Ontologies for Legal Information Serving and Knowledge Management

Joost Breuker Abdullatif Elhag Emil Petkov Radboud Winkels
University of Amsterdam
Department of Computer Science and Law (LRI)
P.O. Box 1030
NL 1000 BA Amsterdam, the Netherlands

Abstract. In this paper we describe the nature and use of various ontologies for the information management of documents of criminal trial hearings. This work is part of the e-COURT European IST project, but it is also based on work in previous and current IST projects aimed at legal knowledge management. We describe these ontologies, in particular an 'upper' ontology –LRI-core– that has the role of providing anchors and interpretation to the various legal domain ontologies. The role of LRI-core is exemplified by an ontology about Dutch criminal law. In the second part of the paper we describe how these ontologies are to be used in tagging and annotating the hearing documents; in searching these documents, and in structuring the return set of retrieved documents. The technology used to represent these ontologies is based on emerging standards of the Semantic Web.

1 Introduction

In this paper we present an overview of the use and development of ontologies for legal domains in the e-COURT project. This overview is based upon experiences and results of various European projects on legal information serving and knowledge management in which we participate(d). [1]

The e-COURT project is a European project [2] that aims at developing an integrated system for the acquisition of audio/video depositions within courtrooms, the archiving of legal documents, information retrieval and synchronized audio/video/text consultation. The University of Amsterdam is responsible for the role of (legal) ontologies in the e-COURT system.

The focus of the project is to process, archive and retrieve legal documents of criminal courtroom sessions. In principle, these documents should be accessible via the web,and should be supported by the semantics-based web-services that will rely on the cascade of languages that will be the standards of the Semantic Web: XML, RDF and OWL. The e-COURT system under development has the following main functions:

- *Audio/Video/Text* synchronization of data from the court trials and hearings.
- *Advanced Information Retrieval.* Multilingual, tolerant to vagueness. Statistical techniques are combined with ontology based indexing and search. Queries are expanded by terms from various ontologies expressed in RDF/OWL(DAML+OIL).

[1] These projects are: CLIME (IST 25414, 98-01, see http://www.bmtech.co.uk/clime/index.html) about legal information serving, KDE (IST 28678, 99-01, see www.lri.jur.uva.nl/kde) about (legal) knowledge management, and E-POWER (IST 28125 see www.lri.jur.uva.nl/research/epower.html, and the MetaLex initiative (see www.metalex.nl).

[2] IST-2000-28199, www.intrasoft-intl.com/e-court

Joost Breuker, Abdullatif Elhag, Emil Petkov and Radboud Winkels, 'Ontologies for Legal Information Serving and Knowledge Management' in T.J.M. Bench-Capon, A. Daskalopulu and R.G.F. Winkels (eds.), *Legal Knowledge and Information Systems. Jurix 2002: The Fifteenth Annual Conference*. Amsterdam: IOS Press, 2002, pp. 73-82.

- *Database management:* multimedia documents supporting retrieval. Documents are annotated and tagged in XML, based upon the ontologies.
- *Workflow management* defines and manages rules for sharing relevant information and events among judicial actors.
- *Security management* plays an important role to protect privacy information and to comply with national and international regulations about the interchange of criminal information.

Our focus, and also the focus of the second part of this paper is on the 'advanced information retrieval' functions and their support by ontologies. We will first discuss in Section 2 the various types of ontologies required to cover the legal domain (criminal law). As law is highly entangled with common sense views on the nature of social events, roles and actions, we need also ontologies that cover the understanding of these concepts (Section 2.1). Besides these high level notions, we need the specific terms to describe the structures in some types of legal documents (transcripts of trial hearings; criminal codes). These are presented in Section 2.2. In the second part of this paper (Section 3) the use of these ontologies is described in the indexing (tagging, annotating, Section 3.2) and in expanding search queries for documents and in clustering result sets (Section 3.3).

