UvA-DARE (Digital Academic Repository)

Proposal for a Dutch Legal XML standard

Boer, A.W.F.; Hoekstra, R.J.; Winkels, R.G.F.; van Engers, T.M.; Willaert, F.

Publication date

2002

Document Version

Author accepted manuscript

Published in

Electronic government : first international conference, EGOV 2002, Aix-en-Provence, France, September 2-6, 2002 : proceedings

[Link to publication](#)

Citation for published version (APA):

Boer, A. W. F., Hoekstra, R. J., Winkels, R. G. F., van Engers, T. M., & Willaert, F. (2002). Proposal for a Dutch Legal XML standard. In R. Traunmüller, & K. Lenk (Eds.), *Electronic government : first international conference, EGOV 2002, Aix-en-Provence, France, September 2-6, 2002 : proceedings* (Lecture notes in computer science; Vol. 2456). Springer.

General rights

It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations

If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: <https://uba.uva.nl/en/contact>, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.

Proposal for a Dutch Legal XML Standard

Alexander Boer, Rinke Hoekstra, Radboud Winkels¹,
Tom van Engers², and Frederik Willaert³

¹ Dept. of Computer Science & Law, University of Amsterdam, Netherlands,

² Dutch Tax and Customs Authority, Utrecht, Netherlands

³ Application Engineers, Belgium

Abstract. This paper⁴ presents a proposal for an XML Standard for legal sources in the Netherlands. The standard intends to provide a generic and easily extensible framework for the XML encoding of the structure and contents of legal and paralegal documents. It differs from other existing metadata schemes for legal documents in two respects; It is language-independent and it aims to accommodate uses of XML beyond search and presentation services.

1 Introduction

This paper presents a proposal for an XML Standard for legal sources in the Netherlands. The research is carried out in the context of the E-POWER and e-COURT⁵ projects and the results will be brought into the LeXML initiative⁶. The standard intends to provide a generic and easily extensible framework for the XML encoding of the structure and contents of legal and paralegal documents. This obviously includes legislation and case law, but also written public decisions, internal and external business regulations (for instance ship classification rules as in [7]), and contracts. XML elements and structure are defined in schemas that can be used to validate a document. Since there is a great variety of legal documents that cannot be covered by one normative standard, the standard consists of multiple schemas defining vocabularies that can be mixed in a document. While the standard aims to cover all possible legal sources, the focus of current work is on Dutch legislation: the 2001 Dutch law on income tax in the context of the E-POWER project, and the Dutch penal code of 1881 in the context of the e-COURT project. Later we will cover the structure of (Italian and Polish) court room transcripts (for e-COURT) and case law. The standard differs from other existing metadata schemes for legal documents in two respects; It is language-independent and it aims to accommodate uses of XML beyond search and presentation services.

⁴ Will appear in: Proceedings of the eGOV2002 Conference: State of the Art and Perspectives. Springer Verlag, Berlin.

⁵ See Acknowledgements at the end of this paper.

⁶ The LeXML initiative is the European equivalent of the Legal XML community in the Unites States.

1.1 Why a Legal XML Standard?

Clearly, legal documents serve a plethora of purposes, and for each purpose contain information that can be grouped in various general and specific categories. Many of the most common classification and reference systems even pre-date the storage of legal information on computers. XML promises dramatic improvements in the efficiency of managing and processing information in legal documents. XML and related technologies are essential parts of the more general idea of an integrated *semantic web*: A machine-readable and machine-understandable version of the current web coupled with new smart applications to exploit it. This machine-readable extension is called metadata: information about information.

Current XML schema and metadata definition efforts in the legal domain are concerned with government document locator services for general use like the British Legal and Advice sectors Metadata Scheme (LAMS) for 'Just Ask!'⁷ and Australian Justice Sector Metadata Scheme (JSMS)⁸. In addition, there are XML schemes used by legal publishers for document locator services. These schemes focus mostly on classification of legal documents using traditional differentiae like:

- The author and legal status of the document
- Creation, modification, and promulgation dates
- The jurisdiction to which it applies, and the language(s) it is available in

These attributes are rather crude in meaning, and the resulting classification is superficial; It leaves out a lot of relevant detail and its quality is questionable for automated reasoning. Identification of documents by jurisdiction, for instance, assumes that the user of a search service knows what jurisdictions he is in – and that jurisdictions can be meaningfully delineated. Establishing jurisdiction is however a domain of legal research in itself. The meaning of the dates is clearer and can be directly used to establish validity of a document, but only relative to a known jurisdiction.

