Managing Large, Structured, and Annotated Documents: A Study of Three Operational Cases in the Field of Environmental Legislation

Michel Treins, Carine Louvion, and Jacques Vaudelin

INERIS, French National Institute for Industrial environment and Risks.
Verneuil-en-halatte, France
Michel.treins@ineris.fr, carine.louvion@ineris.fr,
jacques.vaudelin@ineris.fr

Abstract. Managing legal documents, in the specific context of European environmental legislation, raise specific problems like internationalization and version management of the contents and metadata, and the need to perform tasks as consolidation, annotation, and description of the contents, at the scale of elementary fragment (article or chapter), instead of the whole document. Current standards as METS, or more specialized formats like HL7 / CDA, are not well adapted to answer these specific problems. In this paper, we present a new data model and an innovative structure of document, based on the "object" concept of descriptor. This development is now fully operational, and serves three important knowledge bases totalizing more than 11 millions of requests during the past year.

Keywords: Document management, life-cycle management, METS, CDA, HL7, environment, legislation, container, descriptor.

1 Introduction

Legal documents (laws, decrees, circulars…) have specific characteristics which notably impact on their management and their storage within computerized information systems.

The French national institute for industrial environment and risks (INERIS) plays a leading role in the assessment and prevention of technological and environmental risks in France and in Europe. One of the activities of the institute is the analysis and the periodic review of the legislation of the domain.

Consolidated and annotated legal documents, collected in bases of knowledge, are made available on Internet to a large number of simultaneous users. In consequence, the constraints of keeping the availability, the consistency, the integrity of the documents and all their components, managing their complete life cycle, and their storage in relational databases, have to be handled.

This paper presents the "document container" and the data models we developed to face these constraints. Our challenge was to make a simple, scalable, "object oriented"

R. Meersman, P. Herrero, and T. Dillon (Eds.): OTM 2009 Workshops, LNCS 5872, pp. 39–48, 2009.

model, interoperable with others documentary standards, and easily implementable in usual relational databases management systems.

We took back some of the innovations brought by the HL7 / CDA [2][3] data model, in the medical domain, and by the “Metadata Encoding and Transmission Standard” (METS) [4], by simplifying them, and by bringing an important headway: the concept of "descriptor".

In the sixth section, we present briefly the results of the implementation in three operational knowledge bases, published on Internet, equipped with powerful research functions, totalizing millions of requests during the past year.

In conclusion, we stress future evolution of the model, notably the capability of XML serialization and exportation, complying with documentary standards as METS or DocBook[5].

2 Characteristics of Legal Documents

A legal document (for example, a law) is a complete and inseparable entity: a persistent content established by an act of publication, fixed to a material support, bounded temporarily and spatially [6], tied to a specific context, having a robust logical structure [1], and referenced as a unique object.

Usually, legal documents are constituted by a large and nested hierarchy of fragments, henceforth called “sections”: Titles, Chapters, Articles, and Paragraphs… This generic structure can sometimes present local variations. However, every section within a single document can deal with varied subjects and concepts, sometimes with no semantic link between them. In many cases, during years, the successive amendments of the text will entail the abrogation of several articles and their replacement by new versions. In consequence, the version of the whole document may be not equal to the version of its own sections. In addition, an amendment to a legal text *is necessarily a legal text,* too! These documents, *in their different versions*, are tied together by a semantic link, and create a “solidarity of documents” [8], within which the reader, especially the jurist, may browse hypertextually.

Thus, because of these specificities, usual tasks as version management, consolidation, annotation, creation of relationships between parts (within the same document, or between different documents), and description of the contents (metadata, keywords…) must be done on the scale of the section, and not (only) on the whole document. All contributors, roles and organizations involved in the stewardship of the document and/or sections may be identified and authenticated.

To face these constraints, it was important to use a robust container, having the ability:

- To embed all the different “expressions” of the content (e.g. internationalization), and how to read, play, or consult them.
- To describe the structure of the document and how the different fragments are organized and linked together.
- To describe the concepts pertained to the content (“what we are talking about?”) and in which fragment they are contained (“where do we talk about it?”)

- To describe the semantic links between fragments or between a fragment and other objects: another document, or an external URI…
- To record information about life-cycle management, and administrative metadata.

We focused our study on the data models of two documentary standards: HL7 / Clinical Document Architecture (CDA) and Metadata Encoding and Transmission Standard (METS).

