The construction and use of ontologies of criminal law in the e-Court European project

Joost Breuker
University of Amsterdam
Leibniz Research Institute for law (LRI)
P.O. Box 1030
NL 1000 BA Amsterdam, the Netherlands
breuker@lri.jur.uva.nl

October 25, 2003

Abstract

In this paper we describe the nature and use of various ontologies (knowledge bases) for the information management of documents coming from and related to 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. This upper ontology differs from existing (proposed) upper ontologies (e.g. SUMO or CYC) in several ways. One is that mental and social worlds are under-represented, while legal domains refer to social and mental concepts. The role of LRI-core is exemplified by an ontology about Dutch criminal law. Ontologies play also an important role in providing vocabularies that can be used to index and retrieve documents: in law the most typical kind of document are regulations (laws, provisions, contracts). We report work on the structuring of these legal sources that is part of the MetaLex iniative. 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 is based on emerging standards of the semantic web and we present arguments why AI based technologies reasons why one should not do with apparently simpler, non AI-based technologies.

1 Introduction

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

[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,

The e-COURT project is a European project (IST-2000-28199, www.intrasoft-intl.com/e-court) 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. The accessibility of the trial information is foreseen via the web, and more specifically: to enable forthcoming semantic web services by the use of RDF based ontologies and XML semantic annotation. Audio/video files are synchronised with transcriptions and other documents. Besides issues of work-flow management and security, advanced information retrieval (IR) functions are implemented. These functions are:

- *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 OWL, the language that will be the standard for semantics based web services: the Semantic Web (see www.w3c/2001/sw).
- *Database management:* multimedia documents (audio & video clips, text, pictures, etc.) supporting effective retrieval of (portions of) these. Documents are annotated and tagged (partioned) 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 – in particular European – regulations about the interchange of criminal information.

In this paper we will first discuss in Section 2 the various types of ontologies required to cover the legal domain (criminal law). Ontologies are thesauri of terms that are shared in a particular domain, e.g. criminal law. They differ from classical thesauri because the knowledge that is covered by the terms is specified in machine readable form and expressed in a so-called knowledge representation language. A knowledge representation language allows machines to reason with and about the meaning (semantics) of these terms. These meanings should reflect the shared understanding of these terms – concepts – between humans. Transfered to machines as knowledge bases, they enable computer programs to share this same understanding when they communicate with one another or with humans. Such computer programs are often called (artificial) agents. In particular in managing internet mediated transactions, these ‘agents’ play a more and more important role. However, in this paper the emhasis on the use of ontologies is on the way they may help humans to manage and access large amounts of (legal) information and documentation residing in machines.

---

and E-POWER (IST 28125 see www.lri.jur.uva.nl/research/epower.html

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 some highly 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 Section 2.3 the typical structure of regulations is discussed and how it is used to develop a standard for XML-based-tagging of these kinds of documents as part of the MetaLex initiative (see: www.metalex.nl). Ontologies expressed in RDF/RDF-S are the basis of for identifying the structures of regulations (Section 2.3.1) and of transcripts of criminal trial hearings (Section 2.3.2). 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 (Section 3.3) and in clustering result sets (Section 3.4).

2 Concepts of law and legal documents

Different from medicine, engineering or psychology, law is not "ontologically" founded. For instance, in legal theory or legal philosophy 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 and explaining the world. Therefore, it is no surprise to see that 'core ontologies' about law are rather epistemic frameworks (see e.g. also [Van Kralingen *et al.*, 1999], [Hage & Verheij, 1999]). In particular, FOLaw, the Functional Ontology for Law, developed by Valente (see [Valente *et al.*, 1999]) is to be viewed as a (CommonKADS) inference structure, despite what the authors claim it to be. The dependencies between the various types of knowledge as depicted in Figure 1 in fact constitute the typical structure of argument in solving legal cases.

This framework, which connects types of knowledge with epistemic roles in the way the legal system reasons about normative and responsibility decisions, is particularly useful in analyzing and modelling legal reasoning. It has been used as the basis of several artificial legal reasoning architectures, and worked fine in analyzing regulations for this purpose. It has been the basis of practical applications, in particular in systems for assessing whether provisions are applicable to a case (see [Winkels *et al.*, 2002] which reports its use in legal information serving in the CLIME project). However, 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. Norms are modeled as deontically qualified (generic) descriptions of situations. For instance, article 3 of the Dutch traffic code, which reads "all vehicles should keep to the right hand side of the road", is to be represented as that the situation in which a vehicle is on the right

