Designing medical law ontology from technical texts and core ontology.

Sylvie DESPRES – Bernard DELFORGE
CRIP5
Université René Descartes
45, rue des Saints-Pères
75006 PARIS
email : {sd, delforge} @math-info.univ-paris5.fr

Introduction

This study is part of a pluridisciplinary project which has just started and which will run for three years. The practical aim of this project is to build a portal specialized in medical law. Lawyers, physicians and computer scientists are involved in this project.

This portal will be composed of two parts: texts on line and a search and retrieval engine linked to medical law ontology.

1) The texts on line are:

- texts written by the specialists in their field on particular themes in medical law and which are stored in the server. These texts are called "*studies*". A study includes an abstract, a bibliography and a glossary (legal, technical).
- original texts (legal code, case law) located on different specialized servers on the Web.

2) The search and retrieval engine helps the end-user to search for original texts on the Web. This search is based on medical law terminology and ontology. The engine operates as a meta-engine covering existing Web search engines such as specialized official Web sites (for instance, Legifrance http://www.legifrance.gouv.fr/).

The future end-users of this site are lawyers, health professionals (physicians, nurses, etc.) or non-specialists.

- Lawyers may wish either to update their knowledge of the latest legal texts (legal code, case law), or to get state-of-the-art information in their field. The studies serve this latter purpose. These end-users master the legal vernacular language but not the medical one.
- Physicians are confronted in their practice with cases for which they can be held responsible. They log into the site in order to know what their responsibility is as practicing doctors (medical terminology, work organization, ethics, etc.). These end-users do not master the legal vocabulary.
- The non-specialists are also confronted with concrete cases. They log into the site to find solutions to their problems. Generally speaking, these users master neither the medical nor the legal language. They use a common-sense language.

This paper focuses on the way in which the use of a core ontology in the domain of medical law may facilitate the use of a search engine on the Web. But although ontology exists in law, there is none in medical law. Therefore, it was appropriate to build such an ontology. This ontology was designed from three different modalities: a corpus composed of heterogeneous texts about medical law that are analyzed both with software and manually, expert interviews and reuse of existing ontology in law.

This paper deals with the problems of representing and structuring knowledge in ontology and also with the formalization and re-usability of ontology. The discussion focuses on the knowledge obtained from the specialized texts, and on the complementarity of analyzing texts, interviewing domain experts and reusing existing ontology. This paper is structured as follows.

In section 1, the main problems associated with document retrieval are recalled. In section 2, the textual context of the study is presented, and an analysis of the knowledge obtained from specialized texts and the software used for this study are provided. In section 3, the development of a sub-ontology of law ontology is addressed and, in particular, the links between the initial core ontology and medical law ontology are discussed. In section 4, the role of the lawyers in the acquisition process is described. The discussion and conclusions constitute section 5.

A help tool in medical law request formulation on the Web

As simply put by Winkels [1998], the main problems with conventional text or document retrieval systems are related to:

1. the interaction with the system (typically users have to enter complex queries that combine keywords through Boolean and proximity operators);
2. the quantity and quality of the search result (only a few relevant documents are found);
3. the presentation of the output of the system (typically a list of (ranked) relevant documents).

Some researchers tackled the first problem by improving the query language of databases, e.g. using (restricted) natural language instead of Boolean expressions [Croft and *al*., 1992]. Van der Pole [1996] proposed a device that can handle the composition of Boolean queries, and CLIME [Winkels, 1998] addressed this problem by providing its users with a structured query formulation interface. In our project, the aid in reformulation is provided through a query formulation interface which uses the terminology and ontology of the field, as in OntoSeek [Guarino, 1999].