In addition to these traditional categories documents are usually classified in substantive terms with an attribute identifying a 'domain' of the document in a fixed classification. Viewing these attributes as metadata, or extra information about the document, creates a potential maintenance issue: The values of these attributes may change over the lifetime of a legal document, even if the document itself does not, as the concepts employed in the document change over time and become associated to (disassociated from) other concepts (see e.g. [5]). In reality, the agent of such a change in the web will usually be the creation or modification of a new legal document with certain attribute values. The information about the information is often not extra: It was in the document in the first place, or in another document referring to it. Classification at a 'heuristic level' always places the burden of maintenance on human domain experts.

⁷ See e.g. <http://www.lcd.gov.uk/consult/meta/metafr.htm>

⁸ See e.g. <http://lawfoundation.net.au/olap/guidelines/metaintro.html>

Furthermore, it is hard to conceive of any ‘smart application’ using the classifications as more than filters on search results. This is not because the meaning of ‘jurisdiction’ or ‘appellate court filing’ cannot be made understandable to the computer, it is because the level of the classification presupposes that the user of the classification system can read the document to find out why the classification was attached.

1.2 RDF and Representation of Meaning

The alternative approach to such a domain classification, is the direct identification of statements in the contents of documents. This includes statements about other documents and (fragments of) the document itself. For this purpose the Resource Description Framework⁹ (RDF) was designed. In RDF, statements are encoded as as (*subject, predicate, object*) triples. By describing legal concepts in different jurisdictions in an RDF ‘dictionary’¹⁰ as conceptual prototypes, the LexML initiative assumes that it is possible to identify and describe similarities and differences between legal concepts in different languages. Existing schemes (like JSMS and LAMS) rely on compatibility with HTML’s META tag only allowing RDF-like statements about documents; In RDF terms that means that the subject of a statement is always a HTML document. The META tag can only be used to make statements about a document in the document itself.

The RDF data model is based on the concept of a statement and the concept of ‘quotation’ (or reification) of statements. Quotation is used to make statements about (reified) statements – the statement is treated as the subject or object of another statement. In addition to this data model additional RDF schema statements can be defined, whose meaningfulness depends on the extent to which they can be transformed to other schemas and whether there exist any applications that can validate statements against that schema with query or inference languages. Since RDF is serialized in XML documents, it is important to realize that the validity of an RDF statement is relative to an unspecified set of other statements in the memory of the validating application. There is no central authority guarding universal consistency or coherence of the semantic web and anyone can publish any truth, falsehood, opinion, or judgment.

The problem, obviously, is that of trust. Quotation allows one to create trust, because it allows one to express where something was stated, who stated it, who modeled a statement made by someone, and what level of guarantee they dare to associate with it. ‘Syndication’ of information is based on this notion of trust. You may be willing to pay for advice, for instance, if you trust the party that gives it even if advice on the same subject is also available for free from other parties. It is not merely about good intentions: Heuristic notions like ‘domain’ always remain open for disagreement. The quotation mechanism in RDF allows others to make qualifications about statements, solving another limitation of META tags.

⁹ See <http://www.w3.org/RDF/>. RDF, like XML is an open standard from the World Wide Web Consortium (W3C) that is well-supported with free software.

¹⁰ <http://legalxml.org/Dictionary/>

Fig. 1. RDF representation of legal documents.

Figure 1 shows the relationship we propose between the XML encoding of a document and the RDF encoding of the same document. If a document element is encoded as a set of RDF statements, any element of it can be subject, predicate, or object of statements *about* the document.

1.3 Aims of the Proposed Standard

This proposal is more complicated than comparable standards because it is motivated not just by concerns for ‘smart’ search, but also with document management and maintenance of existing and future decision support software used by public authorities like the fiscal applications developed by the POWER group in the Netherlands (cf. [1]). This proposal aims to standardize legal documents for the purposes of:

- Filtering
- Presentation
- Document Management
- Knowledge Representation
- Search
- Code Generation
- Rule generation
- Classification and Verification

Each meaningful element of the text up to a full sentence can be separately selected or changed by software through an XML API. Each element of the text can be marked with an ID attribute that is unique in the document. An element is easy to transform to (X)HTML (including hyperlinks) with familiar layout by application of XSL transformations. Because each element can have a unique identifier, it is possible to make external statements in RDF refer to the right part of the document. Only RDF, RDF Schema, and HTML support a widely accepted standard for reference anchors embedded in XML. The recent XLink specification released by the W3 Consortium creates a generic facility for

referring unambiguously to parts of XML documents, but is not yet supported by most tools.

