# Proceedings of LOAIT 07

## *II Workshop on Legal Ontologies and Artificial Intelligence Techniques*

Pompeu Casanovas
Maria Angela Biasiotti
Enrico Francesconi
Maria Teresa Sagri (Eds.)

# Proceedings of LOAIT '07

# II Workshop on Legal Ontologies and Artificial Intelligence Techniques

Pompeu Casanovas
Maria Angela Biasiotti
Enrico Francesconi
Maria Teresa Sagri (Eds.)

# Contents

# The Ontology of Legal Possibilities and Legal Potentialities

Pamela N. Gray
*Centre for Research into Complex Systems,*
*Charles Sturt University, Bathurst, Australia*
*pgray@csu.edu.au*

**Abstract.** Ontologies in a legal expert system must be processed to suit all possible user cases within the field of law of the system. From the logical premises of a deductive system of express rules of law, legal ontologies may be implied to encompass the combinatorial explosion of possible cases that may lack one or more of the express antecedents in the deductive rule system. Express ontologies in inductive and abductive premises that are associated with the deductive antecedents, may also be adjusted by implication to suit the combinatorial explosion of possible cases. Implied legal ontologies may be determined to suit the user's case and its legal consequences. The method of this determination and the processing of express black letter law accordingly, is considered by reference to the supplementation of ontology by logic and the supplementation of logic by ontology, in the legal domain; three bases of this method are discussed: law-making power, prior analytics, and the pillars of truth in science and law.

Firstly, law-making authority includes the power to determine the logical category of legal premises, and legal truth tables (c.f. Wittgenstein, 1918); law is laid down as legal ontologies with logic attributes or structures. Thus, three ontological posits of law-makers provide for the logical processing of legal information. Rules of law are Major deductive premises laid down, formally or informally, as conditional propositions which may be systematised for extended deductive reasoning. Material facts in a case are laid down as inductive instances that particularise or define antecedents in rules of law; they also may be used as Minor deductive premises to determine the outcome of the case. Reasons for rules are laid down as and for strong or weak abductive reasoning.

Secondly, legal knowledge engineering requires prior analytics (cf. Aristotle, 1952, originally c.335 BC ) for the acquisition of the expertise; by prior analytics, premises are formalised and systematized for automation of their associated heuristics. Legal epistemology both determines and implements logical structures; through prior analytics it uses ontologies of legal possibilities and potentialities, to comprehensively predetermine premises for its three forms of legal logic: deduction, induction and abduction.

Thirdly, Lord Chancellor Bacon's (1620) reconstruction of legal epistemology as scientific method for expanding knowledge, systematizes the sources of truth in law and science. It is here developed as a method of prior analytics for constructing an ontology of legal possibilities and legal potentialities. Such ontological construction is essential for determining the heuristics of combinatorial explosion derived from express legal rules to meet the possible cases of users; while legal experts need only construct the relevant part of the combinatorial explosion, for a client's case, an expert system must be capable of constructing any relevant part to suit a user's case.

## 1. Ontology of Legal Possibilities and Legal Potentialities

In its meaning, a rule of law is concerned with what will happen if a situation or case exists; this is the nature of a rule because it has the form of a conditional proposition: 'if (antecedent(s)) then (consequent)'. The ontological situations that are explicit in law, might exist; law assumes an ontology of possibilities and potentialities. In the legal domain, reconfigurations of ontology in express rules of law, may produce a range of hypotheticals (cf. Rissland, 1985); the extent of the hypotheticals used by the legal profession is determined by what is the legal consequent if one or more of the antecedents in a rule of law do not exist, which is possible, or are given additions, which are realised potentialities.

## 2. Reconfiguration of Express Ontology

The reconfiguration of legal ontologies is a part of legal epistemology that was adapted for scientific method by Lord Chancellor Francis Bacon in the Second Book of his *Novum Organum* (1952, originally 1620). His system of four Tables, illustrated by the study of heat, allows consideration of (1) the attributes of heat through a range of instances of heat, (2) the attributes of a lack of heat through a range of instances of a lack of heat, (3) degrees or comparative instances of heat and lack of heat with causal observations on increasing and diminishing heat, then (4) the attributes of a lack of heat that are excluded from the attributes of heat. The pattern in the Tables is comparable to the pattern of pleadings in a court case; the *Novum Organum*, which was posed to replace Aristotle's work on ontology and logic, the *Organon*, was written just prior to Bacon's dismissal from office for taking bribes. He died a few years later from a chill suffered during his study of cold. Bacon explains his system as follows:

> The investigation of the forms proceeds thus: a nature being given, we must first present to the understanding all the known instances which agree in the same nature, although the subject matter be considerably diversified. And this collection must be made as a mere history, and without any premature reflection, or too great degree of refinement....
> Negatives, therefore, must be classed under the affirmatives, and the want of the given nature must be inquired into more particularly... (p.141)
> In the third place we must exhibit to the understanding the instances in which that nature, which is the object of our inquiries, is present in a greater or less degree, either by comparing its increase and decrease in the same object, or its degree in different objects... no nature can be considered a real form which

does not uniformly diminish and increase with the given nature. (p.145)
For on an individual review of all the instances a nature is to be found... man...is only allowed to proceed first by negatives, and then to conclude with affirmatives, after every species of exclusion.
We must now offer an example of the exclusion or rejection of natures found by the tables of review, not to be of the form of heat; first premising that not only each table is sufficient for the rejection of any nature, but even in each single instance contained in them. For it is clear from what has been said that every contradictory instance destroys an hypothesis as to form. Still, however, for the sake of clearness, and in order to show more plainly the use of the tables, we redouble or repeat the exclusive. (p.149)
In the exclusive table are laid the foundations of true induction, which is not, however, completed until the affirmative be attained... And, indeed, in the interpretation of nature the mind is to be so prepared and formed, as to rest itself on proper degrees of certainty, and yet to remember (especially at first) that what is present depends much upon what remains behind. (p.150)

Bacon's father was also Lord Chancellor in his time, so Francis, who had studied at Cambridge University and at Gray's Inn, was well imbued with legal epistemology. At the outset of adapting legal method to science, Bacon observed:

Although there is a most intimate connection, and almost an identity between the ways of human power and human knowledge, yet, on account of the pernicious and inveterate habit of dwelling upon abstractions, it is by far the safest method to commence and build up the sciences from those foundations which bear a relation to the practical division, and to let them mark out and limit the theoretical. (p.137)

Bacon set out his method for science to 'superinduce' (p.137) knowledge. Scientific knowledge must look to its inductive instances as the source of truth that can be carried through to establish its Major deductive premises; whereas law looks to law-making power for the 'truth' of its Major deductive premises which then determine the scope of its inductive instances in cases (cf. Ashley, 1990).
A case is now pleaded in a variable Statement of Claim as one or more form(s) of action; this requires a statement of how the case facts of an action satisfy the relevant rules. The facts of the case must particularise the antecedents in the relevant rules and state the Final consequent of those rules in terms of the claim, as well as the orders that thereby are sought. Where several rules that are connected are relied on, the interim conlusions that connect the rules must be set out in the statement as matters that are particularised by the facts of the case. Where there are no rules to rely on, an action on the case may be pleaded, with facts suggesting new rules or a certain exercise of discretion by reference to relevant factors.

Issues of fact and law are resolved through the further pleadings, namely the Defence and Counterclaim, and Reply, if any. The defence will indicate which facts in the Statement of claim are denied and which rules or part of rules in a Statement of claim are joined in issue by the defendant; the defence relies on contradictions of the facts pleaded by the Plaintiff, and the rules that deal with such failures to establish a claim. The defence may plead further facts. If the further facts pleaded by the defendant amount to a claim against the plaintiff, then they must be pleaded as a Counterclaim, which is like a Statement of claim by the defendant. Only pleaded matters may be raised and relied on at the trial; the parties are confined to these matters and issues.

## 3. Legal Epistemology

Law-making authorities, who provide truth to the rules of law as Major premises for the modus ponens deductive syllogism, and truth to premises adopted for inductive and abductive support for the law, lay down law and its associated premises in ontological posits as legal ontologies with integral logic structures. These posits may be compared to the monads of Leibniz (1714), the a priori principles of Kant (1788, 1955), and the epistemes of Foucault (1969); the concept of a paradigm (Kuhn, 1970) also bears a fusion of ontology and epistemology. The fusion reconciles the jurisprudence of legal positivism and analytical jurisprudence. The ontological posits determine the sort of logical use that can be made of the premises in the posits; there are three sorts of ontological posits in the legal domain, where legal ontologies are laid down, namely:

1. deductive premises in the form of rules for use in extended deduction,

2. inductive premises which may be formalised as existential statements that are definitional, and are usually the material facts of cases for use in induction as instances of antecedents or instances of consequents in the deductive rules; inductive instances may be extended by common knowledge and dictionaries of synonyms and antonyms, and

3. abductive premises for use as reasons for rules or reasons for case decisions about rules. Abductive premises may provide strong or weak reasons; there may be abductive premises which are so strong that they displace or justify modification of a deductive rule.

When a conditional proposition, stated formally or informally is said to be a law by a law-making authority, then this description means that it can be treated as true in a syllogistic application of its ontology as a Major premise for deductive application. Material facts of cases that satisfy an antecedent in a rule, are judicially asserted as inductive instances of an antecedent in a rule. If a premise is said to be a reason for a rule or for an accepted instance of an antecedent or consequent in a rule, then it is laid down as an abductive premise that strengthens the deductive or definitional necessity of the application of rules. A strong abductive premise that weakens a rule may break the deductive necessity of the rule and change the rule. What a judgment says of a premise, determines the logical nature of its ontological posit.

## 4. Systematic determination of ontology of legal possibilities

Four steps (cf. Bacon, 1620, 1952) are required to systematically ascertain the full extent of the ontology of legal possibilities: (1) the determination of the extended deductive order of deductive posits, (2) the determination of contradictories and uncertainties in extended deductive order, (3) the determination of inductive posits, their contradictories and uncertainties, and (4) the determination of abductive posits, their contradictories and uncertainties.

### 4.1. Extended deductive order

To establish the possible cases within the scope of the express black letter law, that are the extent of possible legal ontologies, the express legal ontologies of black letter law are initially formalised as the antecedents and/or consequents of the system of rules of law that permit extended deduction; every formalised rule is a Major deductive premise in an extended deductive order whereby rules become linked continuously. Susskind (1987, p.146), the champion of rule base systems, pointed out as crucial, the nature of this linking:

> ... the consequents of some rules function as the antecedents of others.

Thus, if a consequent of one rule is established when all its antecedents are established by the facts of a case, that consequent may be used as an established antecedent in a second rule, to establish, along with further facts of a case that establish any other antecedents in the second rule, the consequent of that second rule, and so on, in a sequence of extended deduction. This phenomenon produces rule hierarchies which prima facie have mixed components of law and fact that may raise issues of fact, issues of law or mixed issues of law and fact

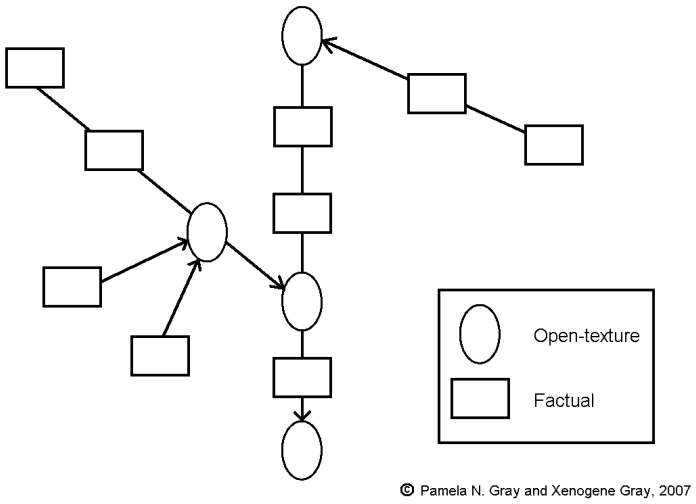



*Figure 1.* River of open texture ovals and factual nodes

in a particular case; the mix of components is evident in the directed acyclical graph of Popple (1996, p. 71), which is further developed in Figure 1, as an eGanges River of open-texture ovals and factual nodes. In this diagram, the oval nodes may raise issues of law and the rectangular nodes may raise issues of fact; in some cases, both an oval and a rectangle may be in issue as a mixed issue of law and fact.
Popple (1996,p. 70-1), explains his directed acyclical graph in terms of his open-texture circles and square leaves:

The circles are parent nodes, representing open-textured concepts; the squares are leaf nodes, representing concepts which are considered to be fully

defined (i.e. answerable by the user). The top level parent node is called the root node.

When it comes to proving a case, law applies not just with the necessity of deduction, but, significantly, with the necessity of extended deduction based on the overlap of common components from different rules. The hierarchy of extended deduction produces the mix of potential issues in a case. Legal ontologies include both the open-textured and factual concepts of the epistemological hierarchy of the rules of law that constitute Major premises for extended deduction; the hierarchy is the epistemological structure that orders the ontology for extended deductive processing of its components by way of application to a client case.

The open texture of some antecedents in law makes extended deduction inevitable. Detailing of antecedents in some rules, with further rules, is the inherent structure of the hierarchy of rule systems in law. Material facts in cases remain the inductive instances of antecedents, as distinct from rules of finer granularity, even if finer rules have only one antecedent. Finer rules are an opportunity to require several antecedents, not just one, and to add a further hierarchy of requirements, not just deductive instances of a legal concept.

### 4.2. Contradictories and uncertainties

To complete the ontology of deductive legal possibilities in a legal expert system, the extended deductive antecedents and consequents that are formalised from black letter law, are expanded by their contradictories and uncertainties, including the contradictories and uncertainties of open-textured ontologies; these additional antecedents and consequents are also structured as rules, further expanding the system of formalised rules that are within the scope of the black letter law. It is possible that a case may occur with the contradictory or uncertainty of an antecedent or consequent that is specified in a rule of law; the effect of this must be clarified, especially where black letter law disjunctions produce alternative rules with the same consequent.

In the legal domain, the contradictories in the ontology of legal possibilities are treated as ontologies, even if they ensure, not some antonym, but the absence of some fact or condition; since rules of law may require the absence of certain antecedents, such absence of an existence is treated as legal ontology. For instance, the absence of rejection of a contractual offer is one of the necessary and sufficient conditions to establish a valid contract. Uncertainties are given legal consequents, according to the rules of burden of proof, so they too are part of the ontology of legal possibilities used by legal experts, pending resolution of uncertainties

in a judgment.
The elements of ontological possibilities in the law are not completed by the deductive contradictories and uncertainties; however, these elements of ontological possibilities are contained in the epistemological structures of rules, as antecedents and/or consequents in an extended deductive structure. Combinatorial explosion of the alternative possible combinations of initial antecedents and consequents, their contradictories and uncertainties, determines the full scope of alternative, consistent, valid deductive legal arguments in the totality of the epistemological system of rules, and all possible case pathways through the full extent of the legal ontology of deductive antecedents and consequents.
Prior analytics, which is one of the six parts of Aristotle's work on logic, the *Organon* (Aristotle, 1952), deals systematically with the formalisation of premises for valid deductive inferences and the extent of necessary logical conclusions in syllogisms; invalid conclusions are also considered within the scheme of invalid inferencing and fallacies. A prior analytics of black letter law is posed in this paper as part of the first step in legal knowledge engineering methodology, namely the acquisition of the expert legal knowledge.
Acquisition of legal expert knowledge must take into account both the black letter law and the further rules to accommodate the contradictories and uncertainties of antededents and consequents, as a matter of logical completeness; these further rules are the implied rules of contradictories and uncertainties that may be relied on by opponents in litigation. Prior analytics may formalise and shape rule hierarchies as continuous Major deductive premises for application by extended deduction to possible cases argued in litigation; the logical extension of the rules of law, and the hierarchies of Major deductive premises, are matters of legal epistemology.
Express law may state a mixture of rules for opposing parties, but extended deductive premises must be streamlined for one party or the other. It may be necessary to use a rule or its contradictory form to complete the streamlining for one side in litigation.

