Digital Legislation: Reflections on the Agora-Lex Project

MARIE-FRANCINE MOENS[1]
MAARTEN LOGGHE[2]

[1] Interdisciplinary Centre for Law & IT
Katholieke Universiteit Leuven
Tiensestraat 41
B-3000 Leuven
Belgium
e-mail: marie-france.moens@law.kuleuven.ac.be

[2] JCS International, Belgium
Interleuvenlaan 62, zone 2
B-3001 Leuven
Belgium
e-mail: maarten@ulyssis.org

Abstract

Legislative databases are a well-established information source for the legal professional. However, the manual updating of the texts and their reference data or metadata that are necessary for an effective consultation of the databases is very expensive and does not always guarantee the correctness of the retrieved texts. A valuable solution to the maintenance problems of the databases is more automation in the lifecycle of legislation. This can be achieved by the official electronic publication of legislation including the texts (original and consolidated) and their reference data, and by incorporating intelligent techniques for drafting, indexing and hypertext linking. Such an approach would allow for an automated update of the texts and the reference data in the databases.

1 Introduction

Legal professionals commonly consult databases that store the texts of legislation. Many of the databases are accessible through the World Wide Web and are freely available for citizens who have Internet access. The databases offer consolidated legislation, i.e., legislation in its actual status, i.e., the laws and regulations in their updated, modified form (which is called *point-in-time access to consolidated legislation*). This primarily refers to the text of the statute that is valid at a certain date. Such a text is built with the statute components (such as the articles, chapters, etc.) that are valid at that moment in time. The system preferably also gives an overview of the history of an article or other statute component and the justifications of the modifications that the article or component has undergone in the form of pointers to the texts of modifying acts. Important requirements of the legislative databases are their completeness, i.e. offering consolidated legislation of all domains of law, the correctness of the retrieved texts, and the information offered being up to date.[1] These requirements are currently not always fulfilled.

In the frame of the Agora-Lex project[2] we have studied legislative databases in Belgium and abroad. This study has revealed the cause of the problems of current databases. The manual acquisition of the data and their input in the database are tedious and expensive tasks requiring skilled specialists. Updating the databases usually does not keep up with the rate legislative texts are published or modified. Moreover, maintenance of the databases by humans is error-prone making the retrieved texts regularly unreliable. Finally, if the bases offer classifications and hypertext links to optimise the search, the links with the classification codes or other texts must be maintained for all the versions, which again is an expensive manual job. These problems are only worsening by the increasing growth and complexity of the body of legislation.

[1] See a recent recommendation Rec (2001)3 of the Committee of Ministers of the Council of Europe (http://cm.coe.int/ta/rec/2001/2001r3.htm).

[2] The Agora-Lex project (1999-2000) has been funded by the Belgian Federal Office for Scientific, Technical and Cultural Affairs. The project was realised by the Interdisciplinary Centre for Law & Information Technology (ICRI), Katholieke Universtiteit Leuven, Belgium in collaboration with the Université Libre de Bruxelles, Belgium. We wish to thank Prof. Dr. Jos Dumortier, Prof. Dr. H. Gilson, Kris Van de Kerckhoven, Dirk Keymis, Caroline Allard, Marc Albert, Anne Spoiden, Serge Declerck and Luc Libert for their contributions or suggestions to the Agora-Lex project.

This paper proposes answers to these problems. It is organised as follows. We shortly describe the work on the Agora-Lex project and the inherent complexity of Belgian legislation. In a next section we investigate the possibilities of the digital production and publication of legislation. Then, we look more closely into the possibilities of an automated classification and linking of the legislative texts. We conclude with the recommendation for more automation in the complete lifecycle of legislation.

2 The Agora-Lex project

Belgian legislation is more complex than in most other countries. As in any civil law country Belgian legislation is subject to multiple modifications resulting in many versions of an article and making the construction of the text of a statute as it is valid at a certain moment in time a complicated task. Moreover, Belgium has different territorial areas of application of statutes and different regional governments. For instance, a regional government amends a statute that is valid in the whole country. Consequently, from that instant onwards certain articles of this statute have two different versions each valid at the same time, but in different communities and regions of the federal state. Finally, Belgium has three official languages: Dutch, French and German.