3 HL7 / RIM and « Clinical Document Architecture »

The HL7 (Health Level 7) initiative plays a predominant role in the standardization and normalization of healthcare information systems. Among its standards, HL7 proposes the CDA (Clinical Document Architecture) which proposes the structure and the semantics of clinical documents for the purpose of exchange between health care actors. The second release of the CDA specifications became an American National Standards Institute (ANSI) approved standard in May 2005. CDA is based on a formal "Reference Information Model" (RIM).

A CDA document has a header and a body. The header identifies and qualifies the document, and provides information on various participants, such as author, authenticator, encounter participants, informants, information recipients, and so on…

The body contains the clinical report and can be either an unstructured blob, or can be composed by a nested hierarchy of sections, each of them containing several attributes, one "narrative" block, and a variable number of "entries", which represent the coded semantics of medical statements: encounters, acts, observations, procedures, substance administrations, supplies, etc.

A CDA document may include texts, pictures and all kind of multimedia contents. It can refer to external documents, procedures, observations, and acts.

However, this structure presents several inconveniences which prevent its usage for our particular need:

- A data model specifically oriented towards healthcare information management. This model is too specialized to be easily applicable to other domains.
- Only one "narrative" (human readable) block per section: no possibility to have simultaneously the same text in several languages.
- A "one level", "flat" (non-recursive) concept of "Entries", whose abstraction need to be improved, as we propose it in this document with our concept of "descriptor".
- A lack of structural description of the document. A CDA document is not described by an explicit "structural map". The logical structure is implicitly contained in the nested serialized hierarchy of XML nodes. In consequence, life cycle management of the document is done on the scale of the whole structure, and not on the scale of its elementary components.

4 Metadata Encoding and Transmission Standard (METS)

Metadata Encoding and Transmission Standard is a XML schema developed on the initiative of the Digital Library Federation (DLF), providing an encoding format for descriptive, administrative and structural metadata for textual and image-based electronic documents. METS is currently maintained by the US Library of Congress. A METS document describes a numeric objet, and is structured in seven sections, which may contain one or several sets of metadata [7]. Both fist ones are mandatory:

- FileSection: list of numeric files constituting the object.
- StructuralMap: presents the physical and/or logical structure of the object.

The five others are optional and repeatable:

- Header: metadata describing the METS document itself.
- Descriptive Metadata: embedded or external descriptive metadata of the object. Multiple instances of both external and internal descriptive metadata may exist.
- Administrative Metadata: information about authors, contributors, intellectual property rights, date of publication, revision, and so on. Administrative metadata may be encoded internally or external to the METS document.
- Structural links: hyperlinks between elements of the structural map.
- Behavior: association of a part of the content with executable code.

METS allows a fine description of the physical and logical structure of the document, and of all of its contents. Each section of metadata is identified by a unique identifier, which can be used in the structural map to link a particular fragment of the document to a particular section of descriptive or administrative metadata.

However, the fragments of document hierarchy, and the metadata pertained to them, are not considered as real objects, which may be specialized, aggregated, composed, especially in a recursive manner. This fact makes difficult certain operations on the document when the contents have multiple expressions, as well as the descriptions associated with these contents.

We may take the example of a European legal directive. Although translated in 27 different languages, it remains the same document. All metadata (keywords, title, summary, date of publication, subject…) apply indifferently to all versions of the document. Furthermore, textual metadata may be expressed in several languages or by a formula (mathematics, chemistry…). They may be described recursively by others metadata (metadata of metadata…).

Such a description is possible with METS, but remains rather heavy to implement.

5 A New Model of Document Container

Because of these specific constraints, we decided to develop a new model of document container which would combine the advantages of the two formats. From

HL7 – CDA, we kept the concept of a document's body constituted by a nested hierarchy of sections. From METS, we kept the idea of the structural map.

In our model, a document is composed by a header, a structural map, and a variable number of sections.

Header and sections have their own small set of encoded attributes, necessary for managing the component's identification, the confidentiality level, the context, and the lifecycle management (new versions, obsolescence…). Mechanisms of inheritance are implanted, so that contextual values can propagate from high level of hierarchy to the nested components.

Descriptors are class of attributes which apply to documents, to sections, or to descriptors themselves, by a relation of aggregation. The application's field of descriptors also extends to thesauri. Descriptors can be repeated, specialized, aggregated as needed.

A descriptor is much more than a simple metadata. Descriptors contain the data, or the references to external data, whatever they are (texts, images, sounds, or any multimedia content, expressed in various encoding formats), AND the metadata associated to these data. For advanced management functions, metadata, which can be notably structured and complex, may be themselves described by others descriptors, without limit in the depth of this recursive schema.