2 Concepts of Law and Legal Documents

Different from medicine, engineering or psychology, law is not "ontologically" founded. For instance, in jurisprudence, but also in legal doctrines, the major questions concern the *justification* of law and legal systems, rather than concepts that cover legal reality. Legal reality is social reality. Justification –which is derived from the term *ius* (law)– is the domain of epistemology; the study of what we can know and believe. Epistemology is about reasoning, argument and evidence, while ontology is concerned with modelling (understanding) and explaining the world. Therefore, it is no surprise to see that 'core ontologies' about law are rather epistemic frameworks (see e.g. also [8], [2]). In particular, FOLaw, the Functional Ontology for Law, developed by [7] is to be viewed as a (CommonKADS) inference structure, despite what the authors claim it to be. FOLaw describes dependencies between the various types of knowledge (e.g. normative and responsibility knowledge) in such a way that they provide a generalized argument or inference structure of legal reasoning. Highly practical for constructing legal knowledge and information systems (see [9]), but it is not a real ontology. Ontologies are not about types of knowledge and reasoning roles, but about identifying concepts. When applied to annotating and 'semantic' tagging and retrieving information in hearings of criminal trials, these epistemic frameworks have little to say.

2.1 Why we Start With a (Legal) Core Ontology

Law is concerned with constraining and controlling social activities using documented norms. Legislation refers to social situations and activities that can be qualified as legal or not. It is the nature of these *social* situations and activities that is the object of ontological modeling of law. The law may provide other, more precise or more 'open texture' kinds of definitions of these entities, but essentially most are left to common sense. That means that for modeling and understanding some legal domain we should be able to include notions about agents, actions, processes, time, space, etc, i.e. some top or upper ontology appears to be indispensable, because the concepts of law are spread over almost the full range of common sense.

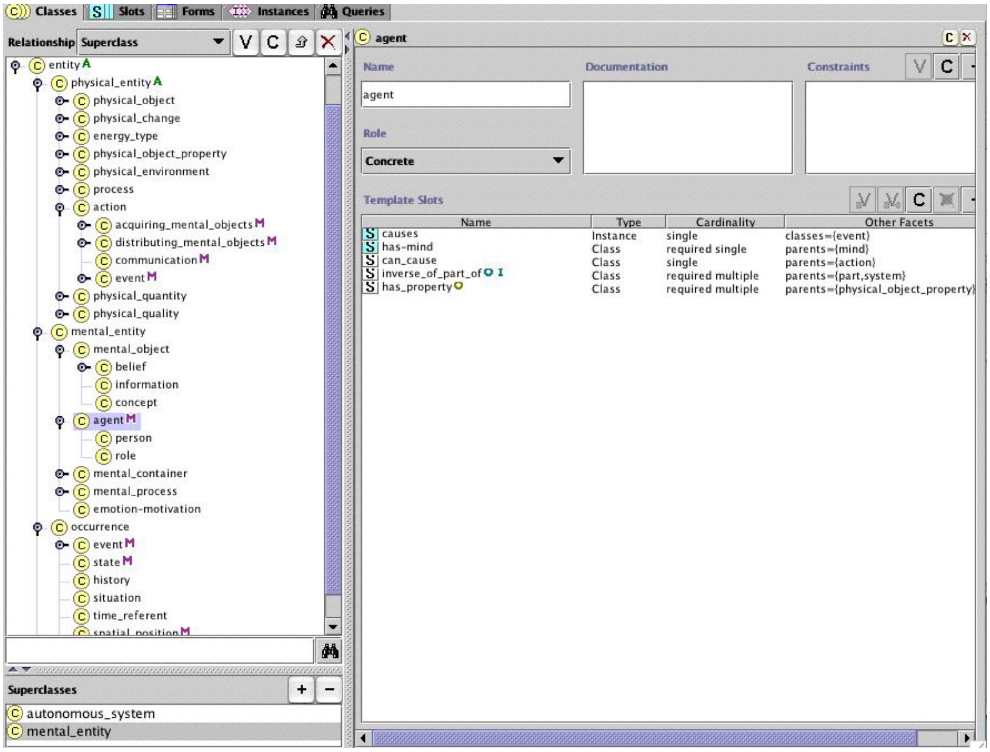

Figure 1: LRI-Core in Protege

We could not reuse currently available upper-ontologies (e.g. [5], the CYC upper ontology or in particular the IEEE-Standard Upper Ontology (SUO) that is under development (http://suo.ieee.org)) because their focus is rather on describing the physical and formal-mathematical world: not the social, communicative world which is more typical for law. Besides this lack of of sufficient covering, we did not agree about the physical parts of these ontologies anyway [1], [4]. What also makes these upper ontologies, in particular the recent SUMO-proposal[3] unsuitable for covering legal documents is that there is no notion of 'document' itself, or of any medium for communication.