The second problem concerns legal information servers which are based on traditional retrieval techniques. Relevant documents can not be found because the keywords the user enters are not mentioned in the text, or because deciding they are relevant requires inferences and legal reasoning. The adding by legal experts of special keywords to the documents or the use of thesauri helps somewhat, but even then recall and precision will remain sub-optimal because not all uses of documents can be foreseen. Artificial Intelligence (AI) techniques have been tried to provide conceptual retrieval which aims to index and retrieve documents according to their content (meaning) rather than by the occurrence of keywords. The technique has been illustrated, most notably by Hafner [1981 in Winkels 1998], Dick [1991], Gelbart and *al*. [1991] and Mariani and *al*. [1992]. A slightly different approach is described by Rissland and *al*. [1996]. They use heuristic search mechanisms with different evaluation functions to retrieve cases from a database. Even conceptual retrieval will not solve all problems in finding relevant documents for legal problems. Legal databases are by nature very incomplete because law is incomplete. Lots of interpretations and inferences are left to the common-sense reasoning of (legal) professionals. Simply adding all missing information is not feasible in some particular domains. Moreover, providing an answer to the question requires a long chain of inferences and legal reasonings. As in CLIME Winkels [2000], the idea is to add knowledge about the domain and inferencing capabilities to help the end-users in the reformulation.

Finally, the third problem deals with the presentation of search results. Typically, information retrieval systems present a list of retrieved documents. The order in which these are presented can be based on the implementation of the system itself (e.g. first found presented first), document time (e.g. most recent cases first), or relevance. In this last, most interesting, case, the idea is that the most relevant documents are presented first. This problem is not tackled here. The relevance will be determined by an evaluation function which is based on the terminology and the ontology of the field.

In fact, the precise aim of this study is to design a specialized legal ontology to help formulate (or reformulate Van der Pole [1996], Desclès and *al*. [1999]) a request in the field of medical law. Reformulation is broken down into two principal phases: a) terminological aid for the end-user to help him decide on the right concept, b) reformulation of the initial request based on the knowledge included in the ontology in interaction with the end-user (chart 1).

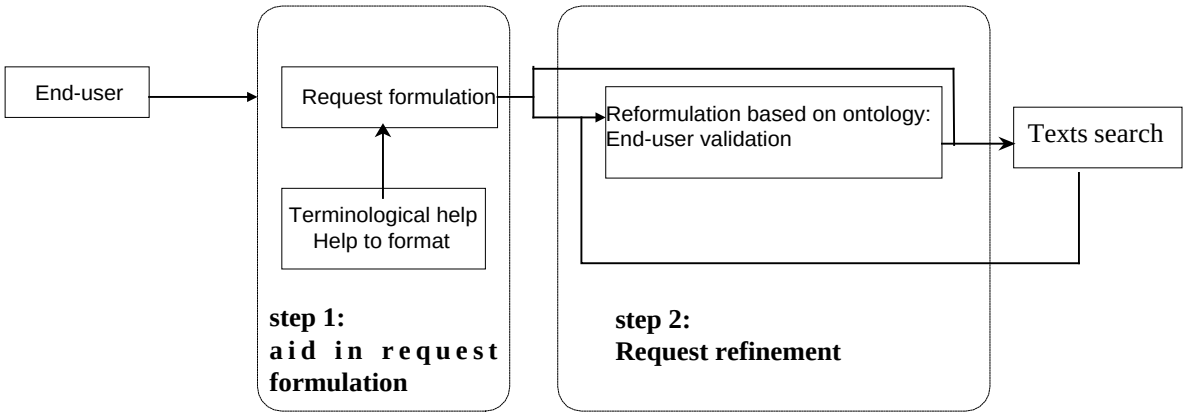

Chart1

Textual context and texts analysis: a cognitivo-discursive analysis

The medical law ontology is built partly from a corpus of texts. To obtain the knowledge from the text a cognitivo-discursive analysis is performed with TROPES software [Ghiglione and *al*., 1998].