There are several types of descriptors:

- The "narrative" descriptors, which contain the textual information of sections. These texts can be expressed in different languages, and according to different characters' codes and styles.
- "Metadata" and "Keywords" descriptors, which describe metadata used to qualify a document or a section: value of metadata, formalism (e.g. Dublin Core), reference to a specific taxonomy, etc.
- "Annotations" descriptors, which contain notes on the document or the section, and the description of the semantic associations[8] which can exist between documents, sections, and/or external resources. Annotations are defined [9] as "particular notes attached to a target by an anchor". Thus the target can be a section, a document, or another descriptor. Annotations may contribute to the content of the component itself [10][11] (for example, a comment or an explanation), or may rise the attention of the reader on a particular fact or information within the document [11]. Annotations are characterized by a "verb of action", which expresses the semantic of the link established between the note and the target. The "note", which is the heart of an annotation, is nothing less than a document [12]: as such, it can be constituted of nested sections, and may be described by a set of descriptors…
- "Anchor" descriptors. This specific tagging allows identifying exactly the portion of text on which is attached the annotation.
- "External reference", which is a link towards an external element, described with sufficient level of formalism to permit semantic sharing.
- "RenderMultimedia" (another "loan" of HL7 / CDA) which reference external multimedia content which is integral to the document, and must be rendered with the document. RenderMultimedia can reference a single ObservationMedia or one or more RegionOfInterests…

– "Structural Map": The nested hierarchy of sections is a kind of description of the document. For this reason, the structural map is a descriptor, too.

Headers, sections, and descriptors (and their successive versions) are all "embedded" in the structure, and are not included in a separate notice.

6 Information Model

Here is an extract of the diagram of class.

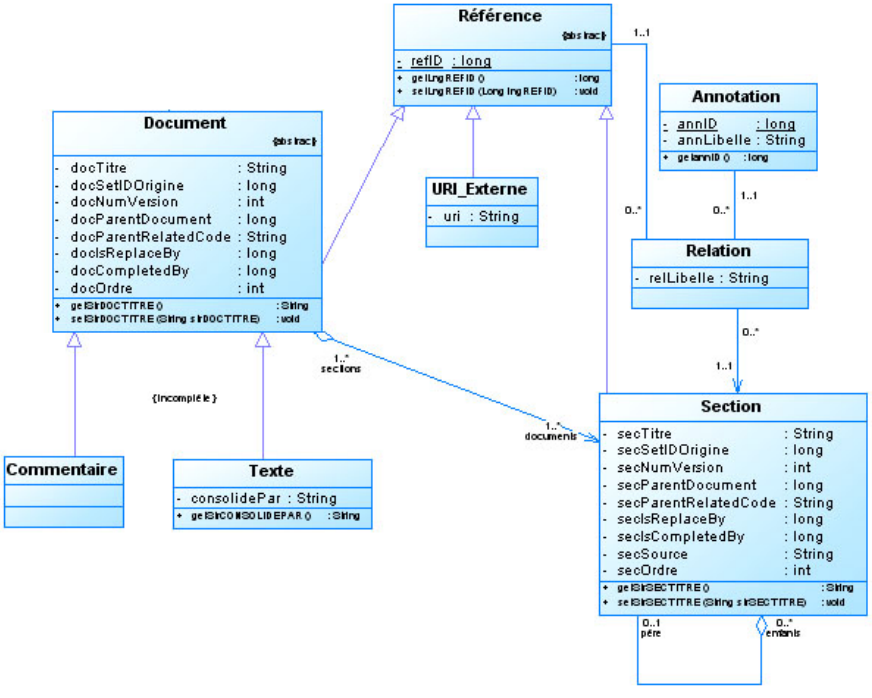

Fig. 1. Static view of the class diagram

6.1 Document / Section Relation

The collection of texts to be set up contains essentially legal texts. The study of this repository put in evidence the logical organization of texts. A lot of them consist of a succession of titles, ordered in chapters divided themselves into articles.

Texts are structured according to a hierarchical set of sections. A document consists of one or several sections. We made the choice to represent this structure with entities "Document" and "Section". The class "Document" contains only one attribute ("Title") and the specific metadata needed for the management of the different versions of the document. "Document" is a virtual class and need to be

specialized in several categories. “Text” represents the class of the corpus that we study. The class” Comment” represents a specific type of annotation.

The sight presented here is incomplete because the virtual class “Document” can be specialized in various types such as “Images”, “Videos” …

The representation of the internal structure of the document is symbolized by the class “Section”. This one contains a recursive relation which allows the management of the hierarchy of sections between them. The class Section is only constituted of information of versions.