The purpose of developing our "LRI-Core" is not to propose yet-another-upper-ontology, but to provide a broad, rather than 'deep' conceptual structure for the typical *legal*, or legally relevant, 'upper' notions.

The major principles applied in this core/upper ontology, called **LRI-core** are:

- Objects and processes are the primary entities of the physical world. In objects energy and matter are distributed, so that objects participate in processes, while processes transfer or transform energy. The participation of objects may change some quantity or quality (transformation) or may change its position (transfer (movement, emission, etc), or its existence.

- Mental entities behave largely analogous to physical objects. In fact, one may argue that the mental world consists largely of metaphores of the physical world. A typical mental object is 'concept', and mental processes affect mental objects. This reflects our folk psychology which assumes e.g. that if one is informed about some fact, this fact is stored in memory. Whether this fact is believed or not is an epistemological issue. Facts of belief and knowledge are mental objects consisting of concepts.

[3]SUMO stands for Standard Upper Merged Ontology, i.e. a merger of a number of published upper ontologies, which is considered by SUO a first draft.

- Communication proceeds via physical objects (documents, sounds) and actions (talk, reading) which represent mental objects (information).
- The mental and the physical world overlap in the concept of 'agent'. It is ambiguous because 'agent' is classified as both a physical object and a mental object.
- Social organization and -processes (e.g. communication) are composed of roles that are performed by agents that are identified as individual persons. The law associates norms to roles. For instance, the traffic regulation provides norms for traffic participants (or its subclasses, eg pedestrian or driver of a motor vehicle). However, when it comes to solving legal cases, the responsibility is with the individual who performed a role.
- Time and space have also an ambiguous status. Related to occurrences, they provide positions of events and situations. However, as physical entities they provide the qualities of extension (size, life-cycle) of objects and processes (field, duration).

This ontology, containing about 200 concepts, is still under development, but has definitions for most of the *anchors* that connect the major categories used in law (person, role, action, process, procedure, time, space, document, information, intention, etc.). This is the major purpose of this ontology. It should not only provide some framework to get a coherent view on a particular legal domain ontology, but it also allows inheritance of well-defined terms e.g. for verifying the domain ontology. **LRI-core** is written in DAML+OIL/RDF using Protégé. In Figure 1 the major structure is presented. Validating an upper ontology is not simple. The representation tools (DAML+OIL with FACT) enable consistency checking (verification) but the real validation is in actual uses. The main intended use is suppporting knowledge acquisition for (legal) domains, but a real test of its semantics should be whether it enables natural language understanding of common sense descriptions of simple events as e.g. in the description of events in legal case documentation. Of course, this is ultimately what the Semantic Web intiative of the W3C organisation is about. An ontology focusses on terminological understanding: for full natural language understanding and reasoning other kinds of knowledge have to be specified and implemented too.

2.2 Criminal Law

The e-COURT project is aimed at the semi-automated information management of documents produced during a criminal trial: in particular the transcriptions of hearings. The structure of this type of document is determined by the debate/dialogue nature of these hearings, but also by specific, local court procedures. Besides tagging its structure, it is also important to identify (annotate) content topics of a document. These vary from case descriptions (e.g. in oral testifying) to topics from criminal law (e.g. in the indictment). Therefore we are currently developing an ontology that covers Dutch criminal law, whose major structure we will discuss below. As the e-COURT solutions are aimed to work for most European countries, in principle we have to develop such an ontology for every jurisdiction that intends to use e-COURT. This Dutch ontology will be the framework for ontologies of Italian and of Polish criminal law.