### 4.3. Inductive posits

Further ontological possibilities in regard to selected black letter law arise from the inductive instances which particularise the deductive antecedents and consequents of the system of rules that is within the scope of the black letter law (cf. Popple, 1996, p.68); these are likely to be factual instances of open texture or factual antecedents, their contradictories and their uncertainties. Inductive instances, which are

existential or definitional in nature, may be iterative, or analogous to each other; they may be devised by reference to dictionary definition, synonyms, facts or dicta of precedent cases, expert evidence, common knowledge or common sense. Thus, further epistemological treatment of ontologies in rules, by induction, further expands the possible ontologies for legal argument or for the goal attainment of legal strategies. The instances are induced ontologies, to be applied through rules, and not directly to case facts. In the legal domain, rules of law are enforced, pursuant to the rule of law, not pursuant to a rule of inductive ontologies or a rule of the functional ontologies of Valente (1995). The contingent nature of rules acts as fair warning of enforcement to subjects of the law.

### 4.4. Abductive posits

A final expansion of possible legal ontologies pertains to the ontologies to be found in abductive premises used in legal argument. These ontologies may also expand and contract as further circumstances come to hand and potentialities for legal invention or law-making, are realised. Abductive premises, their contradictories and uncertainties, may provide strong or weak support for rules of law. They are usually reasons for rules; they may be the deeply rooted customs of moral action referred to by Buchler (1961, p.159). In legislation and explanatory memoranda, abductive ontology may be available. Where case facts are brought to rules as the inductive instances of antecedents, or case dicta establish rules or parts of rules, as envisaged by Branting (1991), reasons for rules might also be given, abductively to the decision in the case (cf. Atkinson, Bench-Capon and McBurney, 2005). Abduction may be a meta-ratio for a *ratio decidendi*.

Abductive posits may have their own separate epistemology, some of which might be a *modus ponens* form of deduction in its own context. Historically, inductive and abductive annotations were made to codes of law as glosses; in modern times, margin notes are customary in statutes, but are treated as extraneous to the statutory law. A stratified appearance of the medieval glosses of the Jewish Code of Laws by Maimonides (1550), which might include his Aristotelian commentary, ex facie indicates an abductive epistemology. Figure 2 is a page in this work; other pages have similar but varied stratification. It would be difficult to provide Aristotelian commentary without retaining its logical structure; there may be transcendent *rationes* for *meta-rationes*. The Bologna glosses of the Roman Code of Laws, which began a century prior to Maimonides (1135-1204), include inductive and abductive annotations, confined to the four simple margins around the text; they

are not complex with stratification in reasoning around central ideas, as is Maimonides glossing.
Where induction and abduction are located, by reference to the components of extended deduction, their strands of annotative reasoning should be kept separate, like glosses, from the strands of extended deductive reasoning. Otherwise the sequence of reasoning may appear non-monotonic. Ontologies that are deemed by law-making authorities to apply to cases by necessity, should not be confused with abductive ontologies that play a different role in legal argument.

## 5. Semantic invalidity in logic

Ontological posits and informal truth tables of law-making authorities solve the problem of semantic invalidities in logical form. Semantic invalidition of a *modus ponens* syllogism which is used in applying law to a case, is described by Waller (1995, p.170), in his first year law text, in the following way:

> Every sentence containing six words is true.
>
> This sentence contains six words.
>
> Therefore it is true.

Waller (1995, p.170-1) also points out that lawyers prefer conditional propositions or propositional calculus to predicate logic, which are interchangeable forms of deduction, as there is less to assert as true in the Major premise:

In any area where people use deduction they may employ one of two kinds of syllogism. They may begin, if the task is of a theoretical kind, by using the word "all". The ancient example is:

1. All men are mortal.
2. Socrates is a man.
3. Therefore Socrates is mortal.

*This method is simple. If the first two propositions are correct, the conclusion is obvious. The first proposition is called the major premise, the second the minor premise. But, of course, you may want proof of either premise. "Is it true that all men are mortal? It is true that Socrates is a man?"In this example long experience shows plainly that both are correct. In any event, the logician would answer that he or she*

*Figure 2.* Page from Maimonides, M., Mishneh Thorah, (c. 1180): Annotation of Jewish Code of Laws with Aristotle's works. (D. Pizzighettone and A. Dayyan (eds), Venice, C. Adelkind for M.A.Giustiniani, 1550)

*is merely making assumption. Consequently then the answer is true as a theory....*
*Lawyers, and most other thinkers, prefer in practice to employ the second kind – the hypothetical deduction. That begins with "if" instead of "all". For example there is this syllogism:*

1. If a person deliberately hits another with a cricket bat that person has committed the crimes of assault and battery.

2. Jane deliberately hit Bill with a cricket bat.

3. Therefore Jane is guilty of these crimes.

The hypothetical method is often superior for use because it does not say "all". It is another kind of assumption, not so hard to prove and likely to be correct....
So "If P then Q" is relevant as a guide – tautological though it may be. It remains the best and most common kind of inference for courts though they rarely use the actual terms: syllogism, major or minor premises. But they do constantly say, "If that is the law, then it follows that the plaintiff was entitled" or "the defendant is guilty".

Of course Waller (1995, p.168), also recognised that precedent cases are inductive examples, even in the formulation of new antecedents or rules; some induction is determined by analogy and some by common sense or authoritative iteration. He also explored the logic used by lawyers that is outside the realms of deduction and induction, especially in keeping rules consistent and providing for new cases. A systems view of legal logic is maintained by Waller (1995, p.181), by reference to Wisdom (1973, p.195):

Professor Wisdom made a penetrating remark: he proposed that lawyers' arguments "are like the legs of a chair, not like links in a chain". Common sense, history, analogy and so on, support one another if the issue is at all complex. This is the type of logic that the ancients knew well and valued highly under the name of rhetoric. It was extensively used in medieval times for practical judgments. Only in the last three years did logic – in a vain effort to make thinking mechanical and perfect – come to include only formal logic. But throughout these centuries lawyers have gone ahead using rhetorical reasoning with excellent results. ("Rhetorical" here is not to be confused with fulsome oratory, unfair appeals to emotions and extravagant language.)

## 6. Limits of logical extensions of legal syllogisms

It is not logically valid to extend a rule of law to its adversarial form. Only the establishment of a contradictory ontology can provide the basis for an opponent's argument in litigation. Thus if there is a rule 'if a then c', it is not thereby logically valid to assume 'if not a then not c'. There may be ways other than a to establish c. However, if the rule 'if not a then not c' is established ontologically, then there is an adversarial provision that is part of the ontology of legal possibilities. The adversarial contradictory will be established from the meaning of the law-maker's language in laying down the express rule; if the antecedents are referred to in terms that they must be established, then this will produce an adversarial contradictory.

Of course, law-making authorities may not lay down the adversarial contradictory rule; instead they may lay down a disjunction: 'if not a then c'. A disjunction of mutually exclusive contradictory antecedents occurs with some qualification in the Australian Spam Act 2004; a message which is not a commercial electronic message is not prohibited and a commercial electronic message which complies with certain conditions also is not prohibited. In legal epistemology, 'not a implies not c' may have ontological validity, even if it does not have logical validity as a derivation from 'a implies c'; epistemological rules may override the meta-rules of logic. Also, contradictories may be common points for both adversaries; it can not be assumed that the contradictory of one party's points is the same as a point for the opponent's case. Authoritative legal ontologies must be considered for each case.
Even though the ontology of the adversarial contradictory may be implied, and extended deduction justifies forward chaining in the direction indicated by the inference arrow that represents 'then' or 'implies' in the conditional proposition, this does not authorise backward inference; the conditional proposition that is a rule of law is only a material implication or an ontologic posit equivalent to the reversed C of Peano, if the law-maker designates it as such, and usually this does not happen unless there is a legal presumption. Prima facie, a consequent in a rule of law does not logically establish its antecedents; antecedents must be established, directly or indirectly, by evidence of material facts in order to establish their consequent.

## 7. Knowledge representation and ontology

In information science, ontology is used as a domain epistemology to acquire vocabulary with meaning mechanisms; Figure 3 is an embellished Porphery tree (2005) which is an epistemological structure that was devised by Porphery (c.232-304) to represent Aristotle's ontology of substance. It also locates inductive instances and the pattern of a taxonomy. The tree categorisation of ontology is useful in information science, as the meaning mechanisms of a Porphery's tree representation founds the epistemology of predicate logic.

It was suggested by Valente that the modelling of functional ontologies would remedy the epistemological shortcomings of earlier legal knowledge engineering to be found in logic systems. Inevitable ontological 'commitments' embedded in logic formalisms were identified by Valente but he did not go on to find the ontology of legal possibilities implied by the express conditional propositions of law:

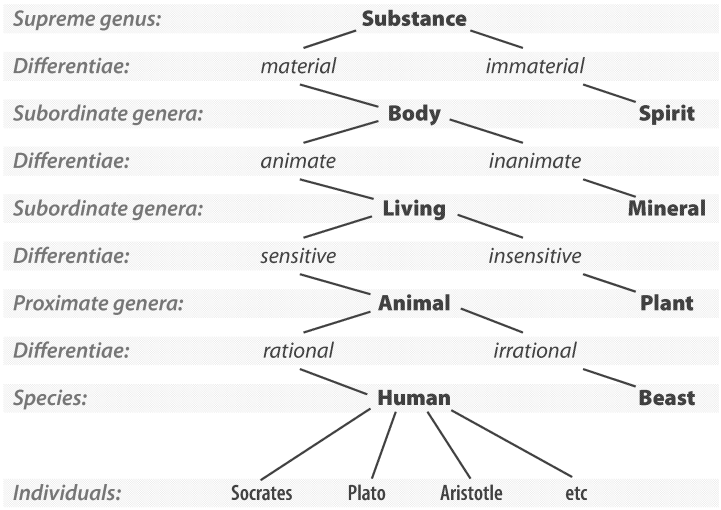



*Figure 3.* Horrocks' Porphyry tree (2005) See http://www.epsg.org.uk/pub/needham2005/

With regard to their role as a representation tool for legal knowledge, the basic problem is that most of the proposed formalisms (which means basically deontic logics) fail to keep track of the epistemological aspects they necessarily involve, i.e. of the (inevitable) ontological commitments embedded in the formalism. (Valente, 1995, p.17).

Certainly Aristotle's ontology of substance intended to capture all possible substance, but the distinction between contradictories that are non-existences and contradictories that are antonyms has not been considered in regard to the Porphery tree epistemology and in the translation of predicate calculus to propositional calculus as suggested by

Waller. Use of an ontology requires logic; the use structure of ontology is required for selection of appropriate logic.

As a solution to the epistemological shortcomings of rule-based and case-based systems, Valente (1995) added ontology modelling to the repertoire of legal knowledge engineering methodology. He recognised that legal ontologies could be extracted from black letter law and modelled in various ways as functional ontologies for legal expert systems. His modelling of extracted ontologies was to be in accordance with models of legal practice ontologies that focussed on tasks, goals and methods. However, he did not consider that such modelling and models were, ipso facto, epistemological; nor did he consider the requirement that the functional ontology be in accordance with sound legal epistemology that had to be found in the legal practice ontologies. Sound legal epistemology plays a role in determining the ontology of legal possibilities for logical processing.

In his criticism of Valente's work, Aikenhead (1996) referred to the oft-quoted point made by Susskind (1987, p.20), in regard to legal expert systems:

> It is beyond argument, however, that all expert systems must conform to some jurisprudential theory because all expert systems in law necessarily make assumptions about the nature of law and legal reasoning.

However, Valente's work filled an important gap in legal knowledge engineering methodology that was not appreciated by Aikenhead. The extraction of a legal ontology and its remodelling, explains the process of formalising rules of law as Major deductive premises; it is the process of prior analytics (cf. Aristotle, 1952 c 330BC) that is required if a deductive antecedent or consequent in a rule of law is varied in accordance with black letter law, for the sake of adversarial completeness. Valente illuminates precisely a step in the reasoning of legal practitioners, not before exposed.

Shannon and Golshani (1988) defined deep models as ones that model meaning and not just words. The meaning of law is adversarially complete with the ontology of legal possibilities and legal potentialities, with their implicit logic.

## 8. Conclusion

Legal knowledge engineering requires the development of its own Jurisprudence of Legal Knowledge Engineering. The use of ontology and epistemology in philosophy is a rich source for the development of legal knowledge engineering jurisprudence. A deep model of legal expertise for legal knowledge engineering may enhance legal practice and further

develop the jurisprudence of legal knowledge engineering.
In order to program an epistemologically sound legal expert system, a legal knowledge engineer must acquire from a legal expert, a knowledge of the substance of express ontological posits that are deductive ontology, inductive ontology and abductive ontology, and then derive from these posits the ontology of legal possibilities. The process of derivation is an epistemological process, like the formulation of a truth table in logic; it provides for possible cases within the scope of the express black letter law and legal information, and the valid legal arguments that apply to those cases.

## References

Aikenhead, M. (1996): Book review, A. Valente, Legal Knowledge Engineering, *International Review of Law, Computers and Technology*, Vol. 10, Issue 2, p351, Oct.

Aristotle (1952): Organon, in R. M. Hutchins (ed.) *Great books of the Western World*, Vol 8, Encyclopaedia Britannica Inc., Chicago, USA.

Ashley, K.D. (1990): *Modeling Legal Argument: Reasoning with Cases and Hypotheticals*, MIT Press, Cambridge, MA.

Atkinson, K., Bench-Capon, T. and McBurney, P. (2005): Arguing about Cases as Practical Reasoning, in *Proceedings of the Tenth International Conference on Artificial Intelligence and Law*, ACM, New York, USA.

Bacon, F., Novum Organum (1620, 1952), in R. M. Hutchins (ed.) *Great books of the Western World*, Vol 30, Encyclopaedia Britannica Inc., Chicago, USA.

Branting, L. K. (1991): Reasoning with portions of precedents, in *Proceedings of the Third International Conference on Artificial Intelligence and Law*, Association for Computing Machinery, New York, USA.

Buchler, J. (1961): *The Concept of Method*, Columbia University Press, New York, USA.

Foucault, M. (1969, 1972): *The archaeology of knowledge*, English translation by A.M. Sheridan-Smith, Pantheon Books, New York, USA.

Gray, P. N. (2005): eGanges: Epistemology and Case Reasoning, in P.E. Dunne and T. Bench-Capon (eds), *Argumentation in Artificial Intelligence and Law*, Wolf Legal Publishers, Nijmegen, Netherlands.

Gray, P.N. and Gray, X. (2003): A Map-Based Expert-Friendly Shell, in D. Bourcier (ed.), *Legal Knowledge and Information Systems*, IOS Press, Amsterdam, Netherlands.

Horrocks, I. (2005): Ontologies and the Semantic Web, http://www.epsg.org.uk/pub/needham2005/Horrocks_needham2005.pdf

Kant, I. (1788, 1955): Critique of Practical Reason, in *Great Books of the Western World*, Vol 42, translated by J. M. D. Meiklejohn, T. K. Abbott and J. C. Meredith, Encyclopaedia Britannica Inc., Chicago, USA.

Kuhn, T.S. (1970): *The Structure of Scientific Revolutions*, University of Chicago Press, Chicago, USA.

Leibniz, G.W. Von, (1714, 1992): *Monadology*, translated and edited by N. Rescher, Routledge, London, England.

Maimonides, M. (originally c. 1180, 1550): *Mishneh Thorah: Annotation of Jewish Code of Laws*, D. Pizzighettone and A. Dayyan (eds), C. Adelkind for M.A.Giustiniani, Venice, Italy.

Popple, J. (1996): *A Pragmatic Legal Expert System*, Dartmouth Publishing Company Limited, Aldershot, England.

Rissland, E. L.(1985): Argument Moves and Hypotheticals, in C. Walter, (ed.), *Computing Power and Legal Reasoning*, West Publishing Co, St Paul, USA.

Shannon, D. T., and Golshani, F. (1988): On the automation of legal reasoning, *Jurimetrics Journal*, Vol 28, no. 3, p305, Spring.

Susskind, R.E. (1987): *Expert Systems in Law: A Jurisprudential Inquiry*, Clarendon Press, Oxford, England.

Valente, A. (1995): *Legal Knowledge Engineering: A Modelling Approach*, IOS Press, Amsterdam, Netherlands.

Waller, L. (1995): *An Introduction to Law*, LBC Information Services, Sydney, Australia.

Wisdom, J. (1951, 1973): Gods, in A. Flew, (ed.), *Logic and Language*, Basil Blackwell, Oxford, England.

