Eunomos is an ongoing piece of work, and we are always interested in finding promising technologies that we can include in our research and products. Eunomos could be improved by applying text categorisation besides text similarity techniques. During the construction phase of the Eunomos database of norms, we did not have much labelled data, and the Cosine text similarity technique was useful for suggesting domains for unclassified norms as well as for finding norms that implicitly override other legislation. In developing and testing the Eunomos system, we are building more and more labelled data, and we will soon be in a position to use this data to bootstrap a new topic-based classifier for paragraphs and articles. Text similarity for finding norms that implicitly override other legislation could also be more useful at paragraph and article level. But the task is more challenging with shorter text, and we need to compare the Cosine Similarity with other algorithms such as Latent Semantic Analysis. The WEKA toolkit [11] contains various machine learning algorithms for text categorisation and text similarity which we can use for our tests. These new requirements place a high performance demand in terms of precision, recall and speed, and careful analysis is required to select the most appropriate technique for each task. Given the size of the database we must firstly take into account efficiency considerations. To cope with the problem of the size of the dataset since laws are considered paragraph by paragraph, we propose to build topic-based classifiers on a small subset of representative norms that have already been classified. We may use results from text categorisation to aid text similarity and vice versa. For example, references to classified norms can be included as a factor in the text categorisation algorithm. Even in cases where a norm refers to a general law containing several topics rather than to an article assigned to a specific category, the reference can be included as a factor and the classifier should be able to take the information into account and assign it the appropriate weight. On the other hand, more efficient retrieval of similar norms could be achieved by limiting the text similarity searches to classified norms within the same domain as assigned by the topic-based classifier.

We will also be evaluating the accuracy of the automated translation of legal text into NIR XML, even if at first sight it seems sufficient for the requirements of the clients. Another development on the NIR XML side is to analyse explicit references. Currently Eunomos can find most explicit references but an expert user needs to specify whether the reference is a simple reference or whether it modifies or overrides other legislation. By incorporating the natural language technologies developed by Mazzei et al. [15], the type of modification can be discerned automatically.

Another area for future development is to exploit Eunomos’s potential to cater for multilingual and multilevel legal research, since some clients may be interested in specialist
databases for foreign legal systems. While Eunomos uses the NormaInRete standard internally, as standards are developed for interchange between different legislative XML formats [5], it should be possible to use Eunomos in other jurisdictions. This would require suitable parsers to structure laws in XML in different languages. It is already possible, however, to model EU directives and their national implementations, and the Legal Taxonomy Syllabus ontology is already multilingual.

The Eunomos software could also be adapted to manage contracts and other legal documents. An integrated document management system that incorporates legislation, ontology and contracts could be very attractive to law firms. After a change in legislation, changes may also need to be made to contracts. In some cases even signed existing contracts need to be checked to ensure that new regulations do not invalidate them. The same mechanism based on reference recognition to find regulations affected by modifications can be applied to contracts.

It may be possible to integrate editors designed for drafting legislation to draft and edit legal contracts. The Norma-System legislative XML editor developed by the Università di Bologna [16] works as a plug-in for Microsoft Word. Additional menu items make it possible to: create an XML version of documents valid for NIR DTDs, mark up the structures of documents with automatic tools, view and compare structures with a text mapping, consolidate documents in fully automated mode or manually, manage integrative, or informative acts as attachments, automatically recognize and mark up normative references. The open-source Bungeni editor for drafting legal text [7] has been designed to work with the Akoma Ntoso standards. The Word-like editor can be integrated with Open Office. The workspace has an attractive and user-friendly interface to enable legislative drafting staff to import and mark up debates and legislation, review metadata and create links between referenced legislation.

6. CONCLUSIONS
Legal informatics is experiencing growth in activity. There is a place for experimentation and cross-fertilisation of ideas from other domains. There is good research within legal informatics, knowledge management, natural language processing and artificial intelligence which can help make the legal process more effective and efficient. Now is the right time to apply this research to products for law firms and not just legislative bodies.

In this position paper we illustrate ongoing work on the Eunomos software. The software is being developed to support the work of law firms, in-house legal offices and law scholars by offering them an environment which makes laws easier to navigate, annotate and understand, using automatically generated hyperlinks to referenced legislation, an extensible and updatable ontology which provides current and previous definitions for norms and concepts within any specific context, and an alert system that specifies existing legislation affected by new legislation.

Eunomos is being developed as a commercial software part of a wider suite distributed by Nomotika s.r.l., a spinoff of University of Torino. Eunomos has a clear business model: a combined software and services package that effectively means that legislation monitoring is outsourced. The roles, permissions and technologies have been carefully selected to address real business needs. The software and related services will be provided by experts with sound technological and business expertise.

7. REFERENCES

http://code.google.com/p/bungeni-editor/