Different software products can extract key-words, segments of texts, and different verbal elements. LEXTER [Bourigault, 1994] and NOMINO[Plante and *al*., 1997] are terminology extraction software. They perform a morpho-syntactic analysis of the corpus and give networks of noun phrases which are likely to be terminological units. In this work, the TROPES software is used. TROPES performs a morpho-syntactic analysis in the same way as the previous software, but the specificity of TROPES is to extract the semantics of the texts i.e. to analyze their content. It seems that this specificity is important because of the nature of legal studies. In fact, the lawyers commit themselves in the topics dealt with in these texts. Nevertheless, at the present time, it is difficult to appreciate precisely the specific contribution of TROPES.

The most useful tools of TROPES to analyze these texts are: automatic classification of words; immediate detection of the context; filtering of the topics according to their relevance; graphs allowing the user to visualize each reference detected in its discursive context; the scenario semantic for interpretation.

The information contained in this corpus, as previously stated, has two origins:

- the codes (essentially civil code and health code) relative to medical law. The civil code is divided into "Livres" (books), "Titres" (titles), "Chapitres" (chapters), "Sections" (sections) and "Articles" (articles). The health code is divided into "Parties" (parts), Livres" (books), "Titres" (sections), "Chapitres" (chapters), "Sections" (sections), "Paragraphes" (paragraphs) and "Articles" (articles). The texts can be found on the Web in integral form or in part. It is possible to retrieve them from a title, a codification and keywords.
- the "studies" are structured by different themes. A study is composed of an abstract on the medical law subject, a glossary, a bibliography and the main judicial decisions. The abstract is indexed by a lawyer who is the text author. The indexing focuses on the terms that the specialists wish to appear in the glossary and in the law sections.

Only the "studies" are preprocessed by TROPES because they contain the specific vocabulary of the medical law field.

The work with TROPES and another software products on a summarized text (composed of around 1200 words) from a study (for instance "Individual's genetic profile" or "Safety precaution", etc.) can be broken down into four steps:

- Step 1: TROPES performed a semantic classification of the 1.200 words in three levels of different granularity (Universe 1, universe 2, used references), organized in a hierarchical structure. The first level, Universe 1, is composed of very general concepts. The second level, Universe 2, is composed of specialized concepts of Universe 1. These reference universes represent the context with respectively 200 and 1000 possible "semantic equivalent" classes. The advantage of this classification in the reference universes is the possibility to target the relevant words. The third level, "used references", performs the analogical grouping and may group 10.000 equivalent classes. For instance, the word "judge" will be associated with the class "lawyer", included in universe 2 "justice" and in universe 1 "law". It is possible to modify or to improve these classifications by using a scenario semantic. The status of the TROPES classification is more than a glossary and is not yet an ontology. In the following steps, the work necessary to reach this status is described.

- Step 2: among these 1.200 words many of them are not relevant. Therefore, a selective sorting carried out on the previous selection of words provides a new set of 400 more relevant words (including the conjugated terms). This sorting is performed by means of a program (not TROPES), and is based on the TROPES labeling. The words labeled "junctor" "modalisation" or "preposition" are automatically removed.

- Step 3: from this new set of words, pruning is performed manually. At the end of this step, for one text, 4 to 5 concepts are selected in Universe 1 (for instance law, family, agreement, people), approximately 10 concepts in Universe 2 (crime, law, justice, sentence, etc.), and around 50 terms in used references (court, family, heredity, etc.). A set of synonyms corresponding to each identified concept is designed. To do that, the lawyers' glossaries are used. At this stage, there is no exploitation of the verb classification.

- Step 4: The classification obtained from the previous steps is refined both by means of the knowledge contained in the reference books (e.g. generic legal works), and by recourse to the lawyers. The knowledge obtained from the reference legal books receives general approval. If necessary, the concepts are renamed or removed. An example of such a classification referring to the concept of Law is described in chart 2.