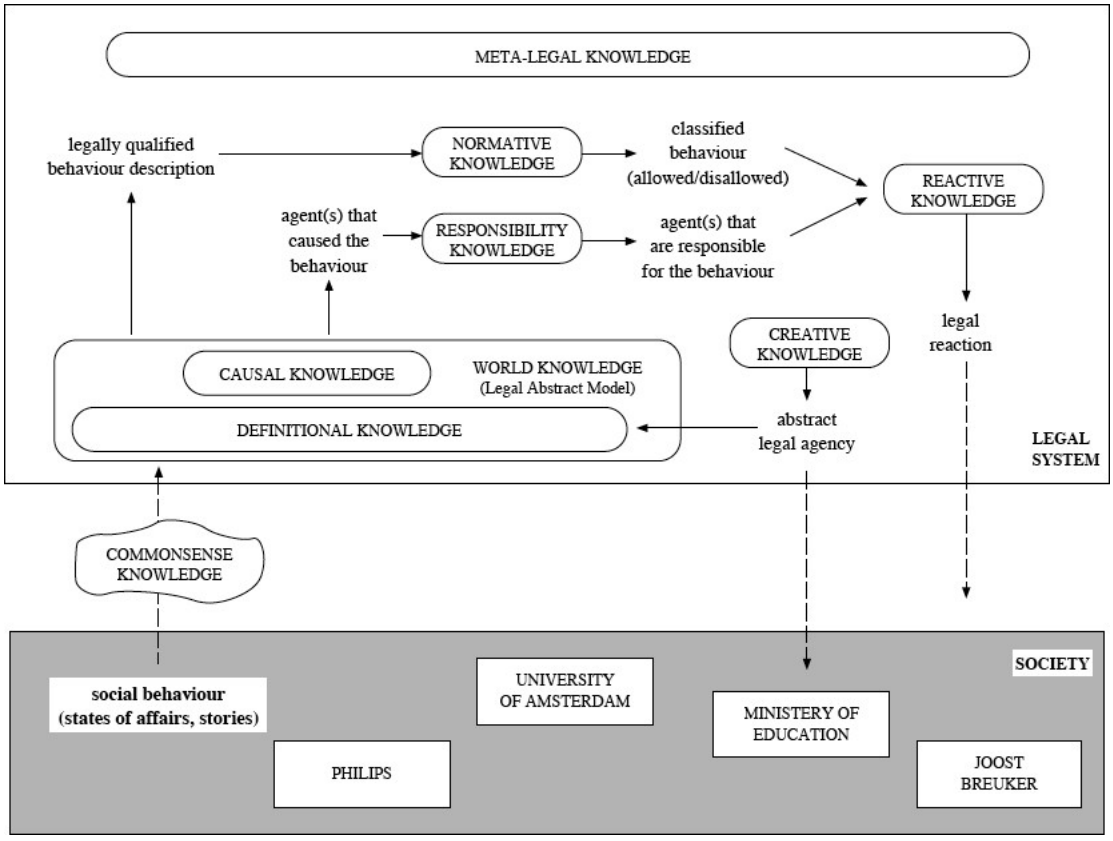

Figure 1: Roles of types of knowledge in FOLaw (see Valente et al, 1999 for details

hand side of the road is obligatory (see [Valente *et al.*, 1999] for details.). A norm is no more or less than a qualification of a generic situation. Legislation refers to social situations and activities in general terms and these situations are qualified as to be desirable 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, either more precise or more 'open texture' kind of definitions of these entities (see 'definitional knowledge' in FOLaw), but essentially most is 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 foundational [2] ontology appears to be indispensable, because the concepts of law are spread over almost the full range of common sense. A foundational ontology contains our understanding of very abstract concepts, like time, space, causality, physical objects, agenthood, etc.

We could not simply start with one of the currently available foundational ontologies (e.g. [Sowa, 2000], 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

[2] Often also known as: top, or upper ontology.