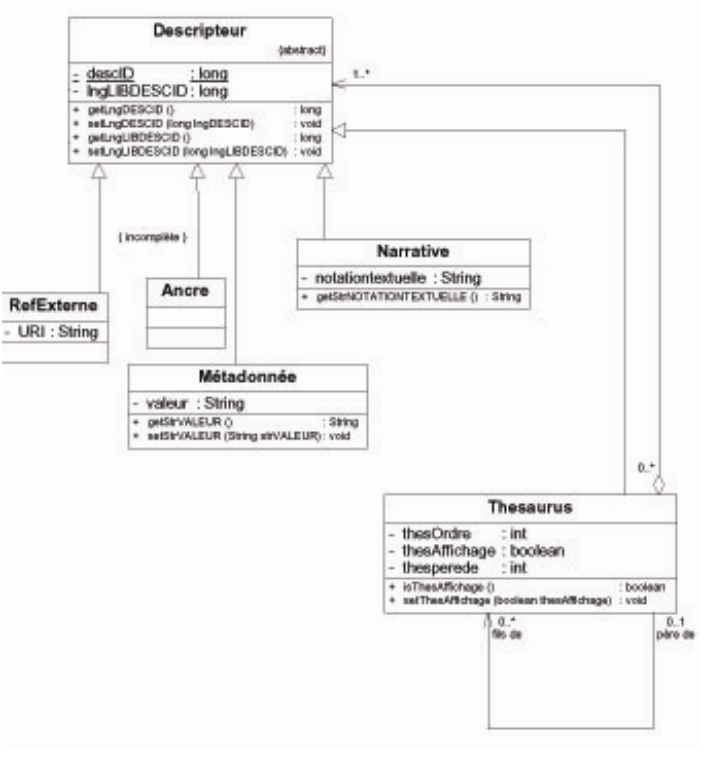

Fig. 2. Part of the class diagram "descriptor"

6.2 Document – Section / Descriptor Relation

The originality of this model is the class “Descriptor”. This class offers a very big flexibility, by allowing the description of all properties of any component (a document, a section, or another descriptor). There is virtually no limit for the different types of descriptors.

A component, at *instance level* and not only at *a class level*, may be described by the exact number of instances of various classes of descriptors. If the descriptor exist, it is always relevant.

The “Section” does not have any textual attribute. The narrative part is described by the eponymic specialization of the descriptor. Therefore, the management of the internationalization of the section is taken into account. In our application, the text can be presented in French or in English.

Metadata may be coded in various formats, as Dublin Core, ISO 19115 [13], EXIF [14] (for the images)…

Different instances of the “Metadata” class have been used: title, date of publication, editor (in our application – see below - the “official bulletin”), author, temporal extension…

Document and section are also described by a set of keywords, coded and referenced in a thesaurus. This thesaurus is composed itself by a set of “narrative” descriptors, answering the same need of internationalization than the content of sections.

6.3 Annotation Relation

The “Annotation” class qualifies, at semantic level, the associations between two references (for example, a “Text”, which represents the target, and a “Comment”, which is the note). The relation is characterized by a “verb of action” (“Is explained by”, “Is referenced by”, “Is overruled by”…). The target is a document or a section. The note is a document, a section, or an external URI.

7 Technical Implementation

An important issue was the ability of the model to be implemented in a standard, operational, traditional web application, accessible by a very large number of simultaneous users over Internet. Too many systems, so clever and so innovative on the paper, are incapable to cross the threshold of the model, even the prototype, while our purpose was to develop several knowledge bases on environmental legislation, serving millions of requests a year.

A Relational DataBase Management System (PostGre-SQL) was chosen for the technical implementation. Indeed, object-oriented databases and natively XML databases are attractive, but they encounter performances limitations and stability problems and remain quartered in niche markets. At the opposite, relational databases offer many advantages, especially in terms of integrity, safety, security, ability to treat large volume of information, handling complex and imbricated requests, and support of the major features of the standard SQL 2004, and ACID transactions (Atomicity, Consistency, Isolation, and Durability).

However, flattening an object-oriented model in an relational model which ignores inheritance raised a lot of problems. We had to create tables for every specialization of our virtual classes. This method offered the best flexibility without using too many empty attributes.