Because it is not apparent where divergences between the concepts of criminal law between legal systems will be found, we want to ground, or 'anchor', the ontology of criminal law at a very abstract level in the LRI-Core. There will be little debate about the fact that criminal actions are physical or symbolic ones; that a verdict is a mental qualification represented by a document; that a person who is being accused will perform in court the role of defendant, and that persons as agents can perform both physical and mental activities etc.

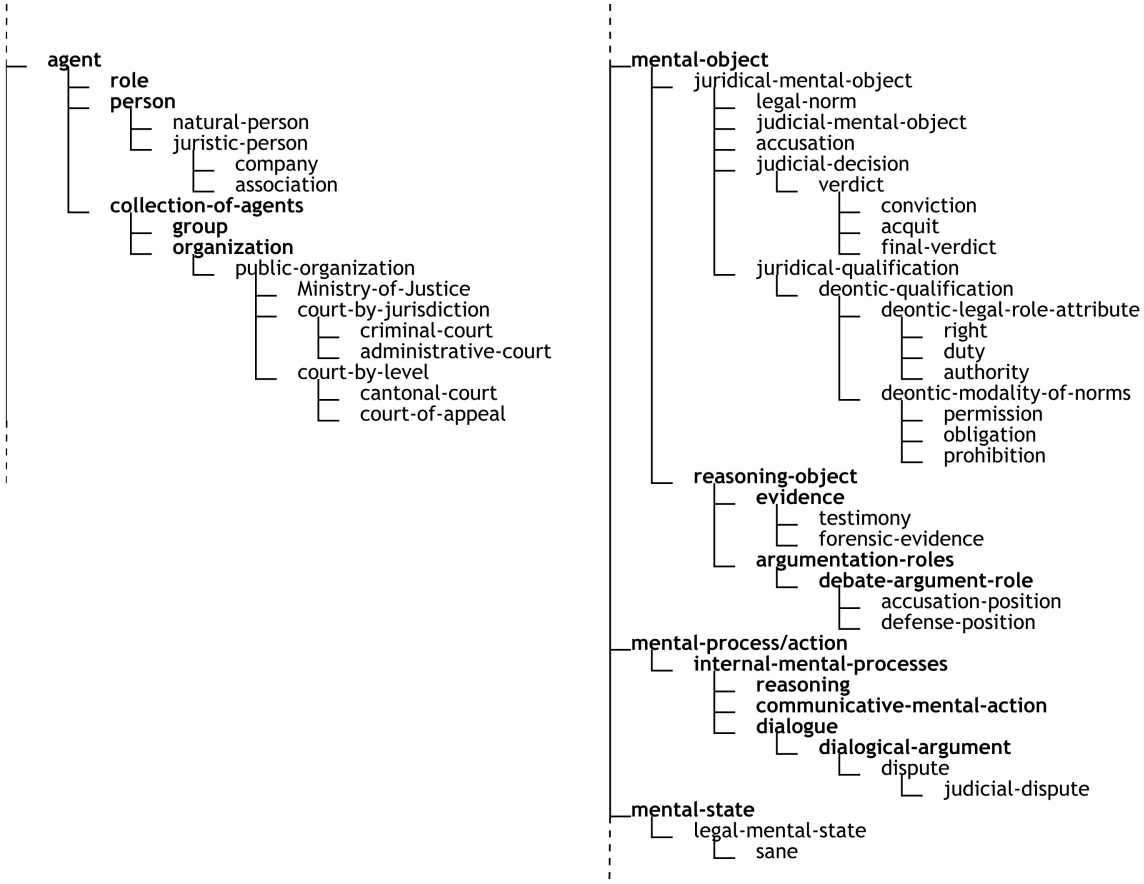

Figure 2: Some agents, mental objects, processes and states in Dutch Criminal Law (OCL.NL) (excerpt)

These anchor-points are not only useful to attach the legal sub-classes and composites. They provide a checklist and, more importantly, they foster the understanding that legal concepts most often imply several perspectives. This is not the same as the classical ambiguities that terms may have. For instance, the most important aspect of a legal document is its content, but the law often requires that the document is physically unique as well, and that it has a prescribed structure or elements. By multiple classification (and inheritance) these points of view can be easily combined and distinguished. Moreover, during the modeling the related points of view may suggest additional classifications and can be used for consistency checking.