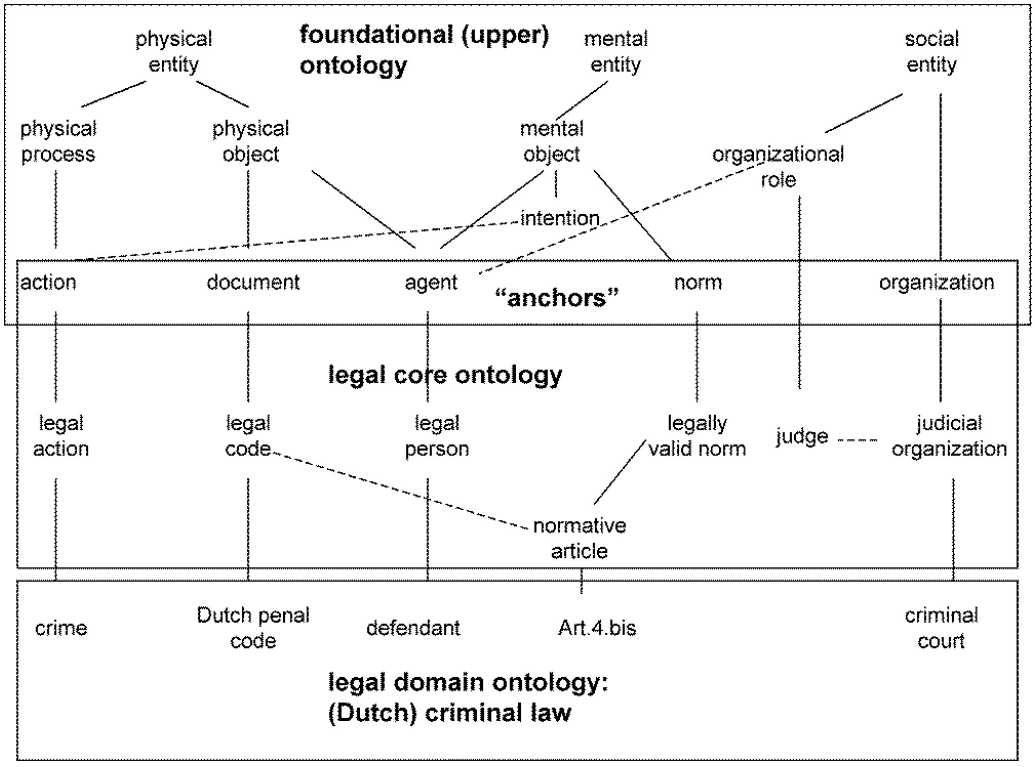

Figure 2: Layers of ontologies illustrated by relations between some typical concepts

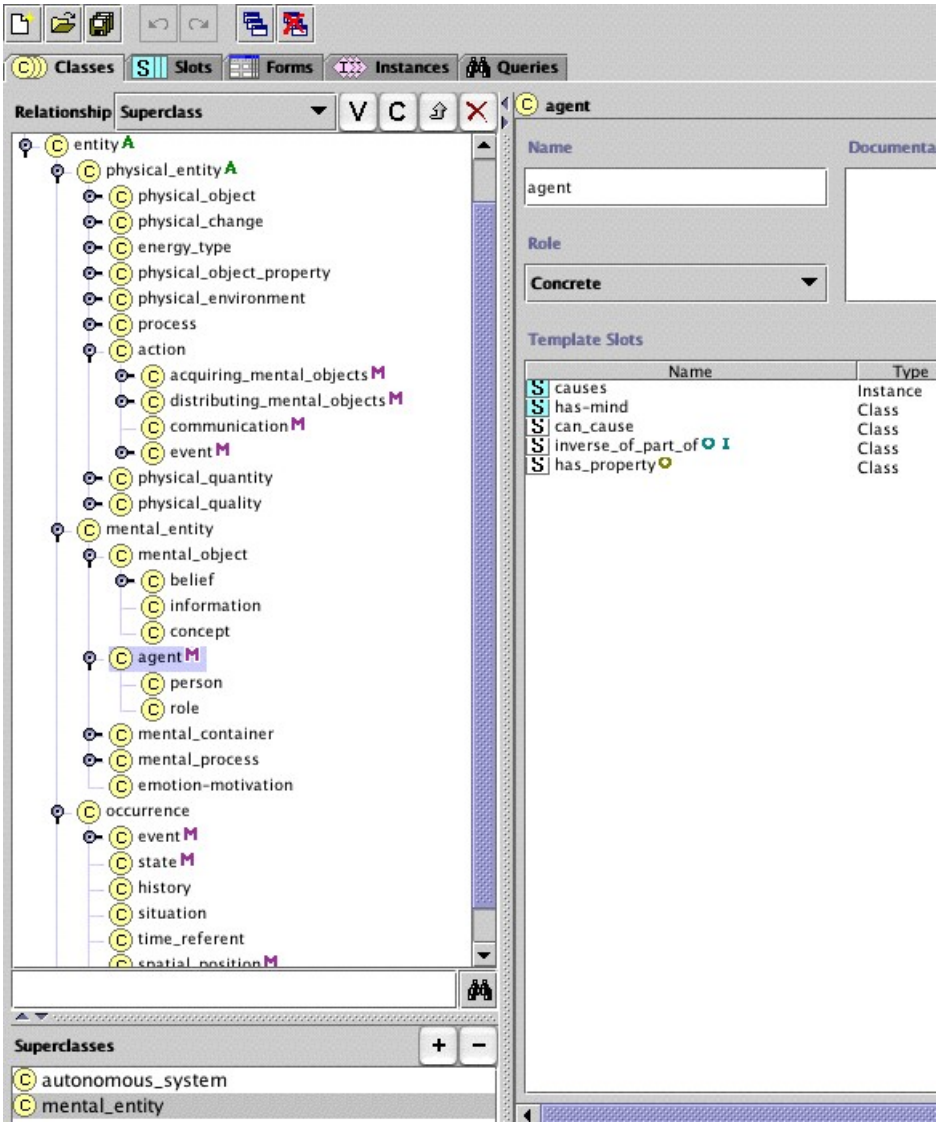

Figure 3: LRI-Core in Protege

more typical for law. Besides this lack of of sufficient covering, we did not agree about the physical part anyway (see [Breuker & Winkels, 2004] for reasons). Disagreeing on the meaning and understanding of concepts that make up a foundational ontology is not new. In fact, since Aristotle's views on the physical world, philosophers a have not stopped critiquing each other views, and in the recent world of knowledge engineering things are not different, despite advances in e.g. our deeper understanding of physics. [3]

Why shouldn't we take the obvious lessons from these experiences and refrain from even trying to construct such a foundational ontology? There are several reasons, but the primary one is that this ontology is not meant to carry the full weight of a complete and comprehensive common sense knowledge base as the CYC system is aiming at. Neither do we have the aspiration that we should provide the basic interlingua for the Semantic Web, as the IEEE-SUO work intends. Our aim is only to get some more semantic structuring for typical legal concepts and We think we can do with no more than about 300-400 concepts. Figure 2 presents in a nutshell how highly abstract concepts help us to define concepts in legal domains. There are three layers of abstraction. The foundational ontology contains concepts that are general for all kinds of domains; the core-ontology contains concepts that are typical for law, and finally at the domain level, the concepts are those that occur in a legal domain, e.g. criminal law.