Wittgenstein, L. (1918, 1922): *Tractatus Logico – Philosophicus*, Routledge and Kegan Paul, London, England., Deontic Logic, Mind, Vol 60, p 1, 1951

# The Legal-RDF Ontology. A Generic Model for Legal Documents

John McClure
*Legal-RDF.org / Hypergrove Engineering*
*jmcclure@hypergrove.com*

**Abstract.** Legal-RDF.org[1] publishes a practical ontology that models both the layout and content of a document and metadata about the document; these have been built using data models implicit within the HTML, XSL, and Dublin Core dialects. Directed Acyclic Graphs (DAGs) form the foundation of all models within the ontology, that is, *DAGNode* and *DAGModel* are the base classes for all other ontology classes, which include a restatement of RDF and OWL classes and properties as well as basic Kellog parts-of-speech. The ontology also represents an explicit semantic model used during its classifications: concrete classes are categorized as some element of a dramatic production, that is, as a subclass of *Actor*, *Role*, *Scene*, *Prop*, *Theme*, or *Drama*; this can be helpful during analyses of semantic perspective and context associated with resource definitions and attribute values. The Legal-RDF ontology distinguishes between predicate verbs and predicate nouns in its models of a Statement to yield an intuitively appealing vocabulary that segregates attributes as past, present, future, or conditional, information. To facilitate development of generic tools, all data and object properties defined in the ontology's models are categorized as a subproperty of one of the 15 Dublin Core properties; provenance data, with emphasis on an asOf timestamp, may be recorded for any attribute of a resource. Legal-RDF's numeric properties derive from the ISO Systeme Internationale measurement systems; algebraic properties derive from XML Schema datatypes; language and currency designations are based upon relevant ISO standards; and time-zone designations are based on a review of local and regional standards (with some modifications necessary to eliminate collisions between the names of these properties and ISO standards). In addition to classes that represent quantities, classes are included that represent qualities that may be used to subtype or otherwise characterize instances.

**Keywords:** Aspect-oriented programming, Dublin Core, Kellog Grammar.

## 1. Status of the Legal-RDF Ontology

Version 2 of the Legal-RDF ontology – which this paper describes – is being documented in a Wiki[2] hosted by LexML.org[3] to encourage the participation of an interested community during its development. The first version of the ontology, at the Legal-RDF.org website, is being

[1] `http://www.hypergrove.com/legalrdf.org/index.html`

[2] The wiki is located at `http://aufderheide.info/lexmlwiki/index.php?title=Legal-RDF\_Ontologies`

[3] http://www.lexml.org/

improved in Version 2 with the adoption of a better class- and property-naming guideline; with the refactoring of base/derived classes; and with the definition of more complete data models.

## 2. Naming Conventions

Facilitating the construction of *dotted-names* is the primary objective of Legal-RDF class and property naming guidelines. A dotted-name defines a path in a Directed Acyclic Graph, e.g., *Person.FullName.FirstName*, intentionally aligned with the ECMA-242 standard for attribute value references. In effect, the Legal-RDF ontology aims to define an object model that is reasonable in the context of software written in an ECMA language, e.g., Javascript, C#, and Eiffel.

Text properties are in lower-case, e.g., *Person.FullName.FirstName.eng* is a reference to a string of English text while *Person.FullName.FirstName* references an object (i.e., a resource) for which the *eng* text property may be present.

RDF triples are accommodated by a defaulting mechanism. When no predicate is specified, the *has* predicate is implied, e.g., the dotted-name above transforms to *Person.has.FullName.has.FirstName.eng*, allowing the use of other predicate verbs such as *Person.willHave.FullName* to describe, perhaps, a bride.

A consequence of this approach is that Legal-RDF separately defines predicate verbs and predicate nouns, specifically eschewing the RDF community practice that concatenates these as single property names, e.g., "hasName". This yields a naming system that is historically more familiar to the software industry, while maintenance economies are had as new predicates are defined over time. [4]

All Legal-RDF classes are named by at least two words so that, when the class is the range of an object property, a single word can be used for the property name.

## 3. The CoreResource[5]Class

All classes in the ontology derive from the *CoreResource* class whose function is to allow Dublin Core attributes to be associated with any resource. This class demonstrates how the Legal-RDF ontology incorporates the principles of aspect-oriented programming. The ISO Dublin

[4] An "RDF quint" composed of subject, predicate verb, predicate noun, object, and node identifier.

[5] http://aufderheide.info/lexmlwiki/index.php?title=Legal-RDF:CoreResource

Core properties are not properties of the *CoreResource* class; instead the Dublin Core properties are inherited from superclasses representing qualities had by instances of *CoreResource*. To the maximum extent possible, all properties in the Legal-RDF ontology are bundled as quality classes, enabling (a) creation of aspect-oriented applications (b) adoption of other namespaces (c) clearer simpler class hierarchies and (d) comparative quality analyses.

Table I. *CoreResource* Inherited Properties

| *CoreResource* **Superclass**[6] | **DublinCore Property** | *CoreResource* **Superclass** | **DublinCore Property** |
|---|---|---|---|
| CategorizableThing | subject | IdentifiableThing | identifier |
| ClassifiableThing | type | ManagableThing | rights |
| CreatableThing | creator | NamableThing | title |
| DerivableThing | source | PublishableThing | publisher |
| DescribableThing | description | RelatableThing | relation |
| EnhanceableThing | contributor | SchedulableThing | date |
| ExpressibleThing | language | ScopableThing | coverage |
| | | StylableThing | format |

The inherited properties (e.g., subject) listed in Table 1 are text-properties. Paired with each superclass, e.g., *ClassifiableThing*, is a subclass representing the **state** of a resource with respect to the quality, e.g., *ClassifiedThing* is a proper subclass of *ClassifiableThing* in that every thing that is deemed 'classified' in some way is, by absolute semantic necessity, a 'classifiable' thing. The *ClassifiedThing* class includes a *Type* property whose range is *ClassNode* (read: *owl:Class*). An instance of a *CoreResource* is therefore not a *ClassifiedThing* until the text within a *type* attribute has been correlated with a class defined by the ontology.
The *CoreResource* class has no object properties and only two text properties: *asOf,* a timestamp to record when a resource was last updated; and *rdf*, a URI for retrieving an RDF representation of the resource.

## 4. The *DAGNode*[7] and *DAGModel*[8] Classes

[6] These classes are subclasses of the *CapabilityFacet* class, and are each paired with a subclass of the *StateFacet* class – both are subclasses of a *FacetNode* class, a subclass of the *DAGNode* class, which is itself a subclass of the *CoreResource* class. The circularity of this hierarchy is an open issue.
[7] http://aufderheide.info/lexmlwiki/index.php?title=Legal-RDF:DAGNode
[8] http://aufderheide.info/lexmlwiki/index.php?title=Legal-RDF:DAGModel

The *rdf:Description* class is the primary superclass for the *DAGModel* class so as to clearly delineate its role as a representation of a Directed Acyclic Graph (DAG). DAGs are composed of nodes and arcs; arcs are classified as either terminating with a node or with a literal text string.

*DAGModel* and *DAGNode* derive from *CoreResource* and from classes that correspond to the Representational State Transfer (REST) protocol:

(a) the *DeletableThing* class, with its *DeletedThing* subclass, correspond to a potential or actual Delete operation;

(b) the *RecordableThing/RecordedThing* classes correspond to a Put;

(c) the *RetrievableThing/RetrievedThing* classes correspond to a Get;

(d) the *UpdatableThing/UpdatedThing* classes correspond to an Update.

All provenance data about creation, retrievals, updates, and deletions of a resource or resource attribute are captured by instances of these classes.
*DAGModel* subclasses are established to correspond with the types of Unified Modeling Language (UML) diagrams. An 'ontology' for example corresponds to a *ClassModel*. Second, subclasses exist for document types; page layouts; and for specifying 'one-off' resource instance models.

Table II. *DAGModel* Class Relations

| **Properties** | **Range** | **Category** | **Superclasses** | **Subclasses** |
|---|---|---|---|---|
| Arc | StatementNode | Coverage | CoreResource | ClassModel |
| LiteralArc | StatementNode | Arc | rdf:Description | DocumentModel |
| Node | DAGNode | Coverage | DeletableThing | EventModel |
| ObjectArc | StatementNode | Arc | RecordableThing | InstanceModel |
| | | | RetrievableThing | PageModel |
| | | | UpdatableThing | ProcessModel |

*DAGNode* subclasses correspond to classes defined by the RDF, RDF Schema, and OWL specifications. Beyond these, two additional subclasses are defined, *ContextNode* and *FacetNode*.

Table III. *DAGNode* Class Relations

| **Properties** | **Range** | **Category** | **Superclasses** | **Subclasses** |
|---|---|---|---|---|
| Arc | StatementNode | Coverage | CoreResource | ClassNode |
| Model | DAGModel | Source | CountableThing | CollectionNode |
| Slot | FacetNode | Description | DeletableThing | ContextNode |
| Template | InstanceModel | Format | RecordableThing | FacetNode |
| | | | RetrievableThing | LiteralNode |
| | | | UpdatableThing | PropertyNode |
| | | | | StatementNode |

## 5. The *LiteralNode*[9] Class

The ontology defines the *LiteralNode* class to represent words, sounds, figures, images, or video clips that can be rendered for presentation. Properties of the *LiteralNode* class allow a concept represented by an instance to be simultaneously expressed in words, sounds, figures, images, and or video. The *Content* super-property is defined for the *ExpressedThing* class, itself a subproperty of the *ExpressedThing* class' *Language* property.

Table IV. *LiteralNode* Class Relations

| **Properties** | **Range** | **Category** | **Superclasses** | **Subclasses** |
|---|---|---|---|---|
| Audio | AudioNode | Content | DAGNode | AudioNode |
| Graphic | GraphicModel | Content | AnalyzableThing | GraphicModel |
| Image | ImageNode | Content | ExpressibleThing | ImageNode |
| Text | TextNode | Content | HidableThing | TextNode |
| Video | VideoNode | Content | | VideoNode |

All document text content and nearly all text properties associated with any resource are represented using the *TextNode* subclass. The *TextNode* class has two groups of subclasses: (a) ones relating to functional types of document text, e.g., strings of text, text tokens, symbolic text, etc. and (b) the union of classes that represent upper-, lower-, and mixed-case text.

*NumericText*, a subclass of *TextNode*, uses the subclass, *RealNumber*, to represent all real numbers found in documents.

The *RealNumber* class defines the *float* text property, which corresponds to the *float* attribute defined by XML Schema Datatypes. Legal-RDF's *float* property is a subproperty of the *value* text property

[9] http://aufderheide.info/lexmlwiki/index.php?title=Legal-RDF:LiteralNode

Table V. *TextNode* Class Relations

| **Properties** | **Range** | **Category** | **Superclasses** | **Subclasses** |
|---|---|---|---|---|
| Language | LanguageFacet | Type | rdf:Literal | NumericText |
| Length | IntegerNumber | Description | LiteralNode | SemanticText |
| text | xmls:string | (none) | LinkableThing | SymbolicText |
| | | | PaddableThing | TextBlock |
| | | | TintableThing | TextString |
| | | | TypesettableThing | TextToken |

Table VI. *NumericText* Class Relations

| **Properties** | **Range** | **Category** | **Superclasses** | **Subclasses** |
|---|---|---|---|---|
| (none) | | | TextNode | ComplexNumber |
| | | | ComparableThing | ImaginaryNumber |
| | | | ConvertableThing | RealNumber |
| | | | EstimableThing | NonNegativeNumber |
| | | | QuantifiableThing | NonPositiveNumber |
| | | | RoundableThing | StatisticalNumber |
| | | | ScalableThing | |

for the *QuantifiedThing* class. In other words, when a numeric value is provided as an attribute, the attribute is then deemed to have entered the 'quantified' state.

Table VII. Systeme Internationale Classes

| **Subclass** | ***QuantityFacet* Subclass' Subclasses** |
|---|---|
| Angle-Measure | ArcDegreeQuantity, ArcMinuteQuantity, ArcSecondQuantity, RadianQuantity, SteradianQuantity |
| Area-Measure | AcreQuantity, AreQuantity, CentiareQuantity, ColumnInchQuantity, CommercialAcreQunatity, HectareQuantity, SquareCentimeterQuantity, SquareQuantity, SquareFootQuantity, SquareInchQuantity, SquareYardQuantity, SquareKilometerQuantity, SquareMeterQuantity, SquareMileQuantity, SquareMillimeterQuantity |
| Capacity-Measure | BarrelQuantity, CentiliterQuantity, CubicCentimeter, LiterQuantity, MilliliterQuantity, CubicKilometer, CubicMeterQuantity, CubicMillimeter |
| Density-Measure | RadQuantity, TeslaQuantity, WeberQuantity |
| Distance-Measure | AngstromQuantity, CaliberQuantity, CentimeterQuantity, DecimeterQuantity, EmQuantity, FootQuantity, FortyFootEquivalentQuantity, FurlongQuantity, GaugeQuantity, InchQuantity, KilometerQuantity, MeterQuantity, MileQuantity, MillimeterQuantity, NauticalMileQuantity, PointQuantity |
| Dry-Measure | BaleQuantity, BoardFootQuantity, BundleQuantity, BushelQuantity, CartonQuantity, CordQuantity, CubicFootQuantity, CubicInchQuantity, CubicMileQuantity, CubicYardQuantity, DozenQuantity, DryQuartQuntity |
| Electrical-Measure | AmpereQuantity, CoulombQuantity, FaradQuantity, GigawattHourQuantity, GigawattQuantity, JouleQuantity, MegawattHourQuantity, MegawattQuantity, WattQuantity, MilliwattHourQuantity, MilliwattQuantity, OhmQuantity, SiemansQuantity, TerawattQuantity, VoltQuantity |
| Energy-Measure | BritishThermalUnitQuantity, CalorieQuantity, GrayQuantity, HorsepowerQuantity, KilopascalQuantity, NewtonQuantity, PoundsPerSquareFootQuantity, PoundsPerSquareInchQuantity |
| Frequency-Measure | CyclesPerMinuteQuantity, CyclesPerSecondQuantity, GigahertzQuantity, HertzQuantity, KilohertzQuantity, MegahertzQuantity |
| Light-Measure | CandelaQuantity, LumenQuantity |
| Medical-Measure | InternationalUnitQuantity, KatalQuantity, SievertQuantity |
| Pressure-Measure | BarQuantity, BecquerelQuantity, DecibarQuantity, DyneQuantity, HectopascalQuantity, KilopascalQuantity, PascalQuantity |
| Sound-Measure | DecibelQuantity, SabinQuantity |
| Speed-Measure | FeetPerMinuteQuantity, FeetPerSecondQuantity, KnotQuantity, KilometersPerHourQuantity, MetersPerSecondQuantity, MilesPerHourQty |
| Temperature-Measure | CelsiusQuantity, FahrenheitQuantity, KelvinQuantity, ThermQuantity |
| Time-Measure | DayQuantity, DecadeQuantity, HourQuantity, MinuteQuantity, MonthQty, QuarterQuantity, SecondQuantity, WeekQuantity, YearQuantity |
| Velocity-Measure | CubicCentimetersPerSecondQuantity, CubicFeetPerMinuteQuantity, CubicMetersPerHourQuantity, CubicMetersPerSecondQuantity, GallonsPerMinuteQuantity |
| Volume-Measure | AcreFootQuantity, CupQuantity, FluidOunceQuantity, GallonQuantity, ImperialGallonQuantity, PintQuantity, QuartQuantity |
| Weight-Measure | AssayTonQuantity, CaratQuantity, CentigramQuantity, GrainQuantity, GramQuantity, KilogramQuantity, MilligramQuantity, MoleQuantity, OunceQuantity, PoundQuantity, PoundFootQuantity, StoneQuantity, TonQuantity, TroyOunceQuantity, TroyPoundQuantity, TonneQuantity |

The *RealNumber* class illustrates another feature of the ontology. It is a superclass of the *QuantityFacet* class, a subclass of the *FacetNode* class (see Table XVI). *QuantityFacet* defines subclasses that correspond to all measures standardized by the Systeme Internationale (SI). Each has a text property – whose name matches SI's standard abbreviation for the measurement – that is a subproperty of the *QuantifiedThing* class' *value* property, e.g., *m2* is the text property defined for the *SquareMeterQuantity* class, where 'm2' is the SI name for 'square meter' measurements.

Finally in the area of numerics, classes exist for each currency recognized by the ISO. For instance, the *UnitedStatesDollarAmount* class derives

Table VIII. *SemanticText* Class Relations

| Properties | Range | Category | Superclasses | Subclasses |
|---|---|---|---|---|
| Abbreviation | AbbreviationToken | Title | TextNode | TextPhrase |
| Acronym | AcronymToken | Title | | TextWord |

from the *CurrencyAmount* class, which derives from *DecimalAmount*, which derives from *ProperFraction*, which derives from *FractionalNumber*, deriving from the *RealNumber* class identified in Table VI as a subclass of *NumericText*.