The standard consist of coupled schemas expressed in RDF Schema and XML Schema that are equivalent in meaning. A two-way translator between the ‘basic’ XML standard and a more flexible RDF standard for software engineers is in development. RDF can be used to encode inverted file indices, association rules and, if time permits, self organizing maps (cf. [3]) to search for document elements using a concept index. A great variety of IR techniques requiring special indexing and active learning algorithms have been used for information retrieval on legal documents (see e.g. [6, 4] for an overview). RDF is syntactically also sufficient to represent description logics (in DAML+OIL) or UML. The standard thus allows use of existing UML and DAML+OIL editors for legal knowledge representation. DAML+OIL is an RDF Schema¹¹ that provides a number of standard ‘logical’ constructs for RDF descriptions (cf. [2]). The content of documents, e.g. norms and validity constraints, are expressed using the DAML+OIL vocabulary, because this vocabulary can be validated by free description logic theorem provers (FaCT and RACER) and translated to an expert system rule engine (JESS).

2 Description of Documents

We roughly distinguish three different viewpoints on how we look at legal documents:

Form In the Netherlands and in general a legal document can be ‘recognized’ and classified by certain required phrases and formulas. Formal requirements on structure mostly reflect considerations of consistency of language and ease of access for the reader, but it also provides a context for the interpretation of the content of the document. This latter role is very specific for jurisdiction and timeframe and not part of the basic standard. Generic structural desiderata are defined in XML schemas.

Role Although we may look at the phrases and formulas in a written decision to classify a document as a law, we know that it is not the structure of the document that makes it a law, but the role the document plays in the activities of public bodies - most importantly the activities that produced the document. This is captured in RDF statements ‘about’ the document.

Content We also classify documents depending on what its content means: It represents a type of decision. If it is just a public decision its meaning is limited to a particular occurrence or case. If it is a norm or policy its meaning extends to general class of occurrences or cases and it postulates a value theory for making and judging decisions. This is captured in RDF statements ‘about’ this content: acts, norms, agents etc.

¹¹ See <http://www.w3.org/SW/> or <http://www.daml.org> for DAML+OIL or e.g. <http://www.ontoknowledge.org/oil/> for its precursor OIL.

Fig. 2. Relations between components of the standard. The white arrows represent possible XSL transformations.

Figure 2 shows the relation between the various XML schemas that are part of the standard. An XML document on the bottom leftside that adheres to schema *Nederlands-document* uses the Dutch (Nederlands) XML element vocabulary for document structure and indirectly imports vocabulary *Nederlands-reference* and the structural requirements of *Standard-document* and *Standard-reference* – the ‘language-independent’ core vocabulary (in simple English). The language-specific schemas contain direct word-for-word translations (with the XML schema substitutionGroup element) to the standard vocabulary. An XML document is first transformed with an XSL transformation sheet to standard vocabulary, and the standard vocabulary can be transformed to HTML or RDF. RDF is used for further description of role and content.

Extensions built on top of the standard consist of relevant language-dependent vocabulary schemas, a standard schema that defines formal requirements, a simple XSL template that translates language-dependent schemas to standard vocabulary, an XSL template that translates the standard schema to standard-document, and optionally XSL templates for specialized presentations in HTML.

3 Terminology

Regardless of whether one sees the written law as an imperfect description of something else (*the Law*) or as a product of an social agreement that requires clear communication and predictable and foreseeable decisionmaking, it is clear that the written law refers to things ‘in the world’. To regulate a world it must be described, after all. To be able to describe the meaning of the document the standard specifies a general vocabulary that can be used to describe it in RDF for specific purposes and domains. This work is in progress.

The conceptual core of these ontologies describes the relations between the various public documents published by public bodies in a fairly generic terminology, based on the general administrative act (AWB) of the Netherlands, that positions public documents in a legal system in terms of:

Public body A public body or body created by an act of law to serve a public interest.

Decision A written decision of a public body to perform a public act using a public power assigned by law.

Power A permission to perform certain public acts in a public interest.

Assignment A public act that attributes a power to perform a public act to a public body.

Delegation A public act of a public body that transfers a power assigned to it to another public body.

Mandate A decision of a public body to allow another person or body to use its power, without transferring the power, to perform certain public legal acts.

