Knowledge acquisition from texts towards an ontology of French law

Guiraude Lame

Centre de Recherche en Informatique, Ecole Nationale Supérieure des Mines de Paris,
35 rue Saint Honoré, 77 305 Fontainebleau France
lame@cri.ensmp.fr
phone : 33 (0)1 64 69 48 50
fax : 33(0)1 64 69 48 47

Abstract. Designing ontologies is a key aspect of knowledge management and knowledge representation.

We introduce a document based ontology building methodology for french legal documents. Unlike others methods, main knowledge will be here directly extracted from texts in two ways : (1) a first knowledge acquisition from texts that extracts relevant terms of the domain and their relations and (2) a second acquisition that extract implicit knowledge from the documents. Our documents are the legal texts published every day in the Journal Officiel de la République française, the french official publication for legal texts.

Our goal, designing an ontology from those texts, is to enable conceptual retrieval of the documents and to formalize the conceptual framework of an information system based on those documents. This ontology has clearly documentary purposes and is dedicated to normative texts.

1 Introduction

The context of this work is the Internet site `http://droit.org` which publishes the Journal Officiel (J.O.) every day. The J.O. gathers all the normative texts of the french Republic : laws, decrees, decisions of various administrations ... When making request to this site, users mainly expect to obtain all the documents related to their request, including those that don't necessarily contain the exact words of the request but which correspond to the same idea, to the same legal concept. Moreover lawyers may appreciate being able to retrieve all documents related to a given juridical point, i.e. all the legal texts regulating a specific topic.

In order to solve these problems, we must consider the knowledge embedded in the documents, the legal concepts herein expressed. An ontology, seen here as a formal description of the domain concepts and relations between them [Gruber], seems then to be the adequate tool to reach that goal. The ontological level has been suggested [Guarino] as the appropriate level for describing the general meaning of a given domain. It is the level at which concepts can be considered. We consider, see also [Valente], ontologies both as epistemological commitments and as a conceptual model of the domain. Therefore, our ontology of law can be considered as a framework for a knowledge retrieval system based on our documents [Valente and Breuker 1]. It is a domain ontology [Valente and Breuker 2] in the sense that it is designed on the basis of this particular field of law that is represented by our legal documents (the normative field).

The method we suggest in this paper is based on the analysis of this corpus which will provide the legal knowledge embedded in those documents [Aussenac-Gilles], [Bourigault]. This approach is partly similar to the goals and techniques proposed by the TIA group ("Terminologie Intelligence Artificielle")[1] merging ideas stemming from terminology analysis, knowledge engineering and artificial intelligence. In our point of view, two kinds of knowledge useful for designing an ontology can be extracted from our documents. First : relevant terms of the domain stored in a terminological knowledge base (TKB) and second : the knowledge implicitely expressed by our documents represented in our three models.

In this paper we start (section 2) by presenting the first step of our method which is a terminological analysis of the corpus leading to a terminological knowledge base. Thereafter (section 3) we introduce three models built upon our documents. Then we suggest a normalization stage (section 4). And finally (section 5) , we discuss one of the main points of our work, the updating of the ontology.

2 First knowledge acquisition : the terminological knowledge base

Creation of a terminological knowledge base (TKB) is the first step of the elaboration of our ontology. This TKB gathers terms of our domain and the lexical relations between them. It can thus be seen as a thesaurus. In order to design

[1] http://biomath.jussieu.fr/TIA

it, we need to extract the domain terms and then identify the lexical relations that may hold between those terms. A human expertise will be needed to identify which terms and relations are relevant to the domain and to our goal of designing an ontology. To make such a process trackable (the total number of terms being very huge), our job here is to suggest a subset of relevant terms and relations to the experts.

2.1 Extraction of the domain terms, relevancy suggestions

Terms, embodied in nouns or noun phrases, are extracted from our documents by specific tools (Lexter, Sylex). Those tools integrate natural language processing techniques to identify noun phrases in documents.