In fact, we have combined the foundational and core levels into one ontology: *LRI-core.* Figure 3 gives a snapshot of some concepts of LRI-core and how they are classified.

The major principles that have motivated this ontology 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. [4] 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.

- Communication proceeds via physical objects (documents, sounds) and actions (talk, reading) which represent mental objects (information). Therefore, LRI-core enables one to see legal documents both from the point of view of their physical existence and layout, and from the point of view of their (mental) content.

[3] In fact, these insights clash with our common-sense understanding of the world. Our intuitions on how the physical world behaves even clash with Newtonian physics.

[4] [Lakoff & Núñez, 2000] present a convincing account of the primacy of conceptual schemas about physical processes that are metaphorized to conceptualize arithmetic, respectively full mathematics.

- The mental and the physical world overlap in concept of 'agent'. It is ambiguous because 'agent' is classified as both a physical object and a mental object. [5]

- 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 provides some framework to get a coherent view on a particular legal domain ontology, but more importantly, by using a knowledge representation language we are able to trace many types of errors, such as contradictions and many omissions. **LRI-core** is written in OWL. OWL is the knowled representation language that is the standard to be used for the Semantic Web; an extension of the world wide web that is now under development (see www.w3c.org/2001/sw for both simple and highly technical information on these developments.) Figure 3 presents a screendump of LRI-Core as it is specified in Protégé, a kind of editor that automatically generates OWL code.

2.2 Criminal law

In the e-COURT project, the focus is on the documents produced during a criminal trial: the most important ones contain the transcriptions of hearings. The structure of this type of document is to some extent determined by the debate/dialogue nature of these hearings, but also by procedural requirements (see Section 2.3). These procedural requirements are in the first place part of criminal procedural (in legal jargon: 'formal') law and refined by specific 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. The case descriptions have a strong common-sense flavour and of course we do not intend to develop here a comprehensive common-sense ontology. [6] 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.

[5] The multiple view evades the classical mind-body problem.

[6] The legal professionals who are the intended users are in the first place interested in the legal aspects of the case. Also, criminal law contains already terms of criminal actions, means and objects. However, we also intend to experiment here with natural language ontologies (vocabularies) like Wordnet.

How much of this framework will be reusable and whether it is easy to map terms from these jurisdictions to one another is difficult to predict. This mapping is more complex than the mapping of the vocabularies of different languages (cf EURO-Wordnet), because criminal law has been the traditional internal concern of nations, and will be the last kind of legislations to get 'harmonized' in the European tuning of legislation. It will be largely an empirical enterprise. Studies of comparative criminal law will not be of much help here as these are rather concerned with characterizing the major principles and histories that make up similarities and differences between criminal legislations. Modelling and inference tools are still unknown to legal scholars and moreover there is a traditional resistance to come to agreement over conceptual matters of law: the nature of legal practice embodies dispute rather than agreement and cooperation. This high level of distinctiveness is another reason we want to ground, or 'anchor', the ontology of criminal law 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 being accused is a role of (legal) person, and that persons as agents can perform both physical and mental activities etc. These anchor-points are not only useful to attach the legal sub-classes and composites. They provide a checklist and more importantly they allow the understanding that a (legal) concept cannot be captured from only one perspective. By multiple classification 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 4 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. 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 [7]. 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 5 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. 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' (see Figure 6).

Figure 6 gives in a nutshell some of the major categories of the mental world. To some extent, the mental world contains many metaphores of the physical world, but it is in no way

[7] 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.

agent
— **role**
— **person**
— — natural person
— — juristic-person
— — — company
— — — association
— — — foundation
— **collection-of-agents**
— — **group**
— — **organization**
— — — public
— — — — Ministry-of-Justice
— — — — courts-by-jurisdiction
— — — — — criminal-court
— — — — — administrative-court
— — — — courts-by-level
— — — — — cantonal-court
— — — — — court-of-appeal
— — — — — Supreme-court

Figure 4: Agents in Dutch Criminal Law (OCL.NL) (excerpt)

role
- **social-function**
  - public-social-function
    - jurisdiction
      - public-prosecution
    - criminal-investigation
      - forensic-investigation
- **social-role**
  - legal-role
    - juridical-role
      - judicial-role
        - judge
          - judge-presiding
        - prosecution-role
          - public-prosecutor
        - defense-role
          - defense-counselor
        - defendant
          - principal-defendant
          - accessory-defendant
          - offender
            - convict
        - witness
      - clerk-of-court
      - lawyer
      - the-Regent
      - the-State
    - public-servant
    - owner-of-goods
    - owner-of-rights/duties
      - creditor
      - debtor