*CurrencyAmount* is the superclass for *EconomicAmount* and *EconomicValue*. These classes measure certain currency flows (*CapitalAmount, ChargeAmount, DeductionAmount, DiscountAmount, DueAmount, Expense-Amount, IncomeAmount, LiabilityAmount, NetAmount, NetWorthAmount, PriceAmount, ProfitAmount, RevenueAmount, WealthAmount*). A number of these subclasses are decomposed by economic factors of production, e.g., C*hargeAmount* has the subclasses *CapitalCharge, LaborCharge, Material-Charge, ProductCharge,* and *ServiceCharge*.

The *SemanticText* class is notable because it unions more than 140 classes that represent each of the languages defined by ISO-639, replicating functionality of the *xml:lang* attribute. For example, the *EnglishText* class defines the *eng* property whose super-property is the *text* property defined for the *TextNode* class.

The *TextPhrase* class segues to Legal-RDF's linguistic model, as it is the superclass for the *TextClause, AdjectivePhrase, AdverbPhrase, Noun-Phrase, Verb-Phrase, PrepositionPhrase,* and *InterjectionPhrase* classes. *TextClause* is the superclass for the *TextSentence* class, which is the superclass for two classes, *CompoundSentence* and *ComplexSentence*.

Table IX. *TextClause* Class Relations

| Properties | Range | Category | Superclasses | Subclasses |
|---|---|---|---|---|
| Adjective | AdjectivePhrase | Content | TextPhrase | TextSentence |
| Nominative | NounPhrase | Content | | |
| Object | NounPhrase | Content | | |
| Predicate | TextClause | Content | | |
| Subject | NounPhrase | Content | | |
| Verb | VerbPhrase | Content | | |
| DirectObject | NounPhrase | Object | | |
| IndirectObject | NounPhrase | Object | | |
| Punctuation | PunctuationMark | Format | | |

The *TextBlock* class represents a *TextNode* that is visually distinct from a simple string of text characters, laid out in a rectangular fashion, and is the gateway to block-elements defined by the XHTML 2.0 and XSL dialects.

Table X. *TextBlock* Class Relations

| **Properties** | **Range** | **Category** | **Superclasses** | **Subclasses** |
|---|---|---|---|---|
| Body | TextBody | Source | TextNode | TextObject |
| Statement | StatementNode | Description | ParsableThing | TextPage |
| Watermark | ImageNode | Format | RectangularThing | |

In the model above, a *TextBlock* is part of a *TextBody*; can have an *ImageNode* specified as a *Watermark*; and may have multiple *StatementNode* instances to describe the contents of the *TextBlock*. Two subclasses are defined, *TextObject* and *TextPage,* whose qualities and properties differ.

Table XI. *TextObject* Class Relations

| **Properties** | **Range** | **Category** | **Superclasses** | **Subclasses** |
|---|---|---|---|---|
| Page | TextPage | Source | TextBlock | TextBody |
| | | | AnnotatableThing | TextColumn |
| | | | ApprovableThing | TextDivision |
| | | | CitableThing | TextHeading |
| | | | DraftableThing | TextImage |
| | | | IllustratableThing | TextIndex |
| | | | PositionableThing | TextLine |
| | | | PrintableThing | TextList |
| | | | ReviewableThing | TextParagraph |
| | | | TranslatableThing | TextQuote |
| | | | VersionableThing | TextRow |
| | | | ViewableThing | TextSection |
| | | | | TextTable |

The class model in Table XI allows a *TextObject* instance to be located on zero or more pages, and its subclasses reflect its coverage of XHTML 2.0 elements. Its qualities indicate typical document actions are permitted, that is, *TextObject* instances can be annotated, approved, cited, drafted, illustrated, positioned, printed, reviewed, translated, versioned, or viewed.

Properties in the *TextPage* class model (Table XII) indicate six layout areas can be formatted. Around the *BodyArea* may be arrayed a *BannerArea*, *FooterArea*, *HeaderArea*, *SidebarArea*, and *SignatureArea*.

Each area has an associated reference to the text, image, sound, figure, or video content to be placed into the area. This "page model" can be customized for a particular document type by a reference to a *PageModel* (a subclass of the *DAGModel* class), which has names and positioning of custom areas that can be displayed in a user. Finally, note that the pagination-specific Cascading Stylesheet (CSS) *size* text attribute can be specified for a *TextPage*, as part of the CSS-2 support provided.

Table XII. *TextPage* Class Relations

| **Properties** | **Range** | **Category** | **Superclasses** | **Subclasses** |
|---|---|---|---|---|
| Banner | LiteralNode | Content | TextBlock | BlankPage |
| BannerArea | RectangularThing | LayoutArea | FoliatableThing | TitlePage |
| Body | LiteralNode | Content | ViewableThing | |
| BodyArea | RectangularThing | LayoutArea | | |
| DotPage | TextPage | Relation | | |
| Footer | LiteralNode | Content | | |
| FooterArea | RectangularThing | LayoutArea | | |
| Header | LiteralNode | Content | | |
| HeaderArea | RectangularThing | LayoutArea | | |
| LayoutArea | PositionableThing | Format | | |
| MappingModel | PageModel | Format | | |
| Sheet | PaperSheet | Source | | |
| Sidebar | LiteralNode | Content | | |
| SidebarArea | RectangularThing | LayoutArea | | |
| Signature | LiteralNode | Content | | |
| SignatureArea | RectangularThing | LayoutArea | | |
| NextPage | TextPage | Relation | | |
| PreviousPage | TextPage | Relation | | |
| size | xmls:string | style | | |

The *TextBody* class is equivalent to HTML's *body* element and allows one to specify the default header, footer, and banner content to appear on pages in the document when it is formatted. The model has properties for front- and rear-matter in the document, and for other functional document parts (colophon, notes, bibliography, etc) not allocated to a quality class.

The qualities associated with *TextBody* instances enable specification of typical functional document parts, including its attachments; its tables of contents, of figures, of tables, and of authorities; its internal divisions and sections; its cover pages; and its introductory and conclusion material.

Table XIII. *TextBody* Class Relations

| Properties | Range | Category | Superclasses | Subclasses |
|---|---|---|---|---|
| Banner | LiteralNode | Relation | TextObject | (none) |
| Bibliography | TextIndex | Source | OntologicalThing | |
| Colophon | LiteralNode | Description | AppendableThing | |
| EndNote | LiteralNode | RearMatter | AttachableThing | |
| Epilogue | LiteralNode | RearMatter | ConcludableThing | |
| Footer | LiteralNode | Relation | IndexableThing | |
| FrontMatter | LiteralNode | Content | IntroducibleThing | |
| Header | LiteralNode | Relation | PaginatableThing | |
| Paragraph | TextParagraph | Content | PrintableThing | |
| RearMatter | LiteralNode | Content | PrependableThing | |
| SubTitle | TextHeading | SecondaryTitle | StaplableThing | |
| | | | SubDivisableThing | |
| | | | SubSectionableThing | |

## 6. The *FacetNode*[10] Class

Document pagination can demonstrate how quality classes play a key role in the specification (and validation) of an instance model. To begin, a *TextBody* is a *PaginatableThing*, that is, its content can be formatted across one or more pages. When formatting occurs, the *TextBody* instance is assigned these *Page* attributes, each referencing a *TextPage* instance; only then does the *TextBody* instance enter the 'state' of being a *PaginatedThing*. A *TextBody* thus has the capacity to be paginated, as is so indicated by its superclass *PaginatableThing*; it is only after pagination that it is explicitly or deducibly a *PaginatedThing*.

Table XIV. *PaginatableThing* Class Relations

| Properties | Range | Category | Superclasses | Subclasses |
|---|---|---|---|---|
| PaginationPlan | PredictiveStatement | Plan | CapabilityFacet | PaginatedThing |
| PaginationPolicy | RequirementStatement | Policy | | |

The classes above derive from *FacetNode*, a subclass of *DAGNode* (see Table III). In the Legal-RDF ontology, a facet is "an instance of an attribute value; a named value or relationship". Five types of resource facet are identified:

(a) its capabilities, e.g., "the resource can be deleted"

[10] http://aufderheide.info/lexmlwiki/index.php?title=Legal-RDF:FacetNode

Table XV. *PaginatedThing* Class Relations

| **Properties** | **Range** | **Category** | **Superclasses** | **Subclasses** |
|---|---|---|---|---|
| Pagination | PaginationEvent | EntryEvent | PaginatableThing | (none) |
| Page | TextPage | Format | StateFacet | |

(b) its existence, e.g., "the resource is an American resource"

(c) its qualities, e.g., "the resource is expressed in English"

(d) its numeric quantities, e.g., "the resource has five attributes" and

(e) its states of existence, e.g., "the resource was validated".

Table XVI. *FacetNode* Class Relations

| **Properties** | **Range** | **Category** | **Superclasses** | **Subclasses** |
|---|---|---|---|---|
| Collection | CollectionNode | Subject | DAGNode | CapabilityFacet |
| Literal | LiteralNode | Language | | ExistentialFacet |
| Object | CoreResource | Relation | | QualityFacet |
| Property | PropertyNode | Title | | QuantityFacet |
| Verb | rdf:Property | Coverage | | StateFacet |

## 7. The *StatementNode*[11] Class

The *StatementNode* class plays a central role in the Legal-RDF ontology in two ways. The class defines components of the proposed "RDF quint", extending the now-classic "RDF quad" with explicit specification of the predicate-verb for an instance. The class additionally defines properties for the context of and the source for the statement. The Legal-RDF process model envisions that: (a) a document when drafted, yields 'content'; (b) content when annotated, yields 'identities'; (c) content when parsed, yields 'sentences' and (d) sentences when normalized using identities, yields 'statements'. This process model implies that document content contains both sentences and statements, the latter being a formally structured characterization of the former.

A role of the *StatementNode* is to package the subclasses that define predicates verbs appropriate to the particular type of statement. The

[11] http://aufderheide.info/lexmlwiki/index.php?title=Legal-RDF:StatementNode

Table XVII. *StatementNode* Class Relations

| **Properties** | **Range** | **Category** | **Superclasses** | **Subclasses, cont'd** |
|---|---|---|---|---|
| Context | ContextNode | Description | DAGNode | FactualStatement |
| Model | DAGModel | Source | rdf:Statement | FictionalStatement |
| Object | CoreResource | Coverage | | HistoricalStatement |
| Predicate | PropertyNode | Coverage | **Subclasses** | HypotheticalStatement |
| TextSource | TextBlock | Source | AcknowledgingStatement | ParentheticalStatement |
| Verb | rdf:Property | Coverage | CautionaryStatement | PermissiveStatement |
| | | | ConclusiveStatement | PredictiveStatement |
| | | | ConditionalStatement | ProhibitiveStatement |
| | | | ConsequentialStatement | RequestStatement |
| | | | DefinitionalStatement | RequirementStatement |
| | | | | ResponseStatement |

*ConditionalStatement* defines three properties – *If, Then,* and *Else* – whose ranges are *StatementNode*. Similarly, the *ConsequentialStatement* defines two properties – *When* and *Then* – having the same range.

Table XVIII. Legal-RDF Predicate Verbs[12]

| ***StatementNode*Subclass** | **Predicate Verbs** |
|---|---|
| CautionaryStatement | mayNotBeA, mayNotHaveBeenA, mayNotHave, mayNotHaveHad |
| DefinitionalStatement | isA, isNotA, wasA, wasNotA, willBeA, willNotBeA |
| FactualStatement | had, hadNot, has, hasNot, willHave, willHaveNot |
| ProhibitiveStatement | shallNotBeA, shallNotHave, shallNotHaveHad |
| PermissiveStatement | canBeA, canBeNotA, canHave, canHaveNot, canHaveHad, canHaveHadNot |
| RequirementStatement | mustBeA, mustHave, mustHaveBeenA, mustHaveNotBeenA, mustHaveHad, mustHaveNot, mustHaveHadNot, mustNotBeA |

## 8. The *ContextNode*[13]Class

The Legal-RDF ontology adopts a 'thematic' perspective in its organization of OWL classes for people, places, and things; the objective is to establish a strong guideline useful during the classification process. We observe a reality, as does an audience. These **concrete ontology classes** are categorized as one does for the elements of a play: we distinguish between the actors and their roles, between scenes and their props,

[12] This table needs refinement.

[13] http://aufderheide.info/lexmlwiki/index.php?title=Legal-RDF:ContextNode

and between the themes communicated by its dramas. An element of a play establishes a context for the interplay of other elements. Accordingly, the *ContextNode* class, a subclass of the *DAGNode* class, has six subclasses: *ActorContext, DramaContext, Prop-Context, RoleContext, SceneContext,* and *Theme-Context.* A perspective that includes the time and venue of the play, its production and other aspects, is already implicit in provenance information relating to the resources described and the attribute values provided.

## 9. The *GenericDocument*[14] and *GenericLegalDocument*[15] Classes

Table XIX shows two sets of subclasses for the *GenericDocument* class: formal subclasses (including quality classes) and second, union subclasses comprised of documents associated with economic industrial sectors[16]. The set of formal subclasses is distinguished by the type and size of the physical sheet of paper used to print and bind the contents of the document.

The *GenericLegalDocument* class represents documents that historically have been printed on legal-type of paper. It has just two properties: *Clause* and *Rider*, both of which have a range of *GenericClause.* The *GenericClause* class provides subclasses for standard types of clauses, e.g., a *DamagesClause* that could appear within a lease contract. The *Generic-Clause* class derives from the *GenericRule* class, which derives from the *LiteralNode* class.

[The *GenericRule* class also has subclasses *GenericLaw, GenericByLaw,* and *GenericRegulation.* The *GenericLaw* class subdivides to *GenericCivilLaw, GenericCriminalLaw, GenericMaritimeLaw*, and *GenericMilitaryLaw.* This class also features a union of *GenericCaseLaw, GenericCommonLaw, GenericStatutoryLaw*, and *GenericTreatyLaw* where classes exist for national, state, local and international entities to classify instances of their laws. Within the subclass for criminal law, the LegalRDF ontology has subclasses for criminal assistance and criminal conspiracy laws, while it unions commercial, economic, interpersonal, offensive, and violent laws; these have all been derived by analysis of the US justice system database structure. Each type of law is then divided into

[14] http://aufderheide.info/lexmlwiki/index.php?title=Legal-RDF:GenericDocument

[15] http://aufderheide.info/lexmlwiki/index.php?title=Legal-RDF:GenericLegalDocument

[16] These sectors coincide with those defined by the North American Industrial Classification System

subclasses for an act of violation, an activity of violation, an attempt at violation, and a threat of violation, of the law.]

Table XIX. *GenericDocument* Class Relations[17]

| **Superclasses** | **Subclasses** | **Unions** |
|---|---|---|
| DocumentModel | GenericBook | AdministrativeDocument |
| ApprovableThing | GenericBooklet | AgricultureDocument |
| ArchivableThing | GenericBlueprint | CommunityServiceDocument |
| FilableThing | GenericCalendar | ConstructionDocument |
| DistributableThing | GenericCard | EducationDocument |
| ReviewableThing | GenericCertificate | EntertainmentDocument |
| | GenericFoil | FinancialDocument |
| | GenericLabel | HealthCareDocument |
| | GenericLedger | HospitalityDocument |
| | GenericLegalDocument | InformationDocument |
| | GenericMagazine | ManagementServiceDocument |
| | GenericMap | ManufacturingDocument |
| | GenericNewspaper | MiningIndustryDocument |
| | GenericPoster | ProfessionalServiceDocument |
| | GenericSlip | PublicAdministrationDocument |
| | GenericStationery | RealtyAndLeasingDocument |
| | | RepairServiceDocument |
| | | RetailTradeDocument |
| | | TransportationWarehousingDocument |
| | | UtilityIndustryDocument |
| | | WholesaleTradeDocument |

[17] This class has just one property, not shown: *Body*, a *TextBody*, categorized as *Content*.