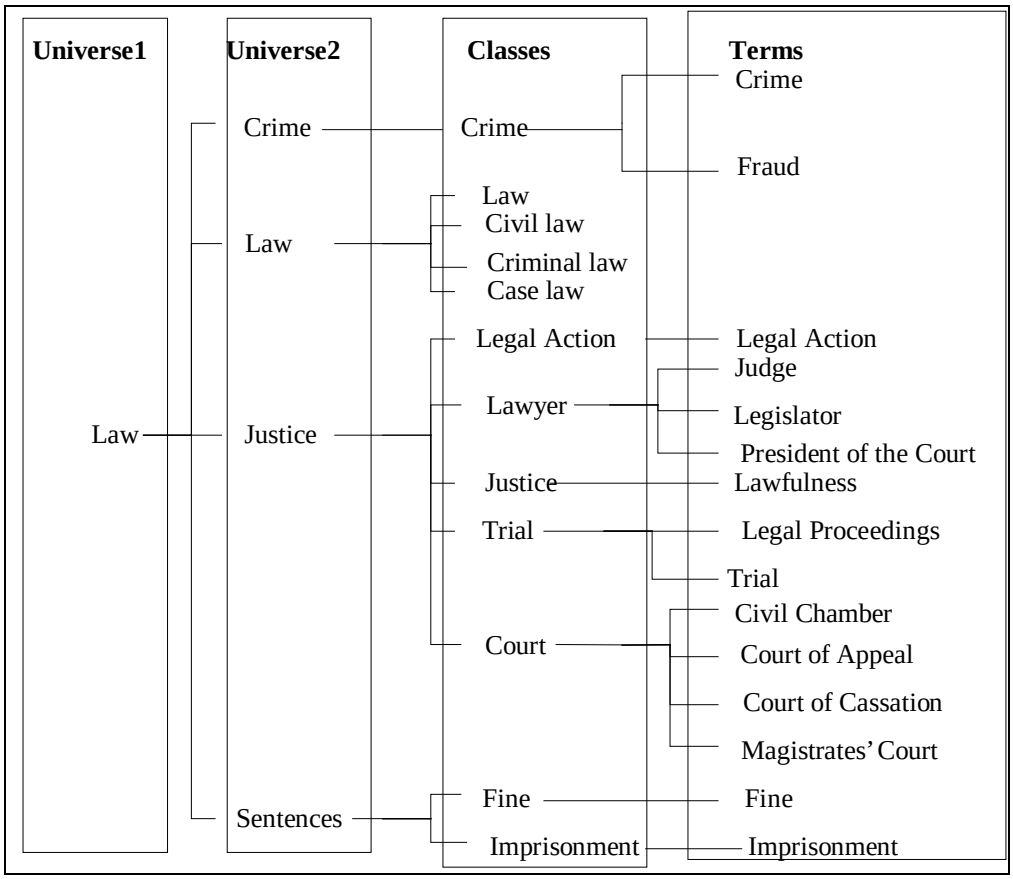

Chart 2

Moreover, some terms and relations between the concepts are refined with the lawyers. For instance, the term "filiation" was refined because it is an important term in the topic "individual's genetic profile and filiation ", but it was not explicitly described in the studies. For instance, the concept of filiation is described in chart 3.