Figure 5: roles and functions in Dutch Criminal Law (OCL.NL) (excerpt)

mental-object
— juridical-mental-object
— — legal-norm
— — judicial-mental-object
— — complaint
— — accusation
— — judicial-decision
— — — verdict
— — — — conviction
— — — — acquit
— — — — final-verdict
— — juridical-qualification
— — — deontic-qualification
— — — — deontic-legal-role-attribute
— — — — — right
— — — — — duty
— — — — — authority
— — — — deontic-modalities-of-norms
— — — — — permission
— — — — — obligation
— — — — — prohibition
— **reasoning-object**
— — **evidence**
— — — testimony
— — — — eye-witness-testimony
— — — forensic-evidence
— — **problem-solving-role**
— — — **solution**
— — — **problem**
— — — **problem-solving-method**
— — **argumentation-roles**
— — — **debate-argument-role**
— — — — accusation-position
— — — — defense-position
mental-process/action
— **internal-mental-processes**
— — **reasoning**
— — **communicative-mental-action**
— — testifying
— — interrogating
— — **argument**
— — **dialogue**
— — — **dialogical argument**
— — — — dispute
— — — — — judicial-dispute
mental-state
— — legal-mental-state
— — — sane
— — — mental-incapacity

a direct mapping. It provides a vocabulary of the folk (naive) psychology and sociology we apply when thinking about and modeling the mental world [Lakoff & Núñez, 2000]. 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 mind-body issues by having both a physical view and a mental view on agents. [8]

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 (new) 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. [9]

An important distinction between physical and mental actions is that for the latter the effects may not be confined to the mental world. Thinking, memorizing etc, are actions that concern only the inner mental world, but talking, pointing, writing etc. are communicative acts that transfer mental objects via symbols to the physical world. If we want to describe these communicative acts including their intended effects we have to add illocutionary acts. When communicative acts interact, we use the term 'dialogue'. Communicative actions are modeled in the first place as mental actions, mediated by symbolic, physical representation. One may argue that this still contains the flavour of epistemology because it states a theory about how we acquire information, but it says nothing about how we justify this information.

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.

[8] At some level of granularity this may pose problems. For instance, folk psychology holds that the container of mental objects -the mind- coincides with a part of the body; the brain. Aside from the fact that this anatomical correspondence is neurologically and physiologically incorrect – the nervous system is highly influenced by non-neural physiological processes – this ontological commitment is not required at all to describe mental and physical activities. It does not matter to model mental actions by reference to body-parts. Traditionally, we think with our brain and feel with our heart, but this (fading-out) notion of folk psychology may be replaced by a more 'embodied mind' view that is emerging in cognitive science. However, for the purpose of interpreting mental actions we do not need some commitment to physical location of processing. Note also that we are talking here about folk psychology; not about the metaphysical question whether there is even such a problem.

[9] There is a strong tendency in law to do away even with these verbs of modality. In the Dutch instructions for drafting legislation, it is advised to avoid these verbs and put the statements as factual description, i.e. this article is expressed in the final version of the traffic code as: "drivers of vehicles keep to the right as much as possible". In terms of discourse this trespass of common sense modality rules conveys a certain arrogance. . . : the command is presented as a fact.

action
— criminal-action
— — offence
— — — felony
— — — — offence-against-the security-of-the-state
— — — — — taking-life-of-regent
— — — — theft-and-stripping
— — — — offences-against-human-life
— — — — — murder
— — — — — manslaughter
— — — — deception
— — — misdemeanour
— — — — lesser-offences-related-to-public-order
— — — — lesser-offences-related-to-public-moral
— — punishment
— — — principal-punishment
— — — — imprisonment
— — — — — for-life
— — — — — for-a-determinate-period
— — — — detention
— — — additional-punishment
— — — disqualification
— — — deprivation-of-a-right
— — — — deprivation-of-an-immunity
— — — — deprivation-of-a-privilege

Figure 7: Criminal actions in Dutch Criminal Law (excerpt)

2.3 Ontologies and the structure of legal documents

In knowledge and information management, there is a large variety of types of document-structures. If we look at texts documents may range from narrative texts (stories, histories, case descriptions, testimony) via 'non-narrative' texts (reports, articles, handbooks, instructions) to fully pre-structured filled-in forms. Legal documents cover this full range. Moreover, the actors in the legal domain deeply believe in the universal adequacy of textual expression. One may say, that legal practioners are text-fetishists. There is a good reason: the concern with evidence makes recording a basic requirement. However, the bad side is that written texts (on paper) are still the almost exclusive trustworthy media of recording and communication. In our bureaucratic society, documents are the universal basis of managing organizations. However, there are some types of documents which are almost unique to law and which play an important part in e-COURT.