3.1 Time and Change

The importance of capturing the relations between public legal documents is made apparent by considering the requirements for correct updating of a collection of documents in time. Changes in laws are announced in separate documents and publishers must keep track of all documents from certain publication channels to be able to reconstruct what the form of an organic law is at some time point. Similarly, if you find a written decision on your doormat its validity status changes when a new written decision retracting it follows two days later.

To keep track of versions the standard provides a number of attributes for every structural XML element in the document that can be identified, selected, and thus changed; The *date-publication* of an element is the time the element is officially published or announced. The *date-enacted*, the time the content becomes applicable in decisionmaking, is always later than or the same as date-publication, but before *date-repealed*, the time the content becomes inapplicable in decisionmaking. Between date-enacted and date-repealed the element and its content is *active*, and outside this interval it is *inactive*. The *date-version* attribute represents the date the correctness of the content and other dates of the XML element was last verified. The XML document loses its value as a normative reference as time progresses and the time-interval between date-version and today increases.

If an XML element in a newly published document refers to another XML element the content of the element may repeal, enact, or change the other element. A XML element may also refer to another XML element to invoke its ‘power’. This is the case if the official author of the document obtains the power to take the decision(s) communicated in the document from a specific assignment by law, delegation decision, or mandate decision. In such cases the element, or an ancestor element, becomes inactive if the element from which the power is obtained becomes inactive (at least in regulations from public bodies in the

Netherlands). The standard aims to provide representational primitives that adequately capture legally relevant acts of this nature on legal documents without commitment to a specific ‘model’ for updating documents.

4 State of the Proposal

We currently have XML schema definitions of the structure (form) of Dutch legislation, Dutch documents in general and their standard counterparts, i.e. effectively all that is presented in Fig. 2. We have tried it out on the new Dutch income tax law and are in the process of applying it on other legislation. We have also created XSL translations from the XML format of a major dutch legal publisher and an XML format supported by the POWER tools (cf. [1]). We are building a large ontology for the legal domain, concentrating first on fiscal and penal law. By the end of June 2002 a substantial part will be finished.

Acknowledgements E-POWER is partially funded by the EC as IST Project 2000-28125; partners are the Dutch Tax and Customs Administration, O&I Management Partners, LibRT, the University of Amsterdam (NL); Application Engineers, Fortis Bank Insurance (B); Mega International (F). e-COURT is partially funded by the EC as IST Project 2000-28199; partners are Project Automation, Ministry of Justice, CNR (I); Ministry of Justice (POL); Sema Group S.a.e. (SP); Intrasoftware International (L); Universit Paul Sabatier (F); University of Amsterdam (NL).

References

1. Engers, T. van, Gerrits, R., Boekennoogen, M., Glass’ee, E., Kordelaar, P.: POWER: Using UML/OCL for Modeling Legislation -an application report. In: *Proceedings of the 8th International Conference on Artificial Intelligence and Law (ICAIL 2001)*, pp. 157–167. ACM, New York, 2001.
2. Fensel, D., Horrocks, I., Harmelen, F. van, Decker, S., Erdmann, M. and Klein, M.: OIL in a nutshell. In: R. Dieng et al. (eds.) *Knowledge Acquisition, Modeling, and Management, Proceedings of the European Knowledge Acquisition Conference*. Lecture Notes in Artificial Intelligence, LNAI, Springer-Verlag, October 2000.
3. Kohonen, T.: Self-Organizing Maps. Springer Series in Information Sciences, Vol. 30, 1995. Springer, Berlin. Third edition 2001.
4. M-F. Moens, ‘Innovative techniques for legal text retrieval’, *Artificial Intelligence and Law*, **9**, 29–57, (2001).
5. E. Rissland and T. Friedman, ‘Detecting change in legal concepts’, in *Proceedings of the Fifth International Conference on Artificial Intelligence and Law (ICAIL-99)*, pp. 127–136, New York (NY), (1995). ACM.
6. H. Turtle, ‘Text retrieval in the legal world’, *Artificial Intelligence and Law*, **3**, 5–54, (1995).
7. R.G.F. Winkels, D. Bosscher, A. Boer, and J.A. Breuker, ‘Generating Exception Structures for Legal Information Serving’, in *Proceedings of the Seventh International Conference on Artificial Intelligence and Law (ICAIL-99)*, ed., Th.F. Gordon, pp. 182–195, New York (NY), (1999). ACM.