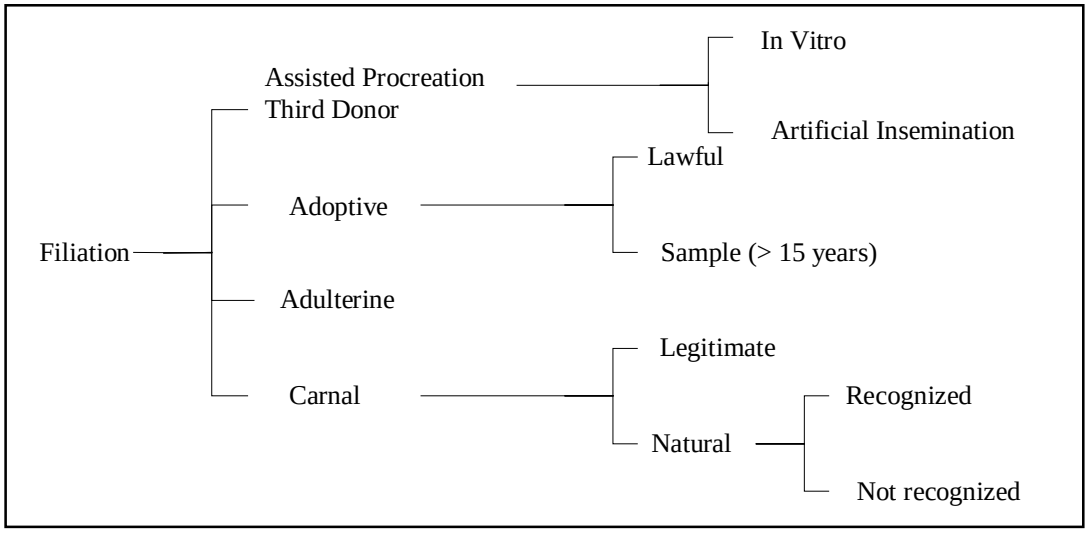

Chart 3

The lawyers also provided a refinement of concept law by adding the notion of legal branches of medical law that did not appear in the classification obtained from the text. The studies are linked to general topics and legal branches of medical law (cf. chart 4).

Legal branches	General topics	Particular studies
Bio-ethics	…	…
Responsibility	… civil responsibility …	… Safety precaution …
Medical practice	…	…

Chart 4

At this point of the study, the ontology obtained from this analysis of the text is not yet definitive. Only a preliminary version with some concepts is available.

Design of medical law ontology: the re-use of a legal core ontology

The use of a core ontology is justified by the need for reasoning in legal knowledge to retrieve searched information. For instance, if the request is "search +paternity +recognition", then in order to lead to "action to establish paternity", it is necessary to have a conceptual model of medical law. The choice of a functional core ontology of law is justified by the need for reasoning in legal knowledge. But the specificity of medical law necessitates the addition of an ontology of this domain. In fact, the codes (civil and health) contain the generally approved knowledge in the domains. Therefore, only this knowledge is globally shared. The text is the absolute reference for the reasoning about law. The difficulty posed by reasoning in law comes from lawyers' interpretations of the texts when they are confronted with a legal case.

A small number of explicit conceptualizations of the legal domain is available (for an overview on legal ontologies for system design, see Visser and Winkels [1997] and Hage and Verheij [1999]. As recalled by Visser and *al*. [1999], four well-known legal ontologies are: (a) McCarty's LLD, (b) Stamper's NORMA, (c) Valente's Functional Ontology of Law and (d) the Frame-based Ontology of Van Kralingen [1995]. This medical law ontology is built from texts written by specialists and from the core ontology established by Valente [1995].

This work is in line with Valente's work because legal knowledge is considered as composed of entities. Valente's ontology of law [1995] is based on a functional perspective of the legal system. The legal system is considered as an instrument to change or influence society in specific directions determined by social goals. Its main function is to react to social behavior. This main function can be broken down into six primitive functions, each corresponding to a category of primitive legal knowledge. Accordingly, Valente distinguishes six categories of legal knowledge Breuker and *al*. [1991], Valente [1995]: (a) normative knowledge, (b) world knowledge, (c) responsibility knowledge, (d) reactive knowledge, (e) meta-legal knowledge, and (f) creative knowledge.