The first ones are **regulations** (codes, legislation, contracts, etc.). They have already played a major role in other projects we are/were involved (E-POWER, CLIME) but they are also the object of international initiatives to arrive at standards for describing the structure of legislation by the use of XML tags. Ontologies play here the role of providing dictionaries for the tagging framework. The framework we developed for the MetaLex iniative, as described in the next section, will also be the basis for tagging documents that contain criminal law in e-COURT in order to be able to associate references to criminal issues with the appropriate articles. In the E-POWER project, this framework is both used to support drafting tax-regulations, as well as for linking it with other documents and also legal reasoning systems that refer to regulations. This framework and tools will be shortly described in the next subsection(Section 2.3.1).

Besides regulations, in e-COURT the major type of legal document are transcripts of hearings of criminal trials. We will only shortly describe our approach: the actual work is still under development (see Section 2.3.2).

2.3.1 Structure of regulations

The legal system presupposes a shared understanding of what norms exist and how they should affect behaviour. The norm communicates a 'social constraint'. It is not necessarily an agreement because the involved parties do not have to agree to the norm. Norms are intended to change people's preferences between choices, to interfere with basic economic behaviour by changing 'the rules' of the game. If one party has the means to punish or reward the other party, he may *de facto* change the behaviour of the other party without agreement. Norms must usually first be communicated in a **document** – a legal source – to come into existence. The document prescribes behaviour to agents assigned a certain role (e.g. the owner and user of an artifact). Sometimes the document posits design constraints for the creation of artifacts (made by agents; e.g. ships, tax forms) or procedural constraints for actions or transactions (by agents; e.g. survey, hearing, purchase) instead. There can be little argument about whether the law posits norms - that is what makes it law - but it is not as clear-cut how norms are to be distinguished from apparently different kinds of information. Legal sources also define, create, and even explain for the purpose of improving compliance with the norms.

Regulatory documents are very strange documents. They are never read from cover to cover; Each article presents a separate ‘discourse’. You can read the contained articles in any order, the resulting ‘discourse structure’, the message, is supposedly always the same. Each of those articles plays an independent role as ‘instrument’ in certain (epistemic) acts. The containing document is a special-purpose container that posits the article in the legal system and provides a position, an identity, by which it can be unambiguously referenced. The ‘legal source’ as an object with a unique identity and history is of course not the same as some paper or electronic copy of it.

A laborious process in both legal publishing and decision making is determining what the contents of legal sources are at some point or interval in time. Changes can be announced in separate documents and publishers keep track of all documents from certain publication channels to be able to reconstruct what the form of an organic law is at some time point[10]. Similarly, if you find a written decision on your doormat its validity status changes when a newly written decision that retracts it follows two days later. Each element, each sentence of a document can go through a complex lifecycle; It exists in some time-interval, it may be ‘active’ or ‘inactive’, and its scope of application may be extended or limited in contained time-intervals. To keep track of this lifecycle each element must have an object identity by which it can be referenced so that it can be positioned in time and in relation to other document elements.

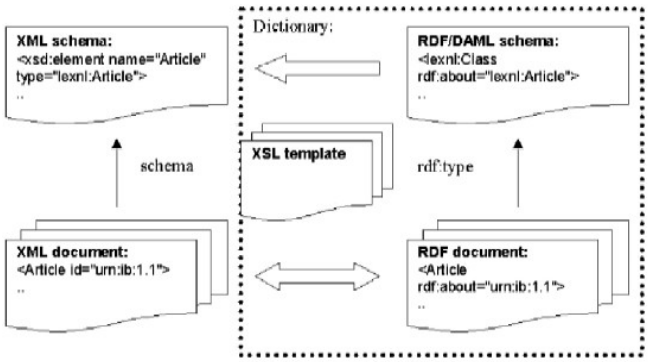

Figure 8: RDF representation of legal documents.

The MetaLex initiative[11] intends to provide a general and easily extensible framework for the XML encoding of the structure and contents of any type of public legal document. By describing legal concepts of different jurisdictions in a single RDF dictionary [12], it is supposedly easier to identify similarities and differences between legal concepts in different jurisdictions. The XML schemas we contribute to MetaLex [Boer *et al.*, 2002] aim to standardize structure

[10]Legal procedure may allow insertion of articles provided that that does not invalidate existing references. An article 3a may be inserted between 3 and 4, for instance, or article 3 may become 3a and member 3b is inserted so that it is subsumed in existing references if intended.

[11]http://www.metalex.nl

[12]http://rdf.MetaLex.de

and designation of identity in legal documents. The standard XML ID attribute can be attached to elements that represent document structure and the structure can be translated with XSL stylesheets to RDF[13] conforming to an RDF Schema. The RDF data model is considered normative for identity matching because it appears to be most suitable for that purpose. Figure 8 shows the relationship we propose between the XML Schema-based and RDF Schema-based encoding of the same document.