Legislative databases have to cope with these difficulties. The Agora-Lex project has built and tested a model system for the version management of legislation.[3] The model allows requesting the valid text of a statute at each moment in the history of the statute as it is valid in a territorial area. The text is retrieved in the selected official language. The model was tested extensively by storing and querying the Belgian Judicial Code and the Law of the Protection of Youth of April 8, 1965. Both statutes that are known for their complex modifications.

We have built a prototype system for the version management of legislation in the form of a relational database. The relational model presents a database as a set of relations. A singular relation is represented as a table of values in which each row (tuple) represents a set of values of related data. These values can be interpreted as the description of an entity or the instance of a relation. The relational model was chosen because of its reputation of being performant, having a good query language, namely of *SQL* (*Structured Query Language*) which can solve complex queries by means of a simple syntax, and of guarantying a good consistency of the answers by means of built-in integrity constraints based upon relational algebra. The Database Management System Informix was chosen because of its availability at our research center.

The most important entities in the model are the statute, the modification (needed for version management), the statute component, the dispositif or body of the document, the rubric and article. The smallest entity that is considered in the management of legislation in the Agora-Lex prototype is the article. We did not model smaller content elements, such as the paragraph, because of the rare modifications on paragraph level. Articles may be hierarchically grouped in rubrics such as sections, chapters and books. Other major components of statutes are the formal parts such as the heading, the salute of the king, introduction, proclamation and signature. Statutes also comprise supplements and errata. Articles, rubrics, supplements and the headings of rubrics may be modified. A modification concerns insertion, deletion and/or updates of the texts and/or the references of a statute component, and causes a new version of the component.

In a constitutional state, it is important that legislation is accessible for every citizen. For the Agora-Lex prototype we have chosen a Web-based technology for querying and maintenance of the system because it has a widespread acceptance in the society and can be used by any institution, company or private person. The Web-based interfaces are developed in Java (Java Servlet technology) in order to guarantee a powerful and portable solution. An HTML (*HyperText Markup Language*) web page in the language of the user guides the querying and maintenance of the database.

The user interface of the Agora-Lex prototype allows querying the statute references (e.g., query based upon the title of the statute, query for the version of the statute that is valid at a given date). The interface can be expanded with other search functionality such as a full text search, a search based on thesaurus class terms and classification codes, or an interface for natural language queries.

[3] We refer to more detailed descriptions of the Agora -Lex project in Logghe, Van de Kerckhove, & Moens (2000), Van de Kerckhove, Logghe, & Moens (2000), Moens, Logghe, & Dumortier (2002).

The Agora-Lex research has built an intelligent maintenance interface and a large-scale anticipation and automation of the data input of complex cases. In order to support all complex cases and to guarantee an efficient search of the data, maintenance with no automatic support would be a tedious job. Intelligent software components make that data input is restricted to an absolute minimum, while automated updates are maximally exploited. Conversion of natural language titles of articles and other statute components to a consistent numbering employed by the system is done automatically. Automated support is given for the input of complex cases (e.g., splitting of versions when two versions become valid in different regions).

The Agora-Lex project has learned us some fundamentals for the management of digital legislation. Legislation is a very complex matter. The drafting of legislation is historically grown based on manual habits and has failed to take into account that because of its current overload legislation must be automatically managed. An important difficulty in building legislative databases is that current legislation is made for use in paper and print and not for an effective and efficient storage in and retrieval from databases. Nor is legislation made for being part of a workflow in which documents are electronically exchanged.