Based on this architecture, an operational application was developed, in the early 2008, as a java “REST” Web-Service within the X86 / Linux - virtualized infrastructure of the institute. This component serves three Web client applications,

developed in Java (the first one), and in PHP (the two others), in the field of environmental legislation:

- REACH-INFO (http://www.ineris.fr/reach-info). REACH is the Regulation for Registration, Evaluation, Authorization and Restriction of Chemicals. It entered into force on 1st June 2007 to streamline and improve the former legislative framework on chemicals of the European Union (EU). The text of REACH is complex; it concerns various categories of industries. National Helpdesk is an information department on REACH, whose mission is to guide companies on the text of REACH, helping them to conform to their obligations.
- AIDA (http://www.ineris.fr/aida). AIDA supplies a selection of European legal texts (regulations, directives, decisions, recommendations, notices), published in the official bulletins, and relative to facilities presenting technological risks.
- RGIE (http://www.ineris.fr/rgie). This site gives information on legislation relative to extractive industries: mines, careers…

The knowledge base server is also indexed and used by a well-known commercial search engine, which builds its index by taking advantage of the specificities of the model (descriptors, metadata, annotations…), and provide more relevant results to user searches.

During year 2008, more than 1.3 millions distinct sessions were counted for these applications, totalizing almost 11 millions of documents requests, and 1.3 Terabytes exchanged, without any problem, and with good performances, sustainability, and fault tolerance.

8 Conclusion

We can ask ourselves on the interest of developing a new model of document – one more – while various standards already exist in this domain. Nevertheless, our model has no vocation to describe a new concept of representation of information, neither a new format of metadata, nor a new type of electronic book. Our model simply implements a “numeric envelope”, an electronic container having the ability to record and exchange various contents, in all their different expressions and versions.

In this particular field, the normative initiative is not so dynamic, nor so advanced. Furthermore, the proposed formats, like METS or HL7/CDA, are often expressed in XML, even when natively XML database management systems have not yet the maturity of their relational counterparts. Usually, the actual standards are dumb on the problems of internal implementation, leaving these points with the discretion of developers or software editors.

For these reasons, we were brought to develop a robust system, based on a very simple “object-oriented” model, capable of a real industrialization.

Recently, a fourth application was added to the three first ones, proving the scalability, the extensibility, and the potential of this architecture.

However, some issues remain to be handled in the next months:

- Ability to export to and import from XML documentary standards, as METS (container for transfer and exchange), or DocBook (for contents themselves).
- Ability of express the "anchor descriptor" in RDF/A, wrapped into the HTML code which usually compose the "narrative" part of the text.

References

1. Pedauque, R.: Document: Form, Sign and Medium, as reformulated for electronic documents. In: Working paper. STIC - CNRS 2003 (2003)
2. CDA, HL7 Clinical Document Architecture - Release 2.0, Committee Ballot. ANSI Standard, Health Level Seven, Ann Arbor, MI, USA (2005)
3. Dolin, R.H., Alschuler, L., Boyer, S., Beebe, C., Behlen, F.M., Biron, P.V., Shabo, A.: HL7 Clinical Document Architecture, release 2. Journal of the American Medical Informatics Association 13(1) (January,Febuary 2006)
4. Cantara, L.: METS: the encoding and transfer Standard. Cataloging & Classification Quarterly 40(3-4), 237–253 (2005)
5. DocBook V5.06b, working draft – june (2008) http://www.oasis-open.org/docbook/specs.
6. Bachimont, B.: Audiovisuel et Numérique. In: Calderan, L., Hidoine, B., Milet, J. (eds.) Métadonnées: mutations et perspectives. ADBS editions, pp. 195–222 (2008)
7. MetaData Encoding and Transfer Standard: Primer and Reference Manuel. Version 1.6 (September 2007), http://www.loc.gov/standards/Mets
8. Treins, M., Curé, O., Salzano, G.: Gestion des annotations dans le dossier médical informatisé. Analyse des apports des normes et standards et propositions pour la conception de solutions. In: Salembier, P., Zacklad, M. (eds.) Annotations dans les documents pour l'action. Hermes Publishing, Londres-Paris (2006)
9. Bringay, S., Barry, C., Charlet, J.: The annotation, a new type of document in electronic health record. In: DOCAM (2004)
10. Lortal, G., Lewkowicz, M., Todirascu-Courtier, A.: Annotation: Textual Media for Cooperation. In: Proceedings of Annotation for Cooperation Workshop, pp. 41–50, November 24-25 (2005)
11. Zacklad, M.: Vers le Web Socio Sémantique: introduction aux ontologies sémiotiques. In: Deuxième journée de la plate-forme de l'AFIA (2005)
12. Treins, M., Curé, O., Salzano, G.: On the interest of using HL7 CDA release 2 for the exchange of annotated medical documents. In: CBMS (2006)
13. ISO/TC 211 19115:2003 Geographic Information – MetaData, http://www.iso.org
14. Exchangeable Image File Format, a standard of Japan Electronics and Information Technology Industries Association (JEITA), http://www.exif.org