At the beginning, this work was centered on world knowledge. World knowledge is legal knowledge that describes the world that is being regulated. It delineates the possible behavior of people, and institutions in society, and thereby it provides a framework to define which behavior ought (and ought not) to be adopted. It can be considered as an interface between the common-sense knowledge of people in society and normative knowledge. Within world knowledge, Valente distinguishes (b.1) definitional knowledge, and (b.2) causal knowledge. Definitional knowledge is the static part. It consists of definitions of (b.1) legal concepts (e.g., agents, objects), (b.2) legal relations (e.g., legal qualifications of actions), (b.3) a case (viz. the problem case under investigation), (b.4) circumstances (viz. the grounded facts or building blocks of a case), (b.5) generic cases (viz. typical generic legal cases), and (b.6) conditions (viz. the building blocks of the generic legal cases). Together these constructs provide vocabulary which can be used to describe the relevant aspects of the world from a specific view adopted by the legislator. Causal knowledge (b.2) is the dynamic part, describing the behavior of people in society in terms of definitional knowledge.

After having identified the concept of medical law, the model of the Legal Abstract Model was specialized. For instance, chart 6 sketches some elements involved in the action to establish paternity.

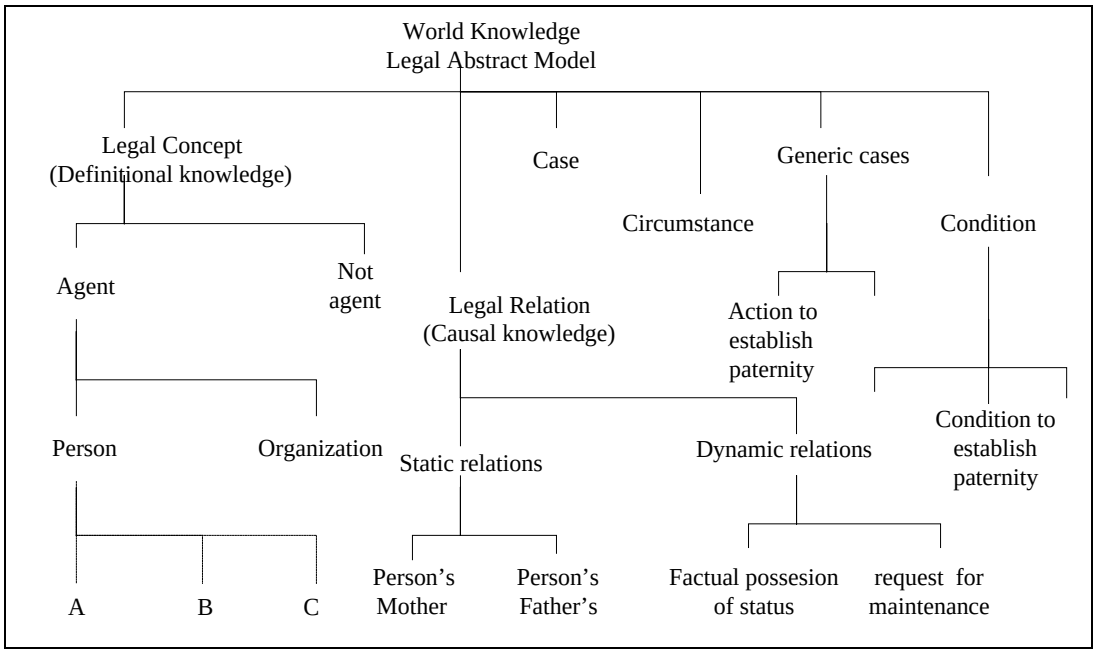

Chart 5

The next step is to build the models for normative and responsibility knowledge.

Knowledge obtained from specialists' interviews

The protocol used to obtain the knowledge from lawyers consisted of work sessions that tackled some subjects: description of the legal studies, discussion about medical law glossary and the current problems confronting medical law.

The "studies" correspond to the lawyers' know-how in medical law. In these "studies", the lawyers address major problems which are evolving quickly, and they have often to actualize the knowledge about these themes. Therefore, these studies produce knowledge that is shared by the most specialized lawyers. For that reason, they are a good corpus because they constitute the knowledge necessary for request formulation.