A substantial requirement for the use of digital legislative documents is that they are uniquely defined. This requires at least using a transparent and uniform identification of the data objects. Statutes, statute components and articles, and versions must be named in a clear and uniform way, which makes their electronic management easier. Using a unique identification frame when drafting legislation refers to the formal aspects of the statutes. Uniform rules for drafting legislation are not commonly used. Belgian institutions (mainly the Council of State) prescribe less variety in the formal aspects of statutes. Their drafting recommendations are rarely applied in practice, also due to the fact that the form of statutes is often part of political compromises. Belgian legislation is characterised by the wide variety in the presentation of the regulations drafted by different law making authorities and institutions. Therefore, one can still find a lot of variation in numbering articles, naming chapters, working with paragraphs and enumeration. Identifying the various documents and within the documents, the different components and articles is one already complex task, but identifying every version of a statute component is even more complex. This makes a clear, transparent method of identifying the data very difficult both manually and automatically.

The identification frame is essential for the basis key structure of legislative databases. This key structure allows accessing the main entities of a database, and exchange of data between bases or in the workflow when preparing legislation. It is attempted that this key structure is isomorphic with the identification frame that the legislator himself provides in order to promote the transparency and acceptance of the model and the communication with existing databases.

Another important finding of the Agora-Lex project is that maintenance problems are a major barrier for constructing comprehensive legislative databases. A study of existing Belgian legislative databases and our own experience of inputting the test data for the Agora-Lex model showed that manual maintenance is a very expensive and error-prone task. Databases need consolidated versions of the texts. The task of manually consolidating the texts based on the modifying statutes is particularly expensive. Maintenance is an important threat to preservation of digital archives in general (McCray & Gallagher, 2001) and especially to legislation, which is subject to multiple and even daily modifications. When no efforts are made to maintain the databases or when maintenance becomes an impossible task due to its magnitude, the databases become obsolete and even will disappear. We belief that legal information systems that manage legislation will always be crippled when nothing is undertaken in the field of the production and publication of legislation and in a larger automation of its complete lifecycle. We look more closely into these topics in the next sections.

3 Drafting and Publication of Legislation

3.1 More Automation in the Complete Lifecycle of Legislation

For an efficient management of documents in digital format, an automated support during the complete lifecycle of the documents is necessary (see figure 1). This support must be incorporated in the creation and drafting process, exchange of documents, publication, archiving in databases and retrieval. For legislation this support is equally needed for the management of the different versions in all of the above tasks. More specifically, the drafting of the texts and their modifications must guarantee the technical qualities of the documents (e.g., correct structure, identification codes, mark-ups for essential reference data) and an appropriate electronic exchange format necessary for a subsequent publishing and data base storage. Legislation preferably is officially published in digital format requiring the necessary support for guaranteeing the quality, authenticity and a long

term preservation of the documents. Upon their publication, the official versions of the texts and of their reference data can immediately be stored in the databases, some of which can be located on public web servers. When a department wants to modify a certain article of an existing statute, it has access to the correct, official, latest version of this article in digital format. Modified text versions and the accompanying legislation that justifies the modification then again can be electronically published and archived.