We can illustrate the use of anchors in the LRI-Core ontology with parts of the ontology for Dutch criminal law (OCL.NL). In Figure 2 the boldface terms are terms from LRI-Core. LRI-Core knows about the distinction between a person as a lifetime identity and roles that a person may perform during this life. Roles and persons are both agents, and agents are both physical and mental objects. We need this perspective to be able to understand what is meant by the generic statement that "drivers of vehicles should keep to the right": drivers are roles that can perform actions [4]. However, we should also be able to interpret a statement from a case description that says that 'Alexander Boer did not keep to the right of the A-5 with his car', in such a way that Alexander Boer is a 'natural person' that acted in the role of 'driver' and performed the actor-role in the 'keeping' (= driving) action.

In Figure 2 a selection of typical legal roles is presented. In LRI-Core we distinguish between social roles and social functions. Social functions are external roles of organizations.

[4]To be precise, there are two kinds of roles involved here: the role of a person to play 'driver' and the actor-role where the driver performs the drive action. These latter roles are roles of actions, while the former roles are roles of agents.

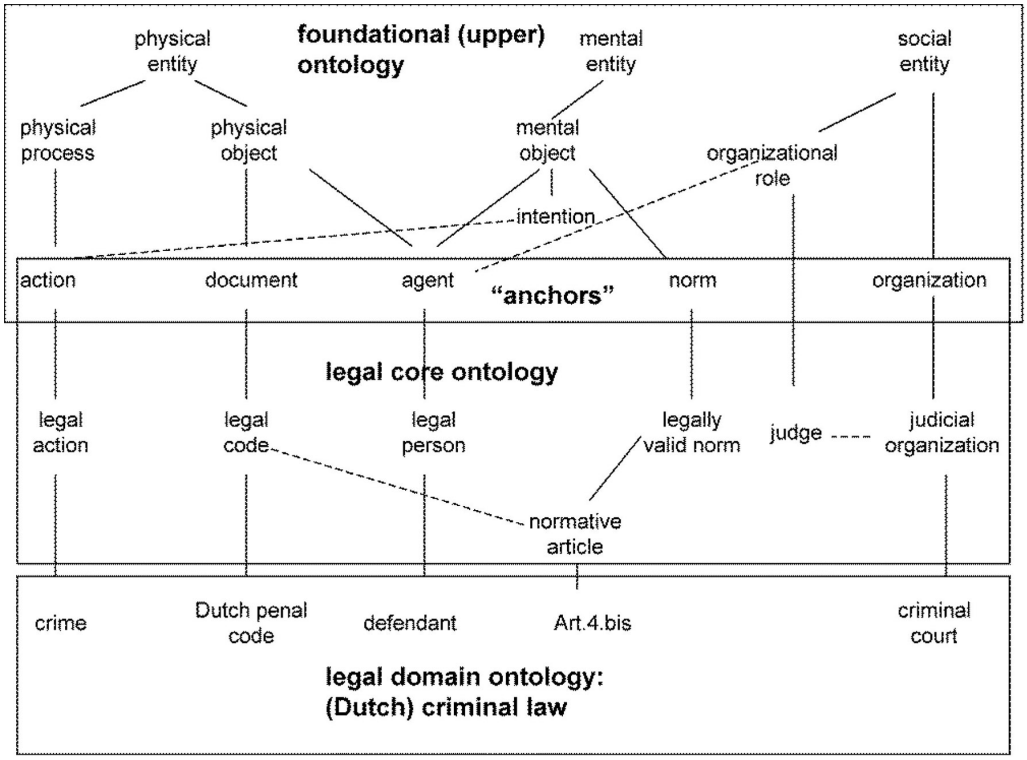

Figure 3: The structure of "cascading" ontologies

Social roles make up the functional internal structure of an organization. In these figures we cannot show multiple classification, nor other relations between classes than subsumption. For instance, an organization has social functions and 'has-as-parts' social roles. This is not the only view on the composition of an organization. The hierarchy of authority is another one, but this hierarchy maps onto the roles: authority is a mental entity: to be precise a 'deontic-legal-role-attribute'.