The lawyers' interviews focused on the character of the requests according to whether they were made by lawyers, health professionals or laymen. The medical law topics considered were: individual's genetic profile and filiation, safety precaution, organ donation. According to the end-user profile, the site access is motivated by different aims: upgrading knowledge for lawyers, inquiring about their responsibility as health professionals, legal recourse for the others. Therefore, through interviews it appeared necessary to organize knowledge in a way providing multiple accesses to the legal information.

Establishing the glossary was a very difficult task for the lawyers. The terms were located on two levels: legal domain, study context. It seems that the difficulties lay in the choice of terms that were judged necessary to be defined by lawyers. The necessity of a double glossary (one in the legal domain and the other in the medical domain) emerged from the discussion. For instance (chart 6), the term "therapeutic cloning" is defined by the physicians, while the term "link of filiation" is defined by the lawyers. In both cases, the definition should be clear for the future end-user. Therefore, the knowledge elicited from the lawyers during these interviews is shared and generalized, but this is restricted to each field. From this work with the lawyers, the necessity of designing a bridge between the two glossaries was confirmed. For instance, the basic lawyer may not know the term "therapeutic cloning" within the general problem of cloning. There may be legal terms in the Convention of the European Council that the physicians don't understand.

Term	Legal interpretation	Medical interpretation	Initial Study	Set of synonymous
Factual Possession of Status	In matters of filiation, factual possession of status is when a person is considered the child of another person …		individual's genetic profile and filiation	presumption of paternity
Link of filiation	Legal relation between a person and his/her parents	Genetic profile	individual's genetic profile and filiation	Heredity Paternity
Therapeutic cloning		…		

Chart 6

During the interviews, the lawyers specified the notion of case law in French Law, and they also discussed some interpretations of the concepts proposed in Valente's ontology. Therefore, a difficult point is to identify the specific elements of French law and to verify how the conceptual model of the core ontology deals with them.

Discussion

The ontology is constructed from a corpus composed of heterogeneous texts about medical law. This approach is opposed to another which is based on psycholinguistics used in WordNet [Miller and *al*., 1997]. It is similar to the approach of Assadi [1998] who worked with a corpus composed of technical documents about a domain. A difference resides in the nature of the corpus: it is not a set of technical texts but texts about specialties in which the lawyers commit themselves. These texts depend on the state of the art which evolves through time along with case law and the new laws, and regularly new concepts may emerge.

Knowledge acquisition from this corpus offers the following advantages:

1. the vocabulary to name the concept is that of the lawyers
2. the links between the concepts and their occurrences in the texts including various linguistic forms of their expression are kept.

One difficulty is related to the choice of the corpus. It seems that the studies constitute a good corpus because they have unity and a legitimacy in the domain of medical law.

Therefore, this study has shown the complementarity of analyzed texts, interviews domain experts and reuse existing ontology in knowledge acquisition from the texts. Three kinds of text were available: the reference legal books, the codes, the studies. Depending on the nature of the texts, globally shared and stabilized domain knowledge was elicited. This knowledge is relevant but remains incomplete. Recourse to other sources of knowledge for modeling the field is absolutely necessary.

The medical law ontology is not finished. Among the reasons that explain the incomplete character of medical law ontology, the first is that all the studies have not been finished, but the major reason is that the processing work on the available texts has not been completed. This can be partially explained by the fact that it is difficult to decide how to integrate the different parts of this ontology into a core ontology. For instance, what are the definitions of core ontology that are also in the ontology of medical law ? How is medical law ontology situated with regard to core ontology ? These questions lead us to study core ontology and to develop the link with it.

The difficulties encountered during the integration of medical law are great. In the field of law, especially for medical law, it is important to remember that new concepts appear regularly. It is the result of the creation of new norms. The design of the conceptual field model is based on the texts containing basic and specific field knowledge. Moreover, the use of individual know-how should simplify the updating of this model.