Sylex extracts 2170243 noun phrases from the 52545 documents used in our experiment (the total of our base at this moment). We developped a suite of techniques to identify potentially relevant terms among this list. For instance, we identified noun phrases that cannot be useful in our goal : those matching standard date patterns[2], those matching cross references between documents ... We also detect noun phrases that certainly have a legal sense since they match predefined legal terms. Using such techniques, we refined our initial list to a sublist of 301957 noun phrases. In a second step, we gather the noun phrases that have the same prefix and perform a plural insensitive morphological grouping of the expansions. We thus obtain, for each prefix known as having at least one legal meaning a list of noun phrases related to this word. From this list our system lists those that may be relevant and those that may be irrelevant to the expert for approval (see below apart of the results on the word abandon-surrender).

NP with abandon as prefix	suggestions
abandon partiel de la production surrender of the production	BON (production) OK (production)
abandon du navire mentionn surrender of the mentionned boat	? ?
abandon de domicile surrender of domicile	BON (domicile) OK (domicile)
abandon des emballages usagés surrender of the packing materials	? ?
abandon du projet du site de voujeaucourt surrender of the Vougeaucourt project	PAS BON BAD
abandon de créance surrender of debts	BON (créance) OK (debt)

2.2 Identifications of the lexical relations between terms

The second step of the process through a TKB is to identify the lexical links that may hold among the selected terms.

[2] one might argue that dates should be preserved, but our goal in this section is to make explicit the future primitives of the domain

Among the terms selected above, we first link terms to all the noun phrases having this term as prefix. This lexical link may semantically be qualified as hyperonymy.

A contextual analysis of the terms is a more accurated way to explicit lexical relations between terms. In such an analysis, the sentence is the relevant unit of study and, via the computation of the cooccurrences of terms, links between them are created. A contextual analysis is generally well admitted when the goal is to build knowledge based or documentary systems [Pietrosanti].

Another way to identify potential lexical links existing among terms is to analyse what [Morin] called lexical syntactic schemas in sentences. Those schemas are based on noun phrases ; they represent pertinent semantic parts of sentences. The semantic analysis of sentences derives from this analysis. The general construction of those schemas are as follows : NP axis-term NP. Two noun phrases are related with an axis-term (usually a verb or a verb group). For example, one could identify in the titles of our documents lexical syntactic schemas such as :

- Norm1 portant modification de Norm2 (Norm1 modifying Norm2)
- Norm1 relatif à Norm2 (Norm1 related to Norm2) ...

These schemas will make explicit the functions of our documents, the motivations that spurred their creation.

A specific kind of schemas is the one devoted to the definition of notions [Rebeyrolle]. In our documents this will correspond to relations between noun phrases (NP) such as :

- NP1 est fixé à NP2 (NP1 is fixed to NP2)
- NP1 comprend NP2 et NP3 (NP1 include NP2 and NP3)
- NP1 résulte de NP2 (NP1 is a result of NP2) ...

The identification of such schemas could allow us to suggest identification relations between terms to the expert of the domain.

We are currently working on improving these identifications of lexical relations between terms. After human validation, we will obtain a hopefully very relevant TKB. This TKB will be the first stage of the process that will lead us to a more formalized model of the domain, i.e. the ontology.

In our search for a formalization of the domain, we considered that we couldn't avoid the juridical aspects of our documents.

3 Second knowledge acquisition : three models of the domain

We now present a second level of knowledge acquisition from texts : the explicitation of the knowledge implicitly embedded in our documents. This knowledge acquisition step tends to three models based on our documents.

The first ontological visions of law separated norms from facts and considered them as separated entities. This viewpoint was pregnant until mid 1990's and has been a preliminary of most of the works on artificial intelligence and law. A different point of view has been taken by [Valente] in 1995. He built upon works of Kelsen, a law theorist, to identify different categories of legal knowledge.

Since each model of the domain is deeply influenced by its own goal [Visser] and our goal is documentary while Valente's was a legal decision system, we decided not to use his approach. Our analysis of the document set led us to introduce three models of the domain that will be integrated in our final model. They have been built considering the specificities of our corpus and always having in mind the documentary finality of our work.