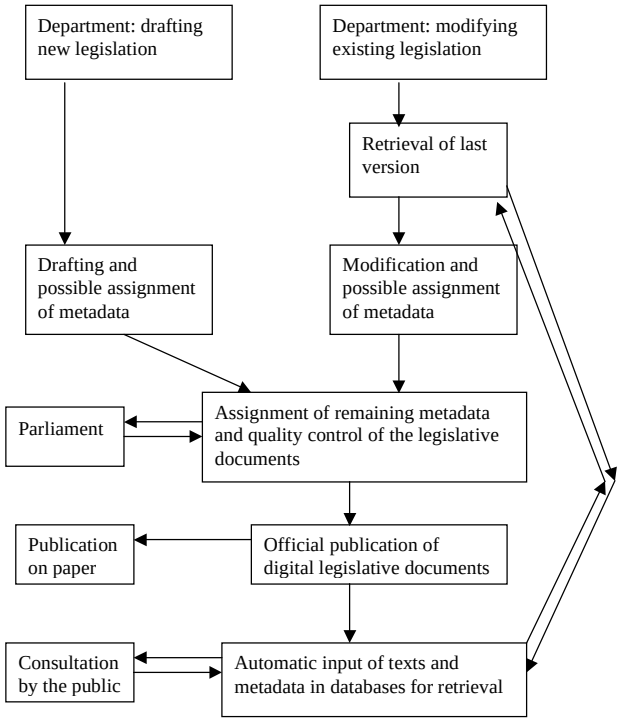

Figure 1: A schematic proposal for the lifecycle of legislation

Such a model is already operational in Tasmania, a federal state of Australia. The EnAct system which is developed by an Australian university, the Royal Melbourne Institute of Technology (RMIT), produces and manages an electronic warehouse of legislation in order to achieve an efficient and effective archiving of historical versions of legislation and retrieval of legislative texts as they are valid at a certain point in time (Arnold-Moore, 1997; Wilkinson et al., 1998). EnAct is a system which, besides the management of a database of consolidated legislation also administers the drafting and production of the legislative texts, workflows of legislative documents, production of camera-ready documents and their publication.[4]

The Tasmanian example is referred to because it shows the general direction in which the drafting, publication and retrieval of legislation should evolve. The example demonstrates that besides the database management and retrieval functions, automated support must also include drafting, exchange and publication of legislative documents. The Tasmanian model is however limited. In civil law countries like Belgium we are confronted with a much larger body of legislation, many modifications (versions) of legislation, a multilingual situation and different application areas.

3.2 Drafting of Legislation

Traditionally the drafting of legal documents has received a lot of attention. Legal documents – especially legislation - are often well structured and are partly composed of routine-like formula's (Moens, Uyttendaele & Dumortier, 1999). This leads to the idea that the task of drafting can be supported or even undertaken by what is known as an automated document assembly or document generation system. Such a system asks its users

[4] See also http://www.thelaw.tas.gov.au

questions, the responses to which insert or delete templates or part of templates, which have previously been set up by legal specialists. The templates provide the structure for inputting the texts and might also provide fixed portions of text together with precise instructions as to when given extracts should be used. Another approach limits the straightjacket into which the author of a document is forced, but supplies checking functions that after the creation of the texts control the structure and language, and suggest the necessary improvements. The checking functions then often incorporate an automated structuring and mark-up of the texts for document exchange (see below). Of course, both approaches can be combined.

The "document grammars" used by the drafting systems might evolve over time and different countries or regions might prefer own document templates requiring flexibility of the implementation and update of their grammar knowledge possibly by means of machine learning techniques (see below).

Despite of almost ten years of research into automated support for legislative drafting and attempts of implementing practical systems, we see that the most successful systems are part of a larger workflow in which the electronic documents are managed, and offer simple interfaces and consistency check functions for drafting the texts (Keymis, De Busser & Moens, 2001). Examples are the EnAct system of Australia (Wilkinson et al., 1998) and the Lexedit (Mercatali, 2000) system in use in Italy. With the exception of the EnAct system, little attention is given to the existence and management of historical versions when automatically supporting the drafting. Historical versions, however, are at the heart of management of legislation. In addition, research into document structures and properties of digital legislation – as well as of other legal documents – has received up until now little attention.

3.3 Document Exchange Formats

Drafting could add extra content to the documents by using mark-up languages such as SGML (*Standard Generalized Markup Language*) or XML (*Extensible Markup Language*). XML defines a subset of SGML and allows for customisation of mark-up languages with application-specific tags. The syntactic conformity of the structure can be validated with a DTD (Document Type Definition). SGML or XML marked documents are in ASCII format and can thus be interpreted on virtually all platforms. Therefore, both formats are extremely suitable for storing the textual content as well as the reference or metadata of a particular document and at the same time they ensure maximal exchange possibilities.