A well known limitation of standard XML is the lack of standardization of *global* object identity of elements and the interpretation of the meaning of references between elements. The ID attribute and standards for namespaces, (X)HTML, XPath, XPointer, XLink, and RDF all offer competing or complementary pieces of solutions to make XML parsing trees represent arbitrary graphs that link distinct individuals. RDF makes this underlying graph explicit and de-couples the identity of elements from the documents in which they are serialized (only positioning the element in a namespace – which may or may not correspond to a document). If a document element is encoded in RDF statements – triples of a *subject*, *predicate*, and *object* – it can be both subject and object of statements regardless of what document it is serialized in. This perspective is certainly more suitable for a world of 'organic' regulations that may never have been entirely published in their present form. A minor disadvantage of the use of RDF is that path-based XPointer references are not transparent. Every target of an XPointer-based link in the XML Schema-based version of a regulation must carry an ID before it can be resolved by the stylesheet that translates it to RDF.

Another notable difference between the MetaLex XML schemas and corresponding RDF schemas for documents is that RDF encoding requires explicit, indexed 'sequences' of e.g. articles, parts, and sentences because RDF is order-independent. Any order of serialization of an RDF model into RDF/XML results in a different XML parsing tree. RDF can for instance represent the existence of an unspecified 'hole' between a first and third sentence in an article. Once the RDF version of a document contains holes, it cannot be written in normal XML Schema-based XML anymore. This notion of a hole in the document representing missing information is not the same as the notion of a hole in an index used for designation in the regulation itself. If article 1 is followed by article 3 that does not imply the 'existence' of an article 2 in a legal source during the time-interval represented by the serialized XML document. Neither does the presence of an article 2 following article 1 contradict the possible existence of an article 1bis[14].

2.3.2 Hearings of criminal trials

Hearing documents reflect in the first place dialogues. Characteristic of dialogues is turn-taking (who is talking). Turn taking identified with the person/role of the one who is talking is a first, low level structuring of these documents. The tagging can be semi-automatically performed by using voice-recognition. These dialogues have different roles and modalities, most often related to phases in the trial. Besides 'ceremonial' steps in the phases, there are typical phases

[13] http://www.w3.org/RDF/

[14] A practice most common when printing was expensive, search engines non-existent, and correcting existing references to articles almost impossible.

such as testimony of eye-witnesses and experts, cross-examination, pleadings by the lawyer, oral verdict, etc. This structure is of a higher level than the dialogue and of more importance. Besides the explicit content of the dialogues, there are many references to other documents; not only to criminal legislation and precedence cases, but in particular to documents that are part of the case (declarations, testimonies from the investigation phase, forensic reports, etc.). For all these entities small scale ontologies are developed.

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.

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. The criminal trial *hearing* documents in e-COURT are not the typical kind of documents that are handled by information systems. These standard documents are written with more or less fixed, often prescribed structures, and strong control from the author(s) who may be able to annotate their documents as part of the authoring process. However, the hearing documents reflect in the first place oral, often 'spontaneous' **dialogue** from the court room. Besides dialogue, the courtroom trial sessions may have more or less formally prescribed **phases**: witnesses are consulted, they may give long accounts of events and interpretation; lawyers may plea and get interrupted; judges may change order of proceeding, etc. Finally, the hearing dialogues have the underlying structure of debate in which the content of the dialogue plays the role of **argument** and its support by (documented) evidence (see [Vreeswijk, 1997] [Prakken & Vreeswijk, 2000], [Gordon, 1995]). The arguments are not produced in a fixed format, nor are they always as explicit to allow e.g. a transcriber to identify the the nature of the argument and what it refers to. Arguments may be presented as analogogies, 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.

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

- The first role is an indirect one: the ontologies provide the structured vocabulary to construct meta-data descriptions and maintain consistent use and semantic distinctions.

The XML-Schemata only provide 'syntactic', structural information, but the ontologies (expressed in RDF-Schema) enable semantic coherence and verification. For instance, legal documents are not only identified by a number of dates, but contain also many dates. Statutes may contain even rather complex notions of (dated) time periods, but also the hearing transcripts are full of dates. To enable reasoning about dates **of** –,and **in** documents, the meta-data should still be linked to their common meaning in an ontology of time as points in time that may mark beginning or ends of periods etc. For the moment such reasoning capabilities are only required for verification purposes at the (semi-automatic) construction of the XML-Schemata, in the same way as they are to be used for the construction of the ontologies themselves. The next step is of course to enable these capabilities to be available at the actual use of the e-COURT system, in the same way as all these XML/RDF/Ontology layering provides the basis of intelligent, semantics-based web services.

- 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 browsers that give options for annotating/tagging (in a context sensitive way) to structure ('to section) 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. [15]

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 (information retriever) has in mind or expects. 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.