3.1 The hierarchical model

Our documents are normative texts and this particularity has to be considered. Normative texts have indeed hierarchical relations through what is called the normative pyramid.

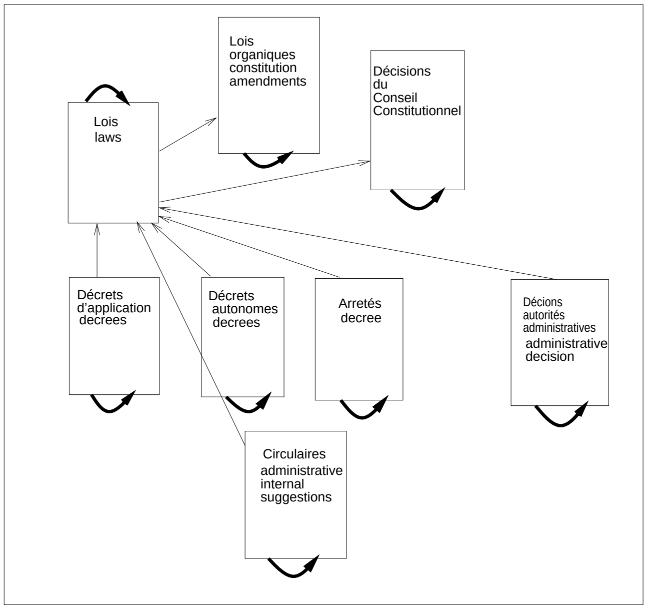

Fig. 1. The hierarchical model

Figure 1 is a diagram of the hierarchical organisation of our documents. Our set of documents include different type of norms. To summerize a norm N1 is above a norm N2 when N2 has to comply with N1. For example, a text (usually a decree) the fonction of which is to detail, to make explicit, another text (a law most of the time) has to comply with it in the sense that it can't change the main purpose of the law.

The hierarchical relations existing between our documents will introduce hierarchical links between the legal concepts expressed in those documents. Those links will be useful in the future for the elaboration of our ontology.

3.2 The structural model

Our documents presentation abides to a specific formalism. This formalism is represented below in Figure 2.

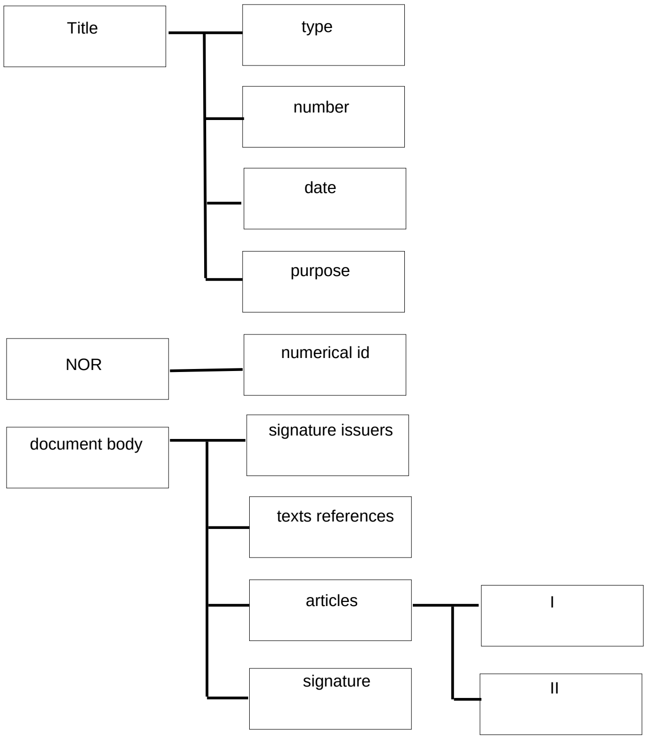

Fig. 2. The structural model

This segmentation constitutes a structure of the document and furthermore enables the identification of the informative parts of the document. Articles are, thus, informative since they contain the legal concepts. But references to other texts related can allow us to find the hierarchical position of the document considered, to include it in the hierarchical model.