There is no question about the usefulness of mark-up languages. SGML has already been used to mark the structure of legislative texts (e.g., Arnold-Moore, 1997; Poulin, Huard, & Lavoie, 1997). There is a current interest to use XML for this purpose. The essential reference data of legislative texts (i.e., the data regarding identification and structure and data needed for version management) can be encoded as metadata in a mark-up language. However, the mark-up languages require that standards are agreed on and are effectively used by the different institutions in order to make interoperability and data exchange possible. For legislation, this means that there have to be agreements on allowable structures, allowable reference data, and their names. Standardization of the properties of documents is unfortunately often perceived as a restriction of intellectual creativity and a limitation of the possibilities of expression. Experiences with digital libraries learned that for standards to be useful they must be relatively simple, well understood and well accepted by all parties involved (Fox & Sornil, 1999). We believe that making agreements on the validity of certain structures and naming schemata is only realistic for the broad structure of legislation and for essential metadata regarding version management.

3.4 Publication and Preservation of Legislation

Legislation in digital electronic form is preferably officially published and consequently automatically stored in databases. While the complete legislation in all domains could be easily covered, users of the databases could be certain of up-to-date and correct legislative texts. There are however a few points that deserve attention. Digital information is fragile in ways that differ from traditional technologies, such as paper or microfilm.

The security and the guarantee of the authenticity of digital legislation are important aspects. Users of legislative databases must be confident of information authenticity, i.e., the information comes from purported sources and is not fraudulently altered. Checking and certifying data integrity is associated with technical processes such as integrity checking, certification, digital watermarking, steganography, and user and authentication protocols.

Another problem to be tackled is the preservation of the data and metadata. It is aimed at preserving and delivering the authentic legislative digital documents to subsequent generations of users. Preserving authentic records entails safeguarding the documents themselves and also their connections to the activities in which they were used: their content, their structure and their context must be preserved.

A long conservation also entails being concerned about changes in technology (McCray & Gallagher, 2001). Technological obsolescence is seen as the greatest threat to digital collections. There is also the lack of durable media: the hardware and software needed for storage, retrieval and communication of the information is subject to rapid changes. In this respect, standards such as SGML and XML that operate regardless the systems' hardware and software are important. The content can remain unchanged but compatible with the advances in database technology, browsers and access technology.

4 Automated Indexing and Hypertext Linking of the Legislative Texts

From the above it is clear that digital legislation needs many reference data or metadata for its efficient processing. There is an increasing interest in automatic generation of metadata, since their manual creation is considered to be the major impediment to digital archiving (Hodge, 2000). Considering the growing amount of legislation and its modifications, the assignment of metadata should as much as possible be automated. Current artificial intelligence techniques developed in the domain of information retrieval and natural language processing are very promising (Moens, 2001). From the viewpoint of legislation, techniques that automatically structure document texts and assign the necessary mark-ups (indexing of structured information) and techniques that automatically assign descriptors and search terms to texts (indexing of unstructured information), techniques that link citations (linking of structured information) and techniques that link texts based upon their content (linking of unstructured information) are valuable.

When legislative texts are stored in databases, they are broken into components (texts of specific articles, texts of titles of chapters, etc.). The structuring into components and assignment of the corresponding markups are best done when drafting the documents. The tasks can be performed automatically based upon knowledge of the structured patterns of the texts. When annotated examples are provided (i.e., example texts of legislation in which the components have been correctly marked manually), the patterns can be automatically acquired by learning from the examples, and subsequently be applied upon other legislative texts. This is a flexible approach that can be used for many types of legislative texts in different countries.

A second paradigm is the automated assignment of descriptors and classification codes to the legislative texts. The manual assignment of descriptors is traditionally applied in limited domains and has proven its usefulness for searching legislation. Automation of the technique can broaden the domains and drastically increase the number of texts processed. Techniques for text classification are actively researched.

Another valuable task is automated hypertext linking. Legislation can be linked to other legislation or to case law when these documents are cited in the texts. If drafters of legislation would cite all legislation in a uniform way, there is no problem of automated linking. Otherwise, an intelligent drafting support tool can translate the citation into a uniform identification code needed for the linking.