Table XX. *GenericLegalDocument* Selected Subclass Hierarchy[18]

| **Subclasses** | **Subclass' Subclasses** |
|---|---|
| GenericAffidavit | GenericAffidavitOfDefense, GenericAffidavitOfInquiry, GenericAffidavitOfMerits, GenericAffidavitOfNotice, GenericAffidavitOfService, GenericAffidavitOfTitle, GenericAffidavitToHoldBail |
| GenericCharter | GenericBankCharter, GenericCityCharter, GenericCorporationCharter |
| GenericLegislation | GenericLegislativeBill, GenericLegislativeLaw, GenericLegislativeResolution |
| GenericWrit | LegalWritOfArraignment, LegalWritOfAttachment, LegalWritOfCertiorari, LegalWritOfDecree, LegalWritOfElection, LegalWritOfDefaultJudgment, LegalWritOfDeficiencyJudgment, LegalWritOfDetinue, LegalWritOfError, LegalWritOfExecution, LegalWritOfFieriFacias, LegalWritOfDecreeOfForeclosure, LegalWritOfHabeusCorpus, LegalWritOfIndictment, LegalWritOfInjunction, LegalWritOfMandamus, LegalWritOfOpinion, LegalWritOfProbateWill, LegalWritOfProhibition, LegalWritOfRight, LegalWritOfScireFacias, LegalWritOfSequestration, LegalWritOfSubpoena, LegalWritOfSummons, LegalWritOfVenireFacias, LegalWritOfWarrant |

## 10. The *GenericInstrument*[19] Class

The *GenericInstrument* class (Table XXI) is another important subclass of *GenericLegalDocument.* This class features both its own subclasses plus a union of instrument types categorized by their subject matter. Its super-classes show that instruments may be amended, attested, delivered, executed, notarized, ratified, subrogated, subscripted, and transferred. Instruments are also temporal entities, meaning they have a beginning 'effective' date time and an ending 'expiration' date time.

The superclasses for the *GenericContract* class (Table XXII) highlight the possible states for a contract, e.g., proposed, offered, accepted, declined, countered, reneged, and defaulted. The subclasses shown decompose in the ontology into contracts specific for industries, types of good, and so forth.

[18] Excluded are GenericBillOfLading, GenericBrief, GenericOrder, GenericPetition, GenericRelease, and GenericTreaty. For GenericInstrument, see Table XXI.

[19] http://aufderheide.info/lexmlwiki/index.php?title=Legal-RDF:GenericInstrument

Table XXI. *GenericInstrument* Class Relations

| **Properties** | **Range** | **Category** | **Superclasses** | **Subclasses** |
|---|---|---|---|---|
| Code | LegalCode | Format | GenericLegalDoc't | NegotiableInstr't |
| GoverningInstr't | GenericInstr't | Relation | TemporalThing | TestamentaryInstr |
| Jurisdiction | LegalJurisdiction | Format | AmendableThing | |
| InterestedParty | GenericActor | Contributor | AttestableThing | **Unions** |
| Party | GenericActor | Creator | DeliverableThing | GenericBond |
| SubordinateInstr't | LegalInstrument | Relation | ExecutableThing | GenericContract |
| | | | NotarizableThing | GenericDeed |
| | | | RatifiableThing | GenericLease |
| | | | SubrogatableThing | GenericWill |
| | | | SubscriptableThing | |
| | | | TransferableThing | |

Table XXII. *GenericContract* Class Relations

| **Properties** | **Range** | **Category** | **Superclasses** | **Subclasses** |
|---|---|---|---|---|
| Consideration | FactualStatement | Description | GenericInstr't | AmendmentAgree'tInst't |
| Obligation | GenericEvent | Description | AcceptableThing | BuybackAgree'tInst't |
| | | | CounterableThing | BuysellAgree'tInstr't |
| | | | DeclinableThing | ExtensionAgree'tInstr't |
| | | | DefaultableThing | **Unions** |
| | | | OfferableThing | AgencyAgree'tInst't |
| | | | ProposableThing | CivilAgree'tInst't |
| | | | RenegableThing | CommercialAgree'tInst't |
| | | | | FinancialAgree'tInst't |
| | | | | OwnershipAgree'tInst't |
| | | | | PublicWorksAgree'tInst't |
| | | | | PurchaseAgree'tInstr't |
| | | | | ServiceAgree'tInst't |
| | | | | UseAgree'tInst't |

## 11. Concluding Remarks

Legal-RDF is creating an ontology useful during semantic annotation of the content of XHTML documents; during exchange of RDF documents; and during execution of ECMA software. The strengths of its ontology are found in its commitment to

(a) integration of basic markup standards (RDF, XML Schema, and XHTML) and international standards (ISO, SI, and NAICS);

(b) segregation of predicate verbs from predicate nouns;

(c) organization of all defined attributes into Dublin Core categories;

(d) adoption of a pronounced perspective for defined concrete classes;

(e) establishment of quality-laden class hierarchies; and

(f) adherence to a statement-based model for legal (document) content.

Support for these requirements are lacking or incomplete in the candidate ontologies reviewed during the initial design of the Legal-RDF ontology. For instance, DOLCE lacks (a), (b), (c), and (f); (d) is implicit in its use of the controversial perdurant/endurant model, and (e) occurs non-rigorously in its set of base classes. The SUMO and LKIF ontologies have similar profiles which also prevented their use towards the goal of the Legal-RDF ontology: to represent the *entirety* of a legal document in a manner practical to both government and industry.

Since this ontology seeks to meet the needs of legal professionals, then its scope needs to encompass all types of documents they encounter. As the Legal-RDF ontology evolves from Version 1 (which has 15,000+ terms) to the expressive models of Version 2, and as it leverages the economies of a public Wiki, this goal is both realistic and attainable.

## 12. Readings

DOLCE Ontologies: `http://www.loa-cnr.it/DOLCE.html`

Legal-RDF Ontologies:
Ver 1: `http://www.legal-rdf.org/` and `http://www.legal-xhtml.org/`
Ver 2: `http://aufderheide.info/lexmlwiki/index.php?title=Legal-RDF_Ontologies`

LKIF Ontology: `http://www.estrellaproject.org/lkif-core/-documentation`

McClure, J., Annotation of Legal Documents in XHTML, V Legislative XML Workshop. June, 2006. `http://www.ittig.cnr.it/legws/program.html`

Niles, I., and Pease, A. 2001. `http://projects.teknowledge.com/HPKB/Publications/FOIS.pdf` Towards a Standard Upper Ontology(`http://projects.teknowledge.com/HPKB/Publications/FOIS.pdf`). In Proceedings of the 2nd International Conference on Formal Ontology in Information Systems (FOIS-2001), Chris Welty and Barry Smith, eds, Ogunquit, Maine, October 17-19, 2001.

Xerox PARC, `http://www.parc.xerox.com/research/projects/aspectj/default.html`

# The LKIF Core Ontology of Basic Legal Concepts

Rinke Hoekstra, Joost Breuker, Marcello Di Bello, Alexander Boer
*Leibniz Center for Law, University of Amsterdam*
*breuker@science.uva.nl, hoekstra@uva.nl, mbello@science.uva.nl, aboer@uva.nl*

**Abstract.** In this paper we describe a legal core ontology that is part of a generic architecture for legal knowledge systems, which will enable the interchange of knowledge between existing legal knowledge systems. This *Legal Knowledge Interchange Format*, is under development in the Estrella project and has two main roles: 1) the translation of legal knowledge bases written in different representation formats and formalisms and 2) a knowledge representation formalism that is part of a larger architecture for developing legal knowledge systems. A legal (core) ontology can play an important role in the translation of existing legal knowledge bases to other representation formats, in particular into LKIF as the basis for articulate knowledge serving. We describe the methodology underlying the LKIF core ontology, introduce the concepts it defines, and discuss its use in the formalisation of an EU directive.

**Keywords:** ontology, legal ontology, legal concept, LKIF, knowledge representation, framework

## 1. Introduction

In this paper we describe a legal core ontology that is part of a generic architecture for legal knowledge systems, which will enable the interchange of knowledge between existing legal knowledge systems. This *Legal Knowledge Interchange Format* (LKIF), is currently being developed in the Estrella project.[1] LKIF has two main roles: enable the translation between legal knowledge bases written in different representation formats and formalisms and secondly, as a knowledge representation formalism that is part of a larger architecture for developing legal knowledge systems. These use-cases for LKIF bring us to the classical trade-off between tractability and expressiveness, as in e.g. KIF (Knowledge Interchange Format, (Genesereth and Fikes, 1992)). An additional requirement is that LKIF should comply with current Semantic Web standards to enable legal information serving via the web: the core of LKIF consists of a combination of OWL-DL and SWRL, offering a classical hybrid solution. How these two formalisms have to be combined still is an important issue in the development of LKIF, and for details the reader is referred to (Boer et al., 2007).

[1] Estrella is a 6th European Framework project (IST-2004-027665). See also: `http://www.estrellaproject.org`. The views and work reported here are those of the authors.

Proposing the OWL-DL subset of SWRL as its core does not make LKIF a formalism tuned to *legal* knowledge and reasoning: how do we get the 'L' into LKIF? To "legalize" LKIF it needs to be constrained in two ways. The first is a *meta-component* that controls the *reasoning* as to gear it to typical legal tasks. For instance, legal assessment and argumentation provide control structures for legal reasoning that put specific demands on the knowledge to be obtained from a legal knowledge base. The second constraint is not specialised to legal reasoning, but to *legal knowledge*. Typical legal concepts may be strongly interrelated and thereby provide the basis for computing equivalencies (paraphrases) and implications. For instance, by representing an obligation as the opposite of a prohibition, a (legal) knowledge system can make inferences that are specialised to these terms. In our view, specialised legal inference should be based on definitions of concepts involved in an ontology. Concept definitions should make all necessary and sufficient interrelationships explicit; the inference engine can then generate all implied consequences. [2]

A legal ontology can play an important role in the translation of existing legal knowledge bases to other representation formats, in particular into LKIF as the basis for articulate knowledge serving. Similar to a translation between different natural languages, a formal, 'syntactic' translation may clash with the semantics implied by the original knowledge representation. An ontology, as representation of the semantics of terms, allows us to keep track of the use of terms in a knowledge base. Furthermore, and more importantly, an ontology can support the process of knowledge acquisition and modelling in legal domains. Defining concepts like 'norm', 'judge', 'liability', 'document', 'claim', etc. helps to structure the process of knowledge acquisition. Earlier experience, as in e.g. (Breuker and Hoekstra, 2004b; Breuker and Hoekstra, 2004a), suggests a commonsense basis for distinguishing main categories in an ontology for law.

The following sections describe the theoretical and methodological framework against which the LKIF core ontology has been developed (Section 2 and 3). Section 4 describes the different modules of the ontology, and introduces its most important concepts. Section 5 gives an example of how the ontology can be used in the formalisation of a regulation.

[2] For an ontology cast in OWL-DL these inference engines are description classifiers, e.g. Pellet, `http://pellet.owldl.com/`

## 2. Frameworks and Ontologies

We adhere to a rather restrictive view on what an ontology should contain: terminological knowledge, i.e. intensional definitions of concepts, represented as classes with which we interpret the world. The distinction between terminological knowledge (T-Box) and assertional knowledge (A-Box) has already been around for a long time. As a rule, terminological knowledge is generic knowledge while assertional knowledge describes the (actual) state of some world: situations and events. However, these asserted states can become generalised into typical patterns related to particular situations. To be sure, if experiences re-occur and have a justifiable structure, it might evidently pay to store these structures as generic descriptions, because they deliver a predictable course of events for free. Eating in a restaurant is a typical example and it served in the Seventies to illustrate the notion of knowledge represented by scripts (Schank and Abelson, 1977) or 'frames' (Minsky, 1975). This kind of generic knowledge is indeed rooted in terminological knowledge, but is structured differently. Where ontologies have a taxonomic structure, frames are dominated by mereological and dependency relationships.

Finally, an important reason to distinguish frameworks from ontology proper is that frameworks often imply epistemic roles which require reasoning architectures that go beyond the services provided by OWL-DL reasoners (e.g. meta-level reasoning). It should be noted that frameworks are generic, i.e. they act as pre-specified patterns that get instantiated for particular situations. We have distinguished the following types of frameworks:

*Situational frameworks* Situational frameworks are stereotypical structures of plans for achieving some goal in a recurrent context. Making coffee may be such a plan. However, the plans may involve transactions in which more than one actor participates. For instance, the definition of Eating-in-a-restaurant[3] shows the dependencies between actions of clients (ordering, paying) and service personnel (noting, serving) as its major structure. This is the internal structure of the concept, but it usually does not make sense to create class-subclass relations between such frame-like concepts. The Eating-in-a-restaurant is not some *natural* sub-class of Eating. It refers to some typical model of how eating is put in the context of a restaurant. We can introduce a proliferation of all contexts of eating, such as Eating-at-home, Eating-with-family, etc. but these contexts do not fundamentally differ, cf. (Bodenreider et al.,

[3] In the following all concepts will start with a capital, properties and relations will not

2004; Breuker and Hoekstra, 2004a). In the legal world, such situational frameworks may be pre-scribed in articles of procedural ('formal') law. Although *stereotypical* plans ('customs') and *prescribed* plans may differ in their justification – rationality vs. authority – their representation is largely analogous. Similarly, legal norms combine generic situation descriptions with some specific state or action. The description is qualified by a deontic term. For instance, the norm that "vehicles should keep to the right of the road" states that the situation in which a vehicle keeps to the right is obliged.

*Mereological frameworks* Many entities, both objects and processes often have parts: they are *composites*. It is tempting to include a mereological (part-of) view in the definition of a concept. For instance, defining a car as having at least three, and usually four wheels, and at least one motor. However, a full *structural* description of all its parts and connections goes beyond what a car *essentially* is. Mereological frameworks appear under a large diversity of names: structural models, configurations, designs, etc. Arguably, the distinction between a mereological framework and a defining description of a term (ontology) is sometimes be very thin. For instance, if we want to describe a bicycle as distinct from a tricycle, it is necessary to use the cardinality of the wheels as defining properties as these are *central* to the nature of the bicycle. On the other hand, the number of branches a tree might have hardly provides any information as to what a tree *is*.

*Epistemological frameworks* Inference structures are often represented as epistemological frameworks of interdependencies between reasoning steps. Typical examples are the problem solving methods (PSM) found in libraries of problem solving components (Breuker and Van de Velde, 1994; Motta, 1999; Schreiber et al., 2000)[4] A problem solving method is not only a break-down of a problem, but also provides control over the making of inferences by assessing success and failure in arriving at the (sub)goals. PSMs have two major components: some method for selecting or generating potential solutions (hypotheses), and some methods for testing whether the solutions hold. Whether they hold may be due to the fact that they satisfy all the specified requirements (constraints) or whether they correspond with ('explain') empirical data.

This focus on the *use* of knowledge, its epistemological *status* (e.g. hypothesis vs. conclusion) and the dependencies between distinct steps in a methodology is characteristic for epistemological frameworks. Epis-

[4] Although the terms 'reasoning' and 'inference' are often used as more or less synonymous, we want to reserve the term inference for making explicit what is implicit in a knowledge base, given some inference engine.

temological frameworks can be more abstract than PSMs. For instance, the Functional Ontology of Law, which is presented as a core ontology, is an epistemological framework that describes the role of law as a control system in society (Valente, 1995; Breuker et al., 2004).

## 3. Methodology

The construction of LKIF followed a combination of methodologies for ontology engineering. Already in the mid-nineties, the need for a well-founded methodology was recognised, most notably by (Gruber, 1994; Grüninger and Fox, 1995; Uschold and King, 1995; Uschold and Grünin-ger, 1996) and later (Fernández et al., 1997). These methodologies follow in the footsteps of earlier experiences in knowledge acquisition, such as the CommonKADS approach (Schreiber et al., 2000) and others, but also considerations from naive physics and cognitive science, such as (Hayes, 1985) and (Lakoff, 1987), respectively.

(Hayes, 1985) describes an approach to the development of a large-scale knowledge base of naive physics. Instead of rather metaphysical top-down construction, his approach starts with the identification of relatively independent *clusters* of closely related concepts. These clusters can be integrated at a later stage, or used in varying combinations allowing for greater flexibility than monolithic ontologies. Furthermore, by constraining (initial) development to clusters, the various – often competing – requirements for the ontology are easier to manage.

Whereas the domain of (Hayes, 1985)'s proposal concerns the relatively well-structured domain of physics, the combination of commonsense and law does not readily provide an obvious starting point for the identification of clusters. In other words, for LKIFcore, we cannot carve-up clusters from a pre-established middle ground of commonsense and legal terms. Furthermore, the field does not provide a relatively stable top level from which top-down development could originate.

In (Uschold and King, 1995), who are the first to use the term 'middle-out' in the context of ontology development, it is stressed that the most 'basic' terms in each cluster should be defined before moving on to more abstract and more specific terms within a cluster. The notion of this basic level is taken from (Lakoff, 1987), who describes a theory of categorisation in human cognition. Most relevant within the context of ontology engineering (Uschold and King, 1995; Lakoff, 1987, p. 12 and 13) are *basic-level categorisation*, *basic-level primacy* and *functional embodiment*. Categories are organised so that the categories that are cognitively basic are 'in the middle' of a taxonomy, generalisation proceeds 'upwards' from this basic level and specialisation

proceeds 'downwards'. Furthermore, these categories are functionally and epistemologically primary with respect to (amongst others) knowledge organisation, ease of cognitive processing and ease of linguistic expression. Basic level concepts are used automatically, unconsciously, and without noticeable effort as part of normal functioning. They have a different, and more important psychological status than those that are only thought about consciously.