The structural approach on our documents allows us to consider the possibility to use the XML standart. We can envision anchors in the document to enable database view of the documents. For instance :

$$< \text{ title } >, < \text{ type } >$$

3.3 The functional model

A functional viewpoint of legal texts is largely admited to be relevant in researches on artificial intelligence and law. [Valente and Breuker 2] indeed suggests a functional ontology of law and makes a distinction between different categories of knowledge, function of their roles. [Pietrosanti] also made a functional analysis of legal documents.

But here again, it can be argued that each model is deeply rooted in its own goals [Guarino]. Our goal is documentary among normative documents. So we consider the function of the norms itselves and therefore the relations between the legal concepts expressed herein. Consequentely the primitives of our final model could take into account this aspect.

While identifying lexical syntactic schemas (section 2.2), we discovered different functions potentially existing between norms. We can distinguish norms that roles of which are to regulate the normative field itself and norms the function of which is "external", ie related to the conditions of existence of the legal entities (artefacts).

Norms regulating the normative fields itselves are assumed to have *organisational functions*. They can be identified through lexical syntactic schemas such as :

- Norm1 abrogeant Norm2 (Norm1 abrogationg Norm2)
- Norm1 modifiant Norm2 (Norm1 modifying Norm2)

Norms determining condition of existence of legal entities have what we call an *organic function*. They create the legal entities (such as a civil servant, a commission ...), specify their conditions of existence and may fix their legal death.

Figure 3 below describes the roles of these norms on the real world and on the normative level.

Those three models will help us structure our final model of the domain : the ontology.

4 The ontology : conceptualizing the TKB using our three models

Using the terms and relations of the TKB and the modelization of our set of documents, we have to link both to achieve our goal : building on ontology. One of the key aspects will be to select the concepts of the domain that will be the future primitives of our ontology.

Conceptualization is the process that, from a TKB, leads to a final model where concepts are structured and relations among them identified. This stage also called *normalization* (Bachimont) exhibits, on the base of the terms lexically linked in the TKB, semantically linked domain concepts.

We don't know how we will implement this stage of conceptualization. We have no doubt that the contextual analysis of the terms of our documents and

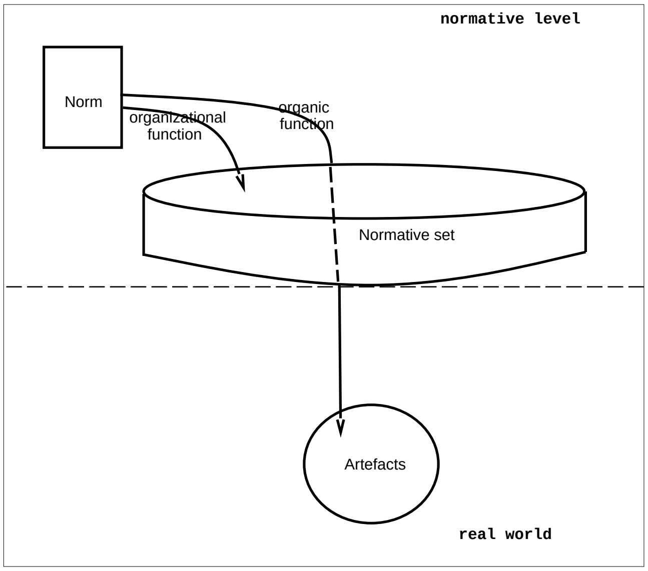

Fig. 3. The functional model

the identification of the roles played by the norms will help us in our task. All three models previously sketched will support the explanation of the domain and documentary task. The hierarchical model can be seen as a potential framework. The functional model describing functional relations among norms may allow us to define functional relation between concepts expressed in those documents. Finally the structural model may suggest an organization of the segments of our documents, and provide clues to refine queries for specific fields of documents.