Linking and identifying similar content to the one at hand is a task that would be very valuable for the management of legislative database. Now too often, a new article is created while already legislation exists on the topic, making consequent expensive modifications and co-ordinations necessary. Finding text similar in content is a classical information retrieval task, but considering the progress in natural language processing it currently gains an important potential.

5 Conclusions

Digital legislation stored in databases is currently a very important source of information for the legal professional and, when accessible via the World Wide Web, also for the ordinary citizen. The correctness of the retrieved texts especially of the modified historical versions is of crucial value. The Agora-Lex project has learned us that current legislative databases suffer important problems mainly caused by their manual maintenance.

Legislation in a digital format requires the interaction of many tasks including drafting, modifications in subsequent versions, quality control (especially of the formal characteristics), the construction of correct reference data or metadata which are essential for naming the objects and for the management of historical versions, document exchange, publication, and archiving in and retrieval from databases. The official electronic publication of the texts (including the modified versions) and of their essential metadata could guaranty a correct and fast input in the databases. Artificial intelligence techniques can assist in the automatic assignment of some of the metadata. Only such a strategy allows us to cope with the growing bodies of complex legislation.

References

Arnold-Moore, T. (1997). Automatic generation of amendment legislation. In *Proceedings of the Sixth International Conference on AI and Law* (pp. 56-61). New York: ACM

Fox E.A. & Sornil, O., (1999). Digital libraries. In R. Baeza-Yates & B. Ribeiro-Neto (Eds.), *Modern Information Retrieval* (pp. 415—432). Harlow, UK: Addisson Wesley.

Hodge, G.M. (2000). Best practices for digital archiving. An information life cycle approach. *D-Lib* Magazine. (http://www.dlib.org/dlib/january00/01hodge.html).

Keymis D., De Busser R., & Moens, M.-F. (2001), *Draftingsystemen voor het opstellen van wetgeving: een evaluatie*. Technical Report TR-*E*-LEX-1 ICRI, K.U.Leuven.

Logghe M., Van de Kerckhove K., & Moens, M.-F. (2000). Automatic version management of legislation: The Agora-Lex project. In *Proceedings of 11th International Workshop on Database and Expert System Applications* (pp. 1051-1055). Los Alamitos, CA: IEEE Computer Society.

McCray, A.T., & Gallagher, M.E. (2001). Principles for digital library development. *Communications of the ACM*, 44 (5), 49-54.

Mercatali, P. (2000). Computer-aided methods and tools for legislative drafting. In S. Debaene & B. Van Buggenhout (Eds.), *Informatietechnologie en de kwaliteit van wetgeving* (pp. 139-156). Antwerpen: Intersentia.

Moens, M.-F. (2001). Innovative techniques for legal text retrieval. *Artificial Intelligence and Law*, 9, 29-57.

Moens, M.-F., Logghe, M., & Dumortier, J. (2002). Legislative databases: Current problems and possible solutions. *International Journal of Law and Information Technology* (in press).

Moens, M.-F., Uyttendaele, C., & Dumortier, J. (1999). Information extraction from legal texts: the potential of discourse analysis *International Journal of Human-Computer Studies* 51, 1155-1171.

Poulin, D., Huard, G., & Lavoie A. (1997), The other formalization of law: SGML modelling and tagging. In *Proceedings of the Sixth International Conference on Artificial Intelligence and Law* (pp. 82-88.). New York: ACM.

Van de Kerckhove, K., Logghe, M., & Moens, M.-F. (2000). Management of consolidated legislation in Belgium. In J. Dumortier, M. Taeymans & F. Robben (Eds.). *A Decade of Research @ the Crossroads of Law and ICT* (pp. 329-342). Gent: Larcier.

Wilkinson, R., Arnold-Moore, T., Fuller, M., Sacks-Davis, R., Thom, J., & Zobel, J. (1998). *Document Computing: Technologies for Managing Electronic Document Collections* (*The Kluwer International Series on Information Retrieval*). Boston: Kluwer Academic Publishers.