For the purpose of the LKIF ontology, we have made slight adjustments to the methodology of (Hayes, 1985; Uschold and Grü-nin-ger, 1996). We established design criteria for the development of the LKIF ontology based on (Gruber, 1993; Uschold and Grü-nin-ger, 1996). These criteria were implemented throughout the following phases: identify *purpose and scope*, ontology *capture* and *coding*, *integration* with existing ontologies and *evaluation*. The following section describes how these phases have materialised in the context of LKIF Core. Furthermore, an example in which the ontology is put to use is described in section 5.

.

## 4. Modules & Outline

This section describes how the methodology described in the previous section was applied to the development of LKIF Core. We first describe the building and clustering phase, followed by a discussion of the existing ontologies we considered for inclusion, and a description of the concepts defined in the different modules of the ontology.

### 4.1. Ontology Capture

The LKIF Core ontology should contain 'basic concepts of law'. It is dependent on the (potential) users what kind of vocabulary is aimed at. We have identified three main groups of users: *citizens*, *legal professionals* and *legal scholars*. Although legal professionals use the legal vocabulary in a far more precise and careful way than laymen, it appears that for most of these terms there is still a sufficient common understanding to treat them more or less as similar (Lame, 2006). Nonetheless, a number of basic terms have a specific legal-technical meaning, such as 'liability' and 'legal fact'. We included these technical terms because they might capture the 'essential', abstract meaning of terms in law, but also because these terms might be used to organise more generally understood legal terms.

The Estrella consortium includes representatives of the three kinds of experts. Each partner was asked to supply their 'top-20' of legal

concepts. Combined with terms we collected from literature (jurisprudence and legal text-books) we obtained a list of about 250 terms. As such a number is unmanageable as a basic set for modelling, we asked partners to assess each term from this list on five scales: level of *abstraction*, *relevance* for the legal domain, the degree to which a term is *legal* rather than *common-sense*, the degree to which a term is a *common legal term* (as opposed to a term that is specific for some sub-domain of law), and the degree to which the expert thinks this term should be *included* in the ontology. The resulting scores were used to select an initial set of 50 terms plus those re-used from other ontologies (see section 4.2), and formed the basis for the identification of clusters and the development of the LKIF Core ontology.

### 4.2. Other Ontologies

We expected to be able to reuse terms and definitions from existing core or upper ontologies that contain legal terms, as e.g. listed in (Casanovas et al., 2006). Unfortunately, it turned out that the amount of re-use and inspiration was rather limited. The following core ontologies for law were consulted, both for their potential contribution for creating a coherent top for LKIF Core, and specifically for legal terms already represented.

The intentional nature of the core concepts for the LKIF ontology (see e.g. sections 4.3.2,4.3.3) emphasises the distinction with other more (meta)physically inclined top ontologies such as SUMO[5], Sowa's upper ontology (Sowa, 2000) and DOLCE[6] (Gangemi et al., 2002)), but shows similarities with the distinction between *intentional*, *design* and *physical* stances described in (Dennett, 1987). As some of these top- or upper ontologies (SUMO, Sowa) do not have a common-sense basis – e.g. mental and social entities are poorly represented – they could neither be used as a top for LKIF Core, nor as a source of descriptions of legal terms. The upper part of the CYC [7] ontology and DOLCE (Gangemi et al., 2003; Massolo et al., 2002) are claimed to have a common-sense view, but this common-sense view is rather based upon personal intuition than on empirical evidence. LRI-Core on the other hand is to a large extent based upon empirical studies in cognitive science, and is intended as a core ontology for law. However, the number of typical legal concepts in this legal core ontology is disappointingly small. Nonetheless, its top structure appeared to be valuable in constructing LKIF as is further described in Section 4. The Language for Legal Discourse (McCarty,

[5] Suggested Upper Merged Ontology; `http://ontology.teknowledge.com`

[6] Descriptive Ontology for Linguistic and Cognitive Engineering; `http://www.loa-cnr.it/DOLCE.html`

[7] www.cyc.com

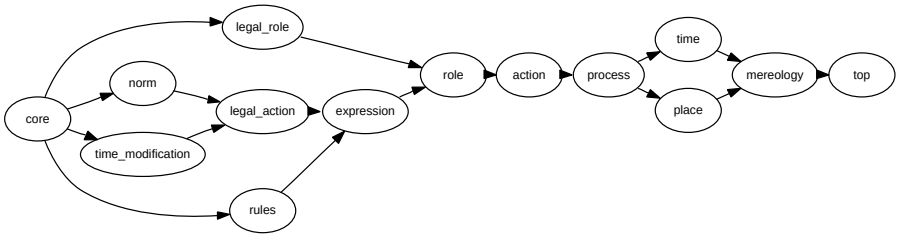



*Figure 1.* Dependencies between LKIFCore modules.

1989, LLD) is a first attempt to define legal concepts in the context of legal reasoning, using formulae and rules. Properly speaking, LLD is not an ontology but a framework but it is a relatively rich source for legal terms and their definitions. The Core Legal Ontology (CLO) is used to support the construction of legal domain ontologies (Gangemi et al., 2005). CLO organises legal concepts and relations on the basis of formal properties defined in DOLCE+. Although purpose and layers are largely similar to those of LRI-Core, the top structures differ considerably.

### 4.3. Ontology Modules

The list of terms and insights from the requirements-phase resulted in a collection of ontology modules, each of which represents a relatively independent cluster of concepts: *expression*, *norm*, *process*, *action*, *role*, *place*, *time* and *mereology* (Breuker et al., 2006; Breuker et al., 2007). The concepts in these clusters were formalised using OWL-DL in a middle-out fashion: for each cluster the most central concepts were represented first.[8]

Discussions, further literature study and the consideration of existing ontologies, led to an extension of the original set of clusters to 14 modules (see Figure 1), each of which describes a set of closely related concepts from both legal and commonsense domains. Nonetheless, we maintained the original views used to identify the clusters, as the explanations and justifications are still valid and applicable to the current version of the ontology. We can distinguish three layers in the ontology: the *top* level (Section 4.3.1), the *intentional* level (Section 4.3.2) and the *legal* level (Section 4.3.3).

#### 4.3.1. *First Things First: The top-level*

The description of any legally relevant fact, event or situation requires a basic conceptualisation of the context in which these occur: the backdrop, or canvas, that is the physical world. Fundamental notions such as

[8] We used both TopBraid Composer (`http://www.topbraidcomposer.com`) and Protégé 3.2/4.0 (`http://protege.stanford.edu`).

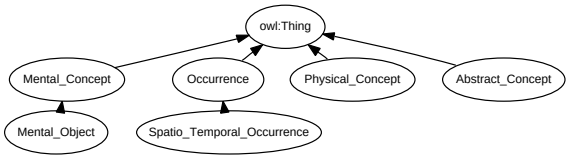



*Figure 2.* Concepts defined in the Top module.

location, time, parthood and change are indispensable in a description of even the simplest legal account. The top level clusters of the ontology provide (primitive) definitions of these notions, which are consequently used to define more intentional and legal concepts in other modules. The most general classes of the LKIF ontology are borrowed from LRI Core. We distinguish between mental, physical and abstract concepts, and occurrences (Figure 2).

Mereological relations allow us to define parts and wholes, allow for expressing a systems-oriented view on concepts, such as functional decompositions, and containment (Figure 3). Furthermore, they form the basis for definitions of *places* (location) and moments and intervals in *time*.

The ontology for places in LKIF Core is based on the work of (Donnelly, 2005), and adopts a distinction between *relative* places and *absolute* places, which goes back to Isaac Newton. Whereas a relative place is defined by reference to some thing, absolute places are part of absolute space and have fixed spatial relations with other absolute places. See figure 3 for an overview of concepts defined in the place module. A Location_Complex is a set of places that share a reference location.

Of the properties defined in this module, meet is the most basic as it is used to define many of the other properties such as abut, cover, coincide etc. See (Breuker et al., 2007; Donnelly, 2005) for a more in depth discussion of these and other relations. The current version of the ontology of places does not define concepts and relations that can be used to express direction and orientation.

Closely related to the theory of places of (Donnelly, 2005) is Allen's theory of time (Allen, 1984; Allen and Ferguson, 1994). We adopt his theory, and distinguish between the basic concepts of Interval and Moment. Intervals have an extent (duration) and can contain other intervals and moments. Moments are points in time, they are atomic and do not have a duration or contain other temporal occurrences (see figure 4).

The relations between temporal occurrences are what defines time. Like (Donnelly, 2005), (Allen, 1984) adopts the meet relation to define two immediately adjacent temporal occurrences. We call this relation

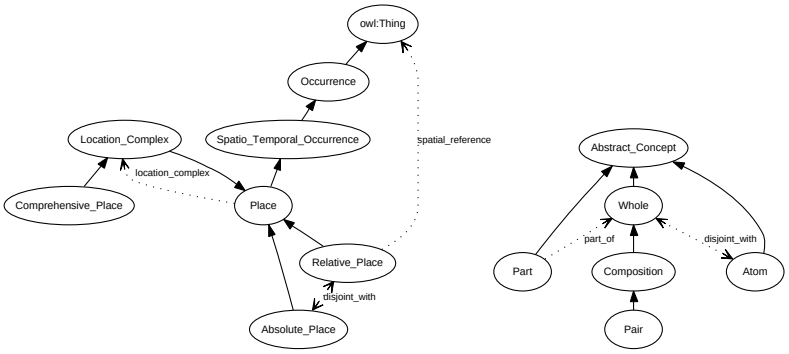



*Figure 3.* Place and Mereology related concepts.

immediately_before, as the temporal meet relation holds only in one direction, and is asymmetric. The property is used to define other temporal relations such as before, after, during, etc.

With these classes and properties in hand, we introduce concepts of (involuntary) change. The process ontology relies on descriptions of time and place for the representation of duration and location of changes. A Change is essentially a difference between the situation before and after the change. It can be a functionally coherent aggregate of one or more other changes. More specifically, we distinguish between Initiation, Continuation and Termination changes.

Changes that occur according to a certain recipe or procedure, i.e. changes that follow from causal necessity are Processes; they introduce causal propagation. Contrary to changes, processes are bound in time and space: they have duration and take place at a time and place. We furthermore distinguish Physical_Processes which operate on Physical_Objects. Furthermore, at this level we do not commit to a particular theory of causation or causal propagation.

4.3.2. *The Intentional Level*

Legal reasoning is based on a common sense model of intelligent behaviour, and the prediction and explanation of intelligent behaviour. It is after all only behaviour of rational agents that can be effectively influenced by the law. The modules at the intentional level include concepts and relations necessary for describing this behaviour (i.e. Actions undertaken by Agents in a particular Role) which are governed by law. Furthermore, it introduces concepts for describing the mental state of these agents, e.g. their Intention or Belief, but also communication between agents by means of Expressions.

The class of agents is defined as the set of things which can be the actor of an intentional action: they perform the action and are

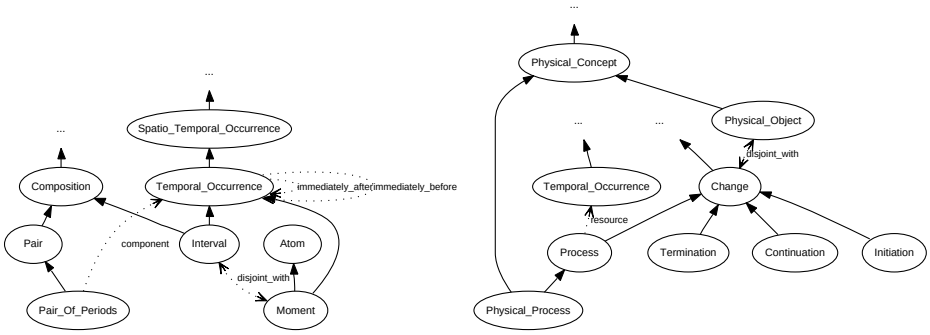



*Figure 4.* Concepts related to time and change.

potentially liable for any effects caused by the action (see figure 5). Actions are processes, they are the changes performed by some agent who has the intention of bringing about the change. Because actions are processes they can become part of causal propagation, allowing us to reason backwards from effect to agent. Actions can be creative in that they initiate the coming into existence of some thing, or the converse. Also, actions are often a direct *re*action to some other action (see figure 5).

The agent is the medium of some intended outcome of the action: an action is always intentional. The intention held by the agent, usually bears with it some expectation that the intended outcome will be brought about: the agent believes in this expectation. The actions an agent is expected or allowed to perform are constrained by the *competence* of the agent, sometimes expressed as *roles* assigned to the agent.

We distinguish between persons, individual agents such as "Joost Breuker" and "Pope Benedict XVI", and Organisations, aggregates of other organisations or persons which acts 'as one', such as the "Dutch Government" and the "Sceptics Society" . Artefacts are physical objects designed for a specific purpose, i.e. to perform some Function as instrument in a specific set of actions such as "Hammer" and "Atlatl"[9]. Persons are physical objects as well, but are not designed (though some might hold the contrary) and are subsumed under the class of Natural_Objects. Note that natural objects can function as tools or weapons as well, the typical example being a stone, but are not designed for that specific purpose.

The notion of roles has played an important part in recent discussions on ontology (Steimann, 2000; Masolo et al., 2004; Guarino and

[9] An atlatl is a tool that uses leverage to achieve greater velocity in spear-throwing, see `http://en.wikipedia.org/wiki/Atlatl`

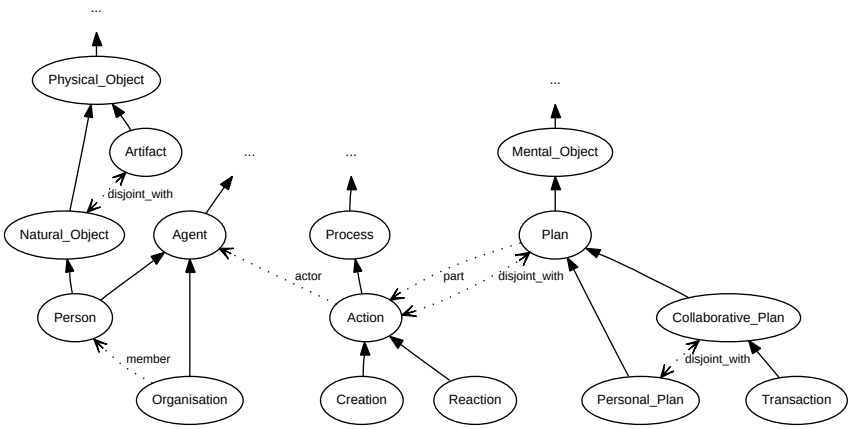



*Figure 5.* Actions, agents and organisations.

Welty, 2002). Roles not only allow us to categorise objects according to their prototypical use and behaviour, they also provide the means for categorising the behaviour of other agents. They are a necessary part of making sense of the social world and allow for describing social organisation, prescribe behaviour of an agent within a particular context, and recognise deviations from 'correct' or normal behaviour. Indeed, roles and actions are closely related concepts: a role defines some set of actions that can be performed by an agent, but is conversely defined by those actions. Roles specify standard or required properties and behaviour (see figure 6). The role module captures the roles and functions that can be played and held by agents and artefacts respectively, and focuses on *social* roles, rather than traditional thematic or relational roles.

A consequence of the prescriptive nature of roles is that agents connect expectations of behaviour to other agents: intentions and expectations can be used as a model for intelligent decision making and planning[10]. It is important to note that there is an *internalist* and an *externalist* way to use intentions and expectations. The external observer can only ascribe intentions and expectations to an agent based on his observed actions. The external observer will make assumptions about what is *normal*, or apply a *normative* standard for explaining the actions of the agent.

[10] Regardless of whether it is a psychologically plausible account of decision making. Daniel Dennett's notion of the *Intentional Stance* is interesting in this context (cf. (Dennett, 1987)). Agents may do no more than occasionally apply the stance they adopt in assessing the actions of others to themselves.

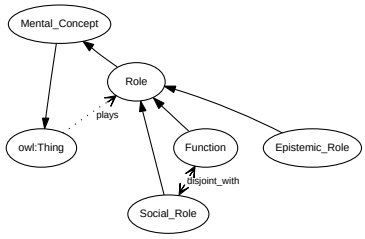



*Figure 6.* Roles.