Figure 2 gives also in a nutshell some of the major categories of the mental world. As said before, the mental world contains many metaphores of the physical world, but it is in no way a direct mapping. It provides a vocabulary of the folk (naive) psychology and sociology we apply when thinking about and modeling the mental world. For instance, [3] present a convincing account of the primacy of conceptual schemas about physical processes that are metaphorized to conceptualize arithmetic, respectively full mathematics. We have to model mental worlds in order to understand one-self, but more importantly to interpret and understand the actions and expressions of others. Note that there is no need to ground this ontology on a biological-physical basis (reductionism). Mental objects are as much real and first class citizens as physical objects. We avoid commitments to the classical mind-body issues by having both a physical view and a mental view on agents (see [6, pp 87 ff])

Many objects of the mental world are reifications of epistemological roles. Terms like 'reason', 'evidence','explanation', 'problem', 'dispute' etc. come from the vocabulary of reasoning methods and are concerned with assessing the (trust in) the truth of beliefs. As stated in the Introduction, law is particularly concerned with terms that act like handles to come to grips with justifying legal decisions. In fact, one may see even terms like 'obligation', 'prohibition', etc. to objectify the imperatives of (illocutionary) discourse. The statement that 'vehicles **should** keep to the right' is reified as an obligation.

The hard core of the OCL.NL consists of actions. There are two major types: the criminal actions themselves (called 'offences'). These are of course the actions executed by the person who is successively acting as suspect, defendant, and eventually convict (if true and

proven. . .). On the other side, the convict may be at the receiving end of the 'punishment' actions, that are declared by the legal system etc. Crime and punishment are the keys to criminal law that is synonym to penal law. In Figure 3 the relationships between LRI-Core and legal domain ontologies (criminal law in various legal systems) are exemplified.

3 Legal Information retrieval in e-COURT

3.1 Outline of the Information Retrieval Process

In e-Court, two user modes of search are used: basic and advanced. The basic search mode allows metadata and/or keyword search by specifying values for one or more metadata fields and/or keywords. The advanced search mode includes possibilities to use linguistic weights and quantifiers with the keywords, to select the language of the query and the searched documents; to choose particular document sections of interest; to use multilingual capabilities (query translation), etc. In this section we describe the specific additional information management functions that are supported by ontologies.

3.2 Annotation and XML Tagging of Legal Documents

In information management the emphasis has been on archiving and retrieving documents by their formal, syntactic characteristics. These structures are abstracted in meta-data: RDBM schemas, DTDs for XML-tags, XML-Schemata, etc. This works fine as long as the structures are rather fixed and the occurrence of parts ('sections') is easy to identify in an automatic way. However, the criminal trial *hearing* documents in e-COURT have a large degree of structural variability. They reflect in the first place oral, often 'spontaneous' **dialogue** from the court room. Some aspects of the dialogue structure can be identified (semi-)automatically by voice recognition (turn-taking). Also, prescribed **phases** may be easy to identify. However, when it comes to the deeper, legal structure of what happens in court, even hand-crafted annotation may sometimes be problematic. The court sessions have two major functions: (1) providing evidence, e.g. in hearing witnesses, and (2) providing argument, e.g. in the plea of a lawyer or in a debate. Both may be produced in a rather unstructured way. Witnesses may give long accounts of events and interpretation; lawyers may plea and get interrupted; judges may change order of proceeding, etc. Arguments are not produced in a fixed format, and may be presented as analogies, as counter-evidence, as (rhetorical) questions, as hypotheticals, and even in the form of irony. Although central to legal hearings, the structure of the debate is the hardest part to make explicit and for long no candidate for automation.

As far as content is concerned, there are two major types of subjects. The accounts of what happened are largely in terms of common sense. These terms are automatically indexable and identifiable by using rather shallow but very useful lexicons like Wordnet. A second type of subject is more technical and concerns the criminal legal concepts and references used. As the users of the e-Court documentation system are in the first place (but not exclusevely) legal professionals, the emphasis in content-support is initially on ontologies of criminal law.