At the end of this formalization stage of our work, we hope to obtain the final model of our documents, ie the ontology.

5 Updating the model

Update management is important in our domain and even more for the normative documents. Every day, new concepts are created in laws, decrees ... Since our model defines the concepts of our documents, frequent updates have to be considered to maintain its relevance and usefulness.

Updates may be divided in two stages : first, an integration of a new concept in the terminologal knowledge base and second, an eventual integration of this concept in the model. In fact, every new legal concepts expressed in our documents has to be present in a thesaurus (the TKB) to enable documents that contain it retrieval. But each new legal concept may not be relevant in the ontology, as a primitive of the model. Consequentely, each new legal concept could be integrated in the TKB, and, if the concept is not sufficiently relevant to the goal of our model and in the field of normative law, it may not be fully integrated in the model but possibly as a descriptor of a pre-existing concept.

Naturaly, this decision of a double-level integration of a new concept has to be taken by an expert. The characteristics of the law field will help him in his task. Law, indeed, can be identified as a relatively static discipline. Legal text writers and, if they don't fulfill the task, legal commentators (judges, law teachers), always try, when creating a new legal concept, to categorize it. And creation of a new category of concepts is unfrequent.

By this legal categorization of new concepts, legal text writers guide ontology designers in their task to choose the appropriate location of concepts in the model.

6 Conclusion

We described method for building an ontology of the normative documents of french law for documentary purposes. This method is a two-tiered bottom-up approach for ontology building : the first step leads us from the documents to a TKB (a semi-formal model) of the domain, the second formalizes in three models the knowledge implicitly embedded in the documents. The ontology is then built by conceptualizing the TKB, using our three models.

References

[Aussenac-Gilles] Aussenac-Gilles, N., Biébow, B., Szulman, S. : Modélisation du domaine par une méthode fondée sur l'analyse de corpus. Actes de la conférence IC'2000 91-104

[Bourigault] Bourigault, D., Charlet, J. : Textes et ontologies. Actes de la conférence IC'2000

[Breuker] Breuker, J., Muntjewerff, A., Bredeweg, B. : Ontological modeling for designing educational systems. Dept. of Computer science & Law, University of Amsterdam, June 17, 1999

[Guarino] Guarino, N. : The ontological level. Proceedings of the 16th Wittgenstein Symposium, Kirchberg, Austrian, August 1993

[Gruber] Gruber, T. : What is an ontology ? A translation approach to portable ontologies. Knowledge acquisition, 5(2), 1993 199-220

[Hwang] Hwang, C. H. : Incompletely and imprecisely speaking : using dynamic ontologies for representing and retrieving information. InfoSleuth, Microelectronics and computer technology corp. (MCC), Austin, Texas, USA, June 1999

[Morin] Morin, E. : Extraction de liens sémantiques entre termes à partir de corpus de textes techniques. PhD IRIN, Nantes, 1999

[Pietrosanti] Pietrosanti, E., Graziadio, B. : Advanced techniques for legal document processing and retrieval. Artificial intelligence and law, vol. 7, Kluwer Academic Publishers, 1999 341-361

[Rebeyrolle] Rebeyrolle, J. : Utilisation de contextes définitoires pour l'acquisition de connaissances à partir de textes. Actes de la conférence IC'2000 105-114

[Valente] Valente, A. : Legal knowledge engineering : a modeling approach. University of amsterdam, The Netherlands, IOS press, 1995

[Valente and Breuker 1] Valente, A., Breuker, J. : Towards principled core ontologies. Proceedings KAW'96, Banff, Canada, November 1996

[Valente and Breuker 2] Valente, A., Breuker, J. : A functional ontology of law. Artificial intelligence and law, vol. 7, Kluwer Academic Publishers, 1999 341-361

[Visser] Visser, P. R. S., Bench-Capon, T. J. M. : A comparison of four ontologies for the design of legal knowledge systems. Artificial intelligence and law, vol. 6, Kluwer Academic Publishers, 1998 27-57