The expression module covers a number of representational primitives necessary for dealing with Propositional_Attitudes (viz. (Dahllöf, 1995)). Many concepts and processes in legal reasoning and argumentation can only be explained in terms of propositional attitudes: a relational mental state connecting a person to a Proposition. However, in many applications of LKIF the attitude of the involved agents towards a proposition will not be relevant at all. For instance, fraud detection applications will only care to distinguish between potentially contradictory observations or expectations relating to the same propositional content. Examples of propositional attitudes are Belief, Intention, and Desire. Each is a component of a mental model, held by an Agent.

Communicated attitudes are held towards expressions: propositions which are externalised through some medium. Statement, Declaration, and Assertion are expressions communicated by one agent to one or more other agents. This classification is loosely based on Searle (cf. (Searle and Vanderveken, 1985)). A prototypical example of a medium in a legal setting is e.g. the Document as a bearer of legally binding (normative) statements.

When propositions are used in reasoning they have an epistemic role, e.g. as Assumption, Cause, Expectation, Observation, Reason, Fact etc. The role a proposition plays within reasoning is dependent not only on the kind of reasoning, but also the level of trust as to the validity of the proposition, and the position in which it occurs (e.g. hypothesis vs. conclusion). In this aspect, the expression module is intentionally left under-defined. A rigourous definition of propositional attitudes relates them to a theory of reasoning and an argumentation theory. The argumentation theory is supplied by an argumentation ontology. The theory of reasoning depends on the type of reasoning task (assessment, design, planning, diagnosis, etc.) LKIF is used in, and should be filled in (if necessary) by the user of LKIF .

Evaluative_Attitudes express an evaluation of a proposition with respect to one or more other propositions, they express e.g. an evaluation, a value statement, value judgement, evaluative concept, etc. I.e. only

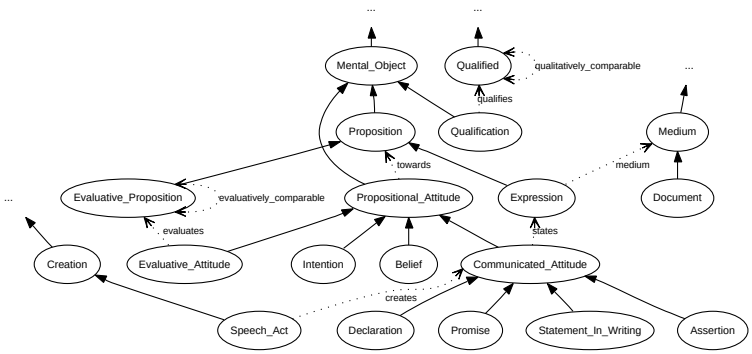



*Figure 7.* Propositions, Attitudes and Expressions.

the type of qualification which is an attitude towards the thing being evaluated, and not for instance the redness of a rose, as in (Gangemi et al., 2002) and others. Of special interest is the Qualification, which is used to define norms based on (Boer et al., 2005). Analogous to the evaluative attitude, a qualification expresses a judgement. However, the subject of this judgement need not be a proposition, but can be any complex description (e.g. a situation).

4.3.3. *The Legal Level*

Legally relevant statements are created through public acts by both natural and legal persons. The legal status of the statement is dependent on both the kind of agent creating the statement, i.e. Natural_Person vs. a Legislative_Body, and the rights and powers attributed to the agent through mandates, assignments and delegations. At the legal level, the LKIF ontology introduces a comprehensive set of legal agents and actions, rights and powers (a modified version of (Sartor, 2006; Rubino et al., 2006)), typical legal roles, and concept definitions which allow us to express normative statements as defined in (Boer et al., 2005; Boer, 2006; Boer et al., 2007).

The Norm is a statement combining two performative meanings: it is *deontic*, in the sense that it is a qualification of the (moral or legal) acceptability of some thing, and it is *directive* in the sense that it commits the speaker to bringing about that the addressee brings about the more acceptable thing (cf. (Nuyts et al., 2005)), presumably through a sanction. These meanings do not have to occur together. It is perfectly possible to attach a moral qualification to something without directing anyone, and it is equally possible to issue a directive based on another reason than a moral or legal qualification (e.g. a warning).

A norm applies to (or qualifies) a certain situation (the Qualified situation), allows a certain situation – the Obliged situation or Allowed

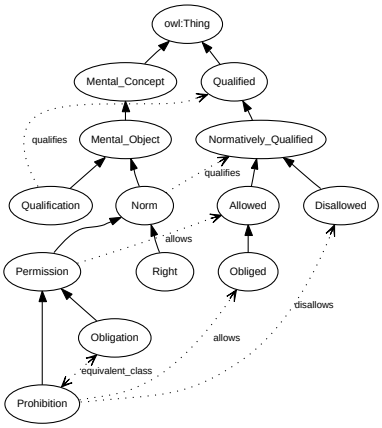



*Figure 8.* Qualifications and Norms

situation – and disallows a certain situation – the Prohibited or Disallowed situation, see Figure 8. The obliged and prohibited situation are both subsumed by the situation to which the norm applies. Besides that they by definition form a complete partition of the case to which the norm applies, i.e. all situation to which the norm applies are either a mandated case or a prohibited situation. This is true of the obligation and the prohibition: they are simply two different ways to put the same thing into words. The permission is different in that it allows something, but it does not prohibit anything. The logical complement of the mandated situation is here an opposite qualified situation, about which we know only that it cannot be obliged.

## 5. Putting the ontology to use: the Traffic domain

The LKIF ontology not only provides a theoretical understanding of the legal domain, but its primary use in practice is as a tool to facilitate knowledge acquisition, exchange and representation: i.e. to formalise pieces of existing legislation. We evaluated the use of the ontology by formalising the EU Directive 2006/126 on driving licences,[11], a relatively straightforward regulation, in which at least two types of normative statement are recognisable—definitional and deontic.

An example of a *definitional statement* from the EU directive is:

> Art. 4(2) **Category AM**: *Two-wheel vehicles or three-wheel vehicles with a maximum design speed of not more than 45 km/h.*

[11] The text is available on-line at `http://eur-lex.europa.eu/`.

The mereo module of the ontology along with a qualified cardinality restriction (available with OWL 1.1) allows us to express that AM vehicles have two or three wheels:

$$\mathsf{AM} \sqsubseteq 2\mathsf{composed\_of.Wheel} \sqcup 3\mathsf{composed\_of.Wheel}.$$

Modelling 'design speed not more than 45 km/h' is more challenging as it requires us to represent the rather common sense domain of speeds, distances etc. Of course, one could introduce the datatype property designSpeed and require its value be expressed in km/h.This choice, however, would not make justice of the conceptual complexity involved in 'design speed not more than 45 km/h', which contains reference to several notions: unit of measurement, number, designed speed, and a no-more relation. In fact, 'design speed not more than 45 km/h' can be rendered by imposing an *linear ordering* relation less-than on the different (instances of the) subclasses of the class DesignSpeed.[12] The ordering allows us to define the class of those DesignSpeeds with a value not exceeding some N45—i.e., $\forall$less-than.DesSpeed-km-h-45.

Let us now look at an example of a *deontic statement*:

Art. 4(2) *The minimum age for category AM is fixed at 16 years.*

Art. 4(2) expresses an obligation whose logical form can be rendered by the implication:

*If $x$ is driving a AM vehicle, then $x$* **must** *be at least 16 years old.*

To fix some terminology, the antecedent is the *context* to which the obligation applies; the consequent (minus the deontic operator **must**) is the *content* of the obligation itself (what the obligations prescribes it ought to be the case). Consistently with this analysis, the LKIF ontology defines obligations as classes (see Section 4.3.3).

In our case, art. 4(2) allows the situation DriverAM $\sqcap$ DriverOlderThan16 and forbids DriverAM $\sqcap \neg$DriverOlderThan16. Suppose that the classes DriverOlderThan16 and DriverAM have already been defined.[13] To model the obligation that drives of AM vehicles must be at least the 16 years older, we introduced the obligation-type class MinAgeAM as follows:

[12] The ordering is linear—i.e, reflexive, antisymmetric, transitive and total—since it mirrors the ordering of the natural numbers. For whenever $n \leq m$, we have that DesignSpeed-km-h-n(a) less-than DesignSpeed-km-h-m(b), with $a, b$ instances.

[13] The class DriverOlderThan16 can be defined by using a more-than ordering relation, roughly along the same lines as the class $\forall$less-than.DesSpeed-km-h-45. The class DriverAM can be easily defined.

$$\begin{aligned}
\mathsf{MinAgeAM} &\sqsubseteq \forall \mathsf{allows}.(\mathsf{DriverAM} \sqcap \mathsf{DriverOlderThan16}).\\
\mathsf{MinAgeAM} &\sqsubseteq \exists \mathsf{allows}.(\mathsf{DriverAM} \sqcap \mathsf{DriverOlderThan16}).\\
\mathsf{MinAgeAM} &\sqsubseteq \forall \mathsf{disallows}.(\mathsf{DriverAM} \sqcap \neg\mathsf{DriverOlderThan16}).\\
\mathsf{MinAgeAM} &\sqsubseteq \exists \mathsf{disallows}.(\mathsf{DriverAM} \sqcap \neg\mathsf{DriverOlderThan16}).
\end{aligned}$$

Other deontic operators, such as permission or prohibition, can be accounted in an alike manner (see (Boer et al., 2007)). Notwithstanding the parsimony of this type of definition, using the LKIF ontology to model normative statements proves to be rather straightforward. Of course, a specialised modelling environment for legislative drafters would need to provide a shorthand for such standard OWL definitions.[14]

The representation of art. 4(2) suggests the LKIF ontology be augmented with a module taking care of quantities, units of measurement, numbers, fractions, mathematical operations, and the like. This is crucial not only for the EU Directive 2006/126, in which most definitional statements contain quantitative features of vehicles (e.g., power, cylinder capacity); quantities and calculations play a central role in any legislative text. Note, however, that the LKIF ontology can only provide a *purely terminological* account, without being able to do mathematical computations. This is unavoidable, given that OWL is a purely logical language. We are currently investigating whether we can import an existing OWL ontology dealing with measurements, such as PhysSys/SUMO or from the Ontolingua server[15].

## 6. Discussion

As LKIF Core was developed by a heterogeneous group of people, we specified a number of conventions to uphold during the representation of terms identified in the previous phases (See (Breuker et al., 2007)). One of these is that classes should be represented using necessary & sufficient conditions as much as possible (i.e. by means of equivalentClass statements). Using such 'complete' class definitions ensures the ability to infer the type of individuals; this does not hold for partial class definitions (using only necessary conditions).

In retrospect, this convention turned out to pose severe problems for existing OWL-DL and OWL 1.1 reasoners as their performance is significantly affected by the generic concept inclusion axioms (GCI):

[14] See e.g. the SEAL project, `http://www.leibnizcenter.org/project/current-projects/seal`

[15] See `http://www.ksl.stanford.edu/software/ontolingua/`

axioms where the left-hand side of a subClassOf statement is a complex class definition. These axioms are abundant when defining classes as equivalent to e.g. someValuesFrom restrictions and in combination with lots of inverse property definitions, this creates a large completion graph for DL reasoners[16] As result of these findings, the LKIF ontology has undergone a significant revision since its initial release.

Using LKIF Core in practice, as e.g. in the traffic example, points to the traditional knowledge-acquisition bottle-neck: for any formal representation of any domain, one needs to build formal representations of adjoining domains. As has been said, this can be largely overcome by including specialised foundational or domain ontologies in a representation based on the LKIF ontology provided that the quality of these ontologies is sufficient. Depending on availability we might consider providing a library of 'compatible' ontologies to users of LKIF Core. This will be of especial use when the ontology vocabulary will be adopted for expressing the LKIF vendor models that will be developed within ESTRELLA.

With respect to coverage of the legal domain, the purpose of the study outlined in Section 4.1 is more ambitious than only the selection of the most basic terms for describing law, but time and effort constraints make it that we could only consider a small pool of referents. The list of terms will be subjected to a more rigourous empirical study, whereby we will consult a group of legal professionals (taking courses in legal drafting), and law students. These empirical studies are planned in the sideline of ESTRELLA. By applying statistical cluster analysis, we might be able to identify the properties of the scales used (e.g. are they independent?) and whether the statistical clusters have some resemblance to the clusters we have identified based on theoretical considerations. The results of this analysis will be used to evaluate the ontology compared to the requirements we identified in the previous chapters.

The LKIF ontology is available online as separate but interdependent OWL-DL files, and can be obtained from the ESTRELLA website at `http://www.estrellaproject.org/lkif-core`. This website also provides links to online documentation and relevant literature.

## References

Allen, J. (1984). Towards a general theory of action and time. *Artificial Intelligence*, 23:123–154.

[16] Thanks to Taowei David Wang for pointing this out, see `http://lists.owldl.com/pipermail/pellet-users/2007-February/001263.html`

Allen, J. F. and Ferguson, G. (1994). Actions and events in interval temporal logic. *Journal of Logic and Computation*, 4(5):531–579.

Bodenreider, O., Smith, B., and Burgun, A. (2004). The ontology-epistemology divide: a case study in medical terminology. In Varzi, A. and Vieu, L., editors, *Formal ontology in Information Systems*, pages 185–198. IOS-Press, Amsterdam.

Boer, A. (2006). Note on production rules and the legal knowledge interchange format. Technical report, Leibniz Center for Law, Faculty of Law, University of Amsterdam.

Boer, A., Gordon, T. F., van den Berg, K., Di Bello, M., Förhécz, A., and Vas, R. (2007). Specification of the legal knowledge interchange format. Deliverable 1.1, Estrella.

Boer, A., van Engers, T., and Winkels, R. (2005). Mixing legal and non-legal norms. In Moens, M.-F. and Spyns, P., editors, *Jurix 2005: The Eighteenth Annual Conference.*, Legal Knowledge and Information Systems, pages 25–36, Amsterdam. IOS Press.

Breuker, J., Boer, A., Hoekstra, R., and van den Berg, K. (2006). Developing content for lkif: Ontologies and frameworks for legal reasoning. In van Engers, T. M., editor, *Legal Knowledge and Information Systems. JURIX 2006: The Nineteenth Annual Conference*, volume 152 of *Frontiers in Artificial Intelligence and Applications*.

Breuker, J. and Hoekstra, R. (2004a). Core concepts of law: taking common-sense seriously. In Varzi, A. and Vieu, L., editors, *Proceedings of Formal Ontologies in Information Systems (FOIS-2004*, pages 210–221. IOS-Press.

Breuker, J. and Hoekstra, R. (2004b). Epistemology and ontology in core ontologies: FOLaw and LRI-Core, two core ontologies for law. In *Proceedings of EKAW Workshop on Core ontologies*, Http://sunsite.informatik.rwth-aachen.de/Publications/CEUR-WS/. Ceur.

Breuker, J., Hoekstra, R., Boer, A., van den Berg, K., Rubino, R., Sartor, G., Palmirani, M., Wyner, A., and Bench-Capon, T. (2007). OWL ontology of basic legal concepts (LKIF-Core). Deliverable 1.4, Estrella.

Breuker, J., Valente, A., and Winkels, R. (2004). Use and reuse of legal ontologies in knowledge engineering and information management. *Artificial Intelligence and Law*, (to appear in special issue on Legal Ontologies).

Breuker, J. and Van de Velde, W., editors (1994). Common*KADS Library for Expertise Modelling*. IOS Pres.

Casanovas, P., Casellas, N., Vallbe, J.-J., maria Poblet, Benjamins, R., Blazquez, M., Pena, R., and Contreras, J. (2006). Semantic web: a legal case study. In Davies, J., Studer, R., and Warren, P., editors, *Semantic Web Technologies*. Wiley.

Dahllöf, M. (1995). On the semantics of propositional attitude reports.

Dennett, D. (1987). *The Intentional Stance*. MIT-Press.

Donnelly, M. (2005). Relative places. *Applied Ontology*, 1(1):55–75.

Fernández, M., Gómez-Pérez, A., and Juristo, N. (1997). Methontology: from ontological art towards ontological engineering. In *Proceedings of the AAAI97 Spring Symposium Series on Ontological Engineering*, pages 33–40, Stanford, USA.

Gangemi, A., Guarino, N., Masolo, C., and Oltramari, A. (2003). Sweetening WORDNET with DOLCE. *AI Magazine*, 24:13–24.