The role of ontologies in indexing and annotating the e-Court hearing documents is threefold:

- The first role is an indirect one: the ontologies provide the structured vocabulary for meta-data descriptions and maintain consistent use and semantic distinctions. The XML-Schemata only provide 'syntactic', structural information, but the ontologies (expressed in RDF/S) enable semantic coherence and verification.

- Although we may design DTDs (or XML-Schemata) in advance to capture dialogue-turns, phasic structuring, and argument-roles, most of these cannot be identified and tagged in an automatic way in the documents themselves. In most cases this can only be performed by a human agent, e.g. the transcriber who is capable of understanding what is (legally) going on in the hearing. This identification process is supported by a browser that gives options for annotating/tagging (in a context sensitive way) to structure the hearing transcripts. In such a way we may obtain multiple structuring of the hearing documents that increases the search options of the user. The identification of dialogue-turns can be (almost) fully automated by the use of simple voice-recognition devices that have only to distinguish voice characteristics of the participants in the dialogue.

- The e-COURT system indexes all documents. A number of these indexed terms correspond with terms of the ontologies. In this way we can link documents automatically with some semantics, i.e. one may gather what the document is about, which is functionally equivalent to (XML)-tagging the document with these terms. [5]

3.3 Query Expansion

The set of keywords used in a query can yield unsatisfactory results because the actual use of terms in a document may not correspond to what the user has in mind. This is obvious in the use of synonyms. However, also more abstract terms may be used to denote a more specific object: e.g. *killing* (synonym: *manslaughter*) for *murder*. A reference to a *murder* may be missed because in the document the terms *killing* and *manslaughter* are used. The reverse may also be relevant in information retrieval. The user may search for the *weapon* that is used in a particular criminal case, but may not know what kind of weapon exactly was used. By browsing a taxonomy of weapons (e.g. as part of an ontology of terms in criminal law) she may specify the query further.

In both search modes (basic and advanced) the ontology repository is consulted for subsumed or subsuming terms with respect to the keywords given. These terms appear in a browser and provide a focussed thesaurus to select additional terms (keywords). More specifically, those terms that occur in the ontologies and which are also in the index of a (set of) document(s) may be selected/highlighted in the browser (see Section 3.2).

Expansion by Subsuming Classes By adding terms for searching that are superclasses [6] of the already specified terms, the search is directed also to the more general, abstract terms. In searching documents that contain regulations (laws, statutes, contracts) where applicable provisions are often formulated in generalized and abstract terms, this IR strategy is in fact the only one to avoid false negatives (i.e. missed applicable provisions). In the CLIME project (IST-25.414) this strategy has been implemented by the University of Amsterdam as part of the MILE demonstrator and it has been used to determine the applicability of norms on the basis of an ontology of about 3,500 terms ([9]).

Expansion by Subsumed Classes The example of the search for a *weapon* above shows the problem when the user is searching for a subclass of a term she may well know. There are two possibilities. The user may allow all subsumed terms to participate as keywords in the search (which may lead to an explosive return of candidates) or she may have already restricted the set of possible documents and have a look at those *weapons* that occur as

[5] For pragmatic reasons we have provisionally opted for this solution, although XML-tagging is the method to be used on the web.

[6] One may also include 'wholes' from part-of hierarchies.

indices of these documents. In fact, the example is typical for the kind of searches where one is looking for additional, very specific information that should answer a question. In those cases, the user usually has specific cases in mind: even has the document already retrieved but has to find the exact information.

Disambiguation of a keyword term is another role of ontologies in IR. Classical ambiguity consists of terms that have different meanings but the same 'orthography'. Except for orthographic coincidences, most ambiguous terms in fact share meaning, besides their differences. Disambiguation occurs in the context of use and is a matter of 'degree'. There may be little ambiguity in the term *car* as an isolated term, but there is little overlap in what it implies between the mechanic's and the salesman's view of *cars*, even if they work for the same company. In ontologies persistent, but context (role) dependent ambiguity is represented as **multiple classification**.