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

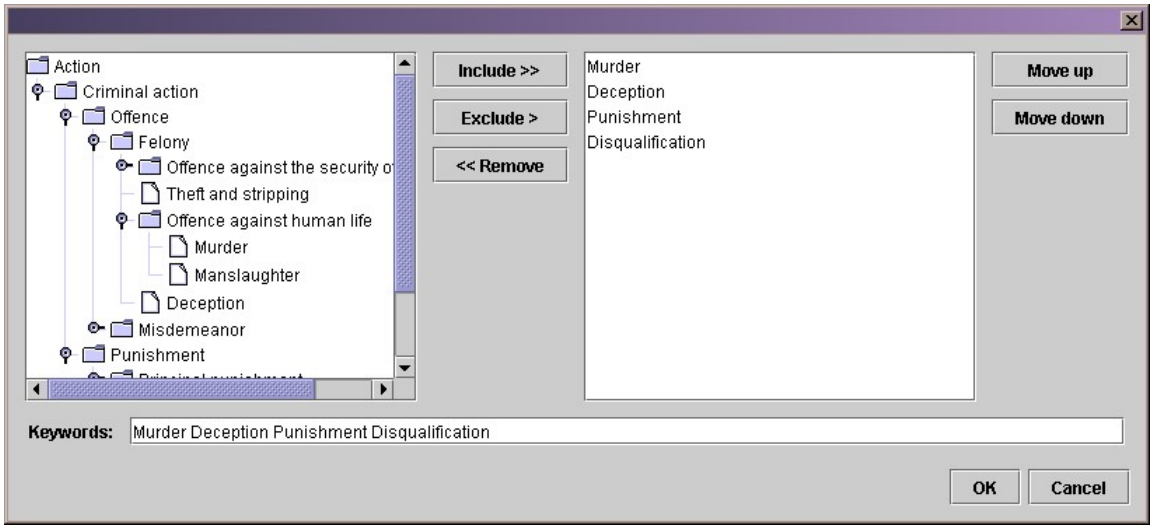

Figure 9: User interface of the e-Court browser for expanding queries based upon an ontology of crminal law

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 [16] 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 15.000 terms ([Winkels *et al.*, 2002]).

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 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

[16] We may also include 'wholes' from part-of hierarchies

those cases, the user usually has specific cases in mind: even has the document already retrieved but has to 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 and the salesman of *cars*, even if they work for the same company. In ontologies persistent, but context (role) dependent ambiguity is represented as **multiple classification**. For instance, the term *date* which denotes the occurrence of an event in time, may refer to an event described in a document, or to the date of creation or modification of a document itself. By showing the user that the keyword submitted is ambiguous in the context of e-COURT documents, misunderstandings or too many false positives may be prevented. Disambiguation works similarly to specification.

Figure 9 shows the browser interface of the eCourt system that enables the user to expand terms for search.

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 for this. The second one is described in the next section.

3.4 Reorganizing 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. We are currently working on clustering algorithms that should make the distinction, but our work is somewhat retarded by lack of a sizeable set of annotated and indexed hearing documents to do experiments and parameterize the algorithms. As long as the e-COURT system is not fully operational on a large scale, we have to do with artificially generated sets.

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 well represented ontologies. The ontologies are developed now in Protege, but it is intended to move as soon as possible to tools that reflect the upcoming standards for semantic-web ontologies, i.e. OWL. One may object that such a relatively 'heavy' apparatus with constrained expressiveness is not really necessary. Most of the information retrieval and storage functions can be supported by relatively simple lexicons. in fact, we will use for a first version such a lexical approach as built in in the latest version of Oracle because we want to start experimenting as soon as possible. However, there are various reasons to use a more richer, formally well grounded knowledge representation formalism. All these reasons are related to the fact that these formalisms allow trustworthy, proven reasoning methods. So why do we need these?

- 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. In later stages of knowledge acquisition, the consistency checking rather gets the role of tracing local errors and mistakes. Still, consistency checking and classification facilities cannot replace a good understanding of the domain.

- 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. In fact, one can see these values as additional query information that enables further distinctions. For instance, for the attribute-values for the car salesman are prices, accessories, etc: for the mechanic they are parts and part-numbers, while both may have to refer to the same models and versions of types of cars.

References

[Boer *et al.*, 2002] A. Boer, R. Hoekstra, R. Winkels, T. Van Engers, and F. Willaert. Proposal for a Dutch Legal XML standard. In *To appear in Proceedings of the 1st International Conference on electronic Government (EGOV2002)*, 2002.

[Breuker & Winkels, 2004] Joost Breuker and Radboud Winkels. Use and reuse of legal ontologies in knowledge engineering and information management. *Ar-*

tificial Intelligence and Law, (special issue on Legal Ontologies), 2004.

[Gordon, 1995] T.F. Gordon. *The Pleadings Game. An Artificial Intelligence Model of Procedural Justice*. Kluwer, Dordrecht, 1995.

[Hage & Verheij, 1999] 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.

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

[Lehmann & Breuker, 2001] 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.

[Prakken & Vreeswijk, 2000] H. Prakken and G. Vreeswijk. Logical systems for defeasible argumentation. In D. Gabbay, editor, *Handbook of Philosophical Logic*. Kluwer, 2000.

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

[Valente *et al.*, 1999] 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.

[Van Kralingen *et al.*, 1999] 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.

[Vreeswijk, 1997] G.A.W. Vreeswijk. Abstract argumentation systems. *Artificial Intelligence*, 90:223–277, 1997.

[Winkels *et al.*, 2002] 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.