Bibliography

Assadi H. – Apprentissage et Acquisition de Connaissances – Construction d'ontologies à partir de textes techniques – application aux systèmes documentaires. Thèse de doctorat de l'Université Paris 6, octobre 1998.
Bourigault D. – LEXTER, un logiciel d'extraction de terminologie.– Application à l'acquisition de connaissances à partir de textes. Thèse de l'EHESS, Paris 1994.
Breuker J.A., Haan N. den – Separating world and regulation knowledge: where is the logic ? In M. Sergot, editor , Proceedings of the third international conference on AI and Law, pp.41-51, New York, NJ,1991 ACM.
Bourcier D., Hasset P., Roquilly C. – Droit et Intelligence Artificielle. Editions Romillat 2000.

Croft W., Turtle H. – Text retrieval and inference. In P. Jacobs, editor, Text-Based Intelligent Systems, pages 127-155. L. Erlbaum, Hillsdale, 1992.
Desclès J.P., Laublet P., Cavanié M., Djiuoa B., Jackiewicz, Naït-Baha, Lespinasse K. – Le projet RAP : Une méthode et un outil interactif pour la collecte et la sélection d'information sur le Web, Rapport technique du CAMS, 1999.
Dick J. – Reasoning with portions of precedents. In Proceedings of the Third International Conference on Intelligence Artificial and Law, ACM, New York (NY), pp. 244-252, 1991.
Gelbart D., Smith J.C. – Beyond Boolean Search: FLEXICON, A legal text-based intelligent system. In In Proceedings of the Third International Conference on Intelligence Artificial and Law, ACM, New York (NY), pp. 244-252, 1991.
Ghiglione R., Landré A., Bromberg M., Molette P. – L'analyse automatique des contenus. Dunod, 154p., 1998
Guarino N., Masolo C., Veter G. - OntoSeek: Content-Based Access to the Web. In IEEE Intelligent Systems, 1999.
Hage J. Verheij B. – The Law as a dynamic interconnected system of states of affairs: a legal top ontology. In International Journal Human-Computer Studies 51, pp.1043-1077, 1999.
Heijst G. van, Schreiber Th., Wielinga B.J. – Using Explicit Ontologies in KBS Development. International Journal of Human-Computer Studies, vol. 45, pp. 183-292, 1997.
Kralingen R.W. van – Frame-based conceptual models of statute law, Computer/Law Series, n°16, Kluwer Law International-The Hague, The Netherlands, 1995.
Mariani P., Tiscornia D., Turchi F. – The formalization of retrieval and advisory systems. In proceedings of JURIX'92, Information Technology and Law, Koniinklijke vermande, Lelystad, NL, pp. 71-79, 1992.
Plante P., Dumas L., Plante A– Nomino, 1997.
Rissland E.L., Skalak D.B., Friedman M.T. – BankXX: Supporting Legal Arguments through Heuristic Retrieval. AI and Law, 4 pp. 1-71, 1996.
Van der Pole R. W. – A device for query composition, Report CS96-02, Université de Maastricht, 1996.
Valente A. – Legal Knowledge Engineering. A Modelling Approach. PhD Thesis, University of Amsterdam. IOS Press Amsterdam, 1995.
Visser P., Winkels R. Eds. Proceedings of the first Workshop on Ontologies in Law. LEGONT-97. Melbourne Australia, 1997.
Visser P., Bench-Capon T. – A comparison of four ontologies for the design of the legal knowledge systems. Artificial Intelligence and law, vol. 6, pp. 27-57, Kluwer Academic Publishers, 1998.
Visser P., Bench-Capon T. – La création d'une bibliothèque ontologique pour les systèmes d'informations juridiques, pp. 28-45. Bourcier D., Hasset P., Roquilly C. – Droit et Intelligence Artificielle. Editions Romillat 2000.
Winkels – R. – CLIME : Intelligent Legal Information Serving. Put to the Test. In AIL98, 10p., 1998.