Except for disambiguation and selective use of terms of subsumed classes, the additional terms are added as disjunctive keywords to the query set, which means that the set of documents that is returned – the *'result set'* – may have increased exponentially. One may find more correct returns, but one must be prepared for a large amount of false positives: the classical problem of information overload we try to avoid and for which the major web stakeholders (at least the W3C) see the solution in the semantic web technology. It appears there is not a free lunch at the web, nor at e-COURT that seeks the same solutions. There are two methods to cope with this problem. The first one is to have the user refine his query. However, this is often a problem because the user may not have enough information to do this.

A second solution consists of *(re)organizing the result set*. The typical problem in (WWW) information search is that the number of returned documents may be unmanageably large and heterogenous. The cause of much heterogeneity is the fact that a term may have multiple senses/views. In particular, the legal (criminal) domain is full of multiple views as we explained in Section 2.2, so we expect that disambiguation may occur by not only matching the indices of the returned documents with the keywords, but also have a second filtering/clustering where we also match indices with associated terms in the ontologies, i.e. the *value(-classe)s* and other related terms in the ontologies.

4 Conclusions

At first sight it may seem that the use of ontologies in legal information retrieval and storage is not paralleled by the effort in creating high level and richly specified ontologies. The ontologies are developed in Protégé using the DAML+OIL knowledge model and the FACT theorem prover. One may object that such a relatively 'heavy' apparatus with constrained expressiveness is not really necessary. A majority of the information retrieval and storage functions can also be supported by relatively simple lexicons. In fact, we will even use for a first version such a lexical approach as built in in Oracle because we want to start experimenting as soon as possible with the retrieval functions. However, there are various reasons to use a formally well grounded knowledge representation formalism and to specify the concepts as much specifically as possible (i.e. using attributes and axioms). This will help us:

- In the first place to verify the consistency of the ontologies created. Informal modeling in ontologies does not give any check on errors other than some kind of visual inspection, For ontologies larger than 200 terms this becomes unmanageable. We do not advocate here strict and formal modeling in a kind of straightjacket, but our experiences show that particularly in designing the basic framework for an ontology, consistency checking plays a very important diagnostic role.

- The second reason is that lexicons do not allow for multiple classification and inheritance. In these legal domains this is certainly required.
- Another reason is that we do not only need the terms as they occur in some classification hierarchy or lattice, but also their attributes, values and relations with other terms. This is required for the disambiguation and for clustering returned documents.

References

[1] J.A. Breuker and A. Boer. Developing ontologies for legal information serving and management. In *Proceedings of the EKAW-2002 workshop on Knowledge Management through Corporate Semantic Webs*, 2002.

[2] J. Hage and B. Verheij. The law as a dynamic interconnected system of states of affairs: a legal top ontology. *International Journal of Human Computer Studies*, 51:1034–1077, 1999.

[3] George Lakoff and Rafael Núñez. *Where Mathematics Comes From*. Basic Books, 2000.

[4] J. Lehmann and J.A. Breuker. On defining ontologies and typologies of objects and processes for causal reasoning. In A. Pease, C. Menzel, M. Uschold, and L. Obrst, editors, *Proceedings of IEEE Standard Upper Ontology*, pages 31 – 36, Menlo Park, 2001. AAAI-Press.

[5] John F. Sowa. *Knowledge Representation: Logical Philosophical, and Computational Foundations*. Brooks Cole Publishing Co, Pacific Grove, CA, 2000.

[6] P.F. Strawson. *Individuals*. Methuen, 1959.

[7] A. Valente, J.A. Breuker, and P.W. Brouwer. Legal modelling and automated reasoning with ON-LINE. *International Journal of Human Computer Studies*, 51:1079–1126, 1999.

[8] R. van Kralingen, P. Visser, T. Bench-Capon, and van den Herik H. A principled approach to developing legal knowledge systems. *International Journal of Human Computer Studies*, 51:1127–1154, 1999.

[9] Radboud Winkels, Alexander Boer, and Rinke Hoekstra. CLIME: Lessons learned in legal information serving. In Frank van Harmelen, editor, *Proceedings of the European Conference on Artificial Intelligence-2002, Lyon (F)*, Amsterdam, 2002. IOS-Press.