Gangemi, A., Guarino, N., Masolo, C., Oltramari, A., and Schneider, L. (2002). Sweetening ontologies with DOLCE. In Gomez-Perez, A. and Benjamins, V., editors, *Proceedings of the EKAW-2002*, pages 166–181. Springer.

Gangemi, A., Sagri, M., and Tiscornia, D. (2005). A constructive framework for legal ontologies. In Benjamins, V., Casanovas, P., Breuker, J., and Gangemi, A., editors, *Law and the Semantic Web*, pages 97–124. Springer Verlag.

Genesereth, M. and Fikes, R. (1992). Knowledge interchange format, version 3.0 reference manual. Technical Report Logic-92-1, Computer Science Department, Stanford University.

Gruber, T. (1994). Towards principles for the design of ontologies used for knowledge sharing. In Guarino, N. and Poli, R., editors, *Formal Ontology in Conceptual Analysis and Knowledge Representation*, pages –. Kluwer Academic Publishers. also: Technical Report KSL 93-40, Knowledge Systems Laboratory, Stanford University.

Gruber, T. R. (1993). A translation approach to portable ontology specifications. *Knowledge Acquisition*, 5:199–220.

Grüninger, M. and Fox, M. S. (1995). Methodology for the design and evaluation of ontologies. In *IJCAI'95, Workshop on Basic Ontological Issues in Knowledge Sharing*.

Guarino, N. and Welty, C. (2002). Evaluating ontological decisions with OntoClean. *Communications of the ACM*, 45(2):61–65.

Hayes, P. J. (1985). The second naive physics manifesto. In Hobbs, J. R. and Moore, R. C., editors, *Formal Theories of the Common Sense World*, pages 1–36. Ablex Publishing Corporation, Norwood.

Lakoff, G. (1987). *Women, Fire and Dangerous Things*. University of Chicago Press.

Lame, G. (2006). Using NLP techniques to identify legal ontology components: Concepts and relations. *Artificial Intelligence and Law*. this issue.

Masolo, C., Vieu, L., Bottazzi, E., Catenacci, C., Ferrario, R., Gangemi, A., and Guarino, N. (2004). Social roles and their descriptions. In *Proceedings of Knowledge Representation Workshop*.

Massolo, C., Borgo, S., Gangemi, A., Guarino, N., Oltramari, A., and Schneider, L. (2002). The WonderWeb foundational ontologies: preliminary report. Technical Report Deliverable D17, version 2, ISTC-CNR (Italy).

McCarty, T. (1989). A language for legal discourse I. basic structures. In *Proc. of the Second International Conference on AI and Law*, pages 180–189, Vancouver. Acm.

Minsky, M. (1975). A framework for representing knowledge. In Winston, P. H., editor, *The psychology of Computer Vision*, pages 211–277, New York. McGraw-Hill.

Motta, E. (1999). *Reusable Components for Knowledge Modelling*. FAIA-Series. IOS Pres, Amsterdam NL.

Nuyts, J., Byloo, P., and Diepeveen, J. (2005). On deontic modality, directivity, and mood.

Rubino, R., Rotolo, A., and Sartor, G. (2006). An owl ontology of fundamental legal concepts. In van Engers, T. M., editor, *Legal Knowledge and Information Systems. JURIX 2006: The Nineteenth Annual Conference*, volume 152 of *Frontiers of Artificial Intelligence and Applications*. IOS Press.

Sartor, G. (2006). Fundamental legal concepts: A formal and teleological characterisation. Technical report, European University Institute, Florence / Cirsfid, University of Bologna.

Schank, R. and Abelson, R. (1977). *Scripts, Plans Goals and Understanding*. Lawrence Erlbaum, New Jersey.

Schreiber, G., Akkermans, H., Anjewierden, A., de Hoog, R., Shadbolt, N., Van den Velde, W., and Wielinga, B. (2000). *Knowledge Engineering and Managament: The CommonKADS Methodology*. MIT Press.

Searle, J. and Vanderveken, D. (1985). *Foundations of Illocutionary Logic*. Cambridge University Press, Cambridge.

Sowa, J. F. (2000). *Knowledge Representation: Logical Philosophical, and Computational Foundations*. Brooks Cole Publishing Co, Pacific Grove, CA.

Steimann, F. (2000). On the representation of roles in object-oriented and conceptual modelling. *Data and Knowledge Engineering*, 35:83–106.

Uschold, M. and Grüninger, M. (1996). Ontologies: principles, methods, and applications. *Knowledge Engineering Review*, 11(2):93–155.

Uschold, M. and King, M. (1995). Towards a methodology for building ontologies. In *Workshop on Basic Ontological Issues in Knowledge Sharing, IJCAI-95*, Montreal, Canada.

Valente, A. (1995). *A Modelling Approach to Legal Knowledge Engineering*. PhD thesis, University of Amsterdam.

# Design patterns for legal ontology construction

Aldo Gangemi
*Laboratory for Applied Ontology, ISTC-CNR, Roma, Italy*
*aldo.gangemi@istc.cnr.it*

**Abstract.** Ontology design is known to be a difficult task, requiring much more than expertise in an area; *legal* ontology design, due to the complexity of its domain, makes those difficulties worse. That may be partly due to poor requirement analysis in existing tools, but there is also an inherent gap between the purely logical constructs and methods that are expected to be used, and the actual competences and thought habits of domain experts. This paper presents some solutions, based on *content ontology design patterns*, which are intended to make life of legal ontology designers easier. An overview of the typical tasks and services for legal knowledge is presented, the notion of ontology design pattern is introduced, and some excerpts of a reference ontology (CLO) and its related patterns are included, showing their utility in a simple legal modeling case

**Keywords:** legal knowledge engineering, ontology design, design pattern

## 1. Introduction

A new breed of semantically-explicit applications is getting momentum through the Semantic Web programme and beyond. Legal practice can take advantage from them e.g. in the form of dynamically integrated Semantic Web Services (SWS) (Motta et al., 2003), directed towards citizens, institutions, and companies.

The core of semantically-explicit applications is constituted by so-called *ontologies*, which are strongly-typed logical theories that formalize the assumptions underlying various kinds of knowledge, including physical and social objects, as well as legal procedures, norms, roles, contracts, etc. Ontologies are usually expressed in first-order languages or fragments of them, although some typical modal and meta-level primitives are usually added to them, e.g. in *description logics* like OWL(DL) (McGuinness and van Harmelen (editors), 2004).

Ontologies can be designed by means of various methodologies (e.g. (Gruninger et al., 1994)(Gangemi et al., 2004)), encompassing top-down expertise elicitation from humans, bottom-up learning from documents, and middle-out application of content patterns (specialized from domain-independent ontologies, elicited in a top-down way, or learnt from patterns found in experts' documents), which can be called *Content Ontology Design Patterns* (CODeP, also known as "conceptual" ontol-

ogy design patterns) (Gangemi et al., 2005). In large-scale, realistic applications, CODePs are core components for ontology design.

The legal domain is very complex compared to others, because it involves knowledge of the physical and social worlds, as well as typical legal knowledge that actually creates a novel layer over the social world (Moore, 2002).
Due to the autonomy (on one hand) and dependence (on the other hand) of the legal knowledge on both physical and social knowledge, legal reasoning tasks have evolved in a peculiar way, which include e.g. the norm structure based on CODePs like *Requirement→Consequence* (if the factual knowledge is $P$, then the legal knowledge is $Q$), *Obligation→Right* (if $A$ has an obligation towards $B$, then $B$ has a right towards $A$), *Norm↔Case* (if a situation fulfils a the conditions for violating a norm, it becomes a legal case), *CrimeScenario* (a crime is committed by a perpetrator and comes to the attention of authorities that pursue a criminal process), etc. The CODePs that are assumed by legal experts can be formalized by specializing or composing other existing patterns for the social world.

This work introduces some use cases for legal ontologies, as well as some CODePs that can be specialized to support ontology-driven solutions to those use cases. The CODePs have been defined on top of the DOLCE foundational ontology library (Masolo et al., 2004)(DOLCE, -), the Core Legal Ontology (CLO) (Gangemi, Sagri and Tiscornia, 2004) (CLO, -), and JurWordNet (Sagri, 2003).

In section 1. some ontology design/engineering use cases in the legal domain are introduced. In section 2. the CODeP idea is presented. In section 3. The Core Legal Ontology is briefly summarized and some more legal CODePs sketched with an example on a use case.

## 2. Legal Ontology Engineering: Functionalities and Techniques

Within ICT, ontology design is dependent on (ontology) engineering applications, which involve the statement of functionalities and their implementation as techniques and tools. For a comprehensive framework of ontology design, and its relations to content, related data, formal languages, design patterns, social practices, organizations, teams, and functionalities, see (Gangemi et al., 2007). The ontology encoding of a metamodel for describing collaborative ontology design activities and data can be found at: `http://www.loa-cnr.it/ontologies/OD/codolib.owl`.

Ontology engineering deals with designing, managing, and exploiting ontologies (to be intended as *strongly-typed logical theories*) within information systems. Ontologies are usually hybridated with other components in order to build semantically-explicit applications; e.g., when used jointly with:

- theorem provers, *consistency checking* can be performed to logically validate the set of assumptions encoded in an ontology
- subsumption and instance classifiers against a logical language of known and manageable complexity, like OWL (in the Lite and DL species), *automatic inferences* can be derived from *taxonomical reasoning* as well as for the *classification of instances and facts* (Gangemi et al., 2001) (Gangemi, Sagri and Tiscornia, 2004)
- computational lexicons, NLP tools, and machine-learning algorithms, legal ontologies can enhance *information extraction* from semi-structured and non-structured data, adding a new dimension to knowledge management and discovery in Law (Gilardoni et al., 2005)
- planning algorithms, ontologies can assist or automatize *negotiation* or *execution* e.g. for *contracts, regulations*, *services*, etc. (Gil et al., 2005)
- case-based reasoners, ontologies can *formalize case abstractions* within more general frameworks, or can *classify cases* according to predesigned descriptions (Forbus et al., 2002)
- rule-based engines, facts can be inferred e.g. for *causal responsibility assessment*, *conformity checking, conflict detection* and in general for *fact composition* (Gangemi et al., 2001).

Ontology engineering techniques are exploited in the context of "generic use cases" defined for a domain of application. The main types of use cases that can be implemented or assisted by means of semantically-explicit applications in the legal domain (Fig. 1) are summarized in the following.

*Intersubjective agreement and meaning negotiation*
*Definition*: the task of getting consensus (or of discovering disagreement) about the intended meaning of a legal term, legal text unit, etc.
*Approach*: the formal encoding of (part of) the intended meaning assumed by each of the parties involved in the task.
*Issues*: given the traditional practices of consensus reaching in Law,

this task is usually considered intrusive, and could require a mindset shifting in order to acquire some relevance. Nonetheless, encoding intended meaning of a legal text is preliminary to all other tasks presented here. This observation seems paradoxical, since, if we cannot consider the formal encoding of legal meaning as an interpretation with legal validity, all the other tasks result to be based on an arbitrary (in the worst case) or a weak (in the best case) set of assumptions. Due to the current state of legal ontology, most tasks are carried out *as if* that formal encoding had legal validity, thus providing results that can be considered only as heuristical means for legal professionals or citizens.

*Knowledge extraction*
*Definition*: the process of extracting concepts, relations, named entities, and complex knowledge patterns from a database, a document, or a corpus.
*Approach*: data- and text-mining, machine-learning, and NLP algorithms that can extract linguistic objects from a corpus, and semi-automated methods that match them to semantic objects.
*Issues*: this task is highly incremental, because the approaches need a *training* phase or an extensive *data entry* procedure, so that the extracted knowledge can be used to build a repository of patterns that can be used to improve further extraction processes. Best results can be achieved on very large corpora (for statistical reasons), or on well-delimited, possibly semi-structured corpora and tasks; for example, typical expressions that are found in legal drafting can be used to formalize expectation patterns in corpora consisting of homogeneous texts (Basili et al., 2005).

*Conformity checking*
*Definition*: the task aimed at verifying if a social situation (known in some way compliant with legal regulations) satisfies a legal description (norm, principle, regulation, etc.). In the generalized case, also situations already known to be legally relevant for some reason (e.g. a crime situation) can be checked for conformity against a further legal description (e.g. an appeal judgment procedure).
*Approach*: the representation of social or legal situations as well as of legal regulations. Reasoners to cluster/classify/abstract situations.
*Issues*: Reasoning: the typical inferences supported by semantic web engines for OWL (McGuinness and van Harmelen (editors), 2004) (e.g. FaCT++ (Horrocks, )) currently include only *concept subsumption* and *instance classification*. The expressivity is also limited so that e.g. *fact classification* (also called *materialization*) is only performed by some engines (e.g. Pellet (Sirin et al., 2007)). Moreover, *compositions bet-*

*ween facts* cannot be inferred unless an additional rule engine is added (extensions for rule languages supporting fact composition are provided by additional languages: SWRL(Horrocks et al., 2005), SPARQL (Prud'hommeaux et al., 2005), F-Logic (Hustadt et al., 2004), and their related implementations, e.g. KAON2 (Hustadt et al., 2004). Moreover, clustering and abstraction of situations requires different reasoners, e.g. induction engines (Basili et al., 2005) and case-based reasoners (Forbus et al., 2002). Finally, approximate inferences (Domingos and Richardson, 2004) should also be supported in the generalized case of partial knowledge about situations.

Representation: a homogeneous language to represent both situations and the constraints from a legal description is highly desirable, otherwise a higher-order logic would be required to express constraints on contraints on constraints etc. on situations. The proposal in (Gangemi, Sagri and Tiscornia, 2004), briefly summarized in next sections, shows a viable approach to represent both constraints and instance data in a same, partitioned first-order domain of quantification.

*Legal advice*

*Definition*: the investigation of the relations between legal cases and common sense situations.

*Approach*: subsumption/instantiation classification, or case-based reasoning.

*Issues*: in large scale applications, legal advice involves crucial problems such as causality and responsibility assessment, open-textured concepts, interpretation aspects, which are still being investigated from an ontological perspective. A typical scope for legal ontology design is to encode only *weak constraints* for terminological clarification. Legal advice requires more than that.

*Norm comparison*

*Definition*: the matchmaking between different norms. Norm comparison includes tasks such as: (i) *normative conflict checking* and handling between norms about a same situation type, (ii) *discovery of implicit relations* between a norm and other norms from a known corpus.

*Approach*: approximate classification algorithms (i-ii), including legal text annotation and classification on large corpora (ii). (Gangemi, Sagri and Tiscornia, 2004) and (Gangemi et al., 2001) show simpler approaches to the classification of norm dynamics and conflicts within a finite set of norms after their first-order encoding.

*Issues*: in Civil Law corpora, the task (ii) is sometimes relevant as much as in Common Law corpora, because of the stratification of laws that do not explicitly delete or even refer to previous ones (e.g.

in Italy). In Common Law, implicit relations can be discovered more easily, because case abstraction has always a clear reference to a case, while in Civil Law, implicit relations appear at the purely normative level.

*Norm rephrasing*
*Definition*: expression of norms' content in different terms, which can be either translations in a different natural language, or in a different form within a same natural language, e.g. for the purpose of popularization.
*Approach*: translation between different languages requires a preliminary mapping between terms, like the EuroWordNet-oriented work performed in the LOIS project (Peters et al, 2006), classifications based on statistical NLP techniques, and subsumption classification for a close matching between content patterns and linguistic patterns.

*Contract management and execution*
*Definition*: a service assisting parties in the tasks of managing contract agreement and definition, and of following contract execution.
*Approach*: the semantic specification of contract content, as well as algorithms to manage the matching of parties' constraints and preferences, and a planning algorithm for the generation of optimal obligations that parties could undertake (Gil et al., 2005).

*(Information) service matchmaking and composition*
*Definition*: operations carried on the description of services, in order to check e.g. if an offered service matches the requested service, or to orchestrate two services to get a more complex one.
*Approach*: in the legal domain, these tasks require the semantic specification of services with reference to the legal knowledge involved in the execution of the service. Appropriate reasoners and planners are required.


*Figure 1.* A taxonomy of ontology-driven tasks and related techniques for legal information

## 3. Content Ontology Design Patterns

Semantically-explicit applications in the legal domain present us with conceptual analysis and integration problems that require appropriately designed legal formal ontologies. Part of the design problems can be simplified by creating or extracting "Conceptual Ontology Design Patterns" (CODeP) for a domain of application (Gangemi, Sagri and Tiscornia, 2004) (Gangemi et al., 2004). An intuitive characterization of CODePs is provided here:

– A CODeP is a template to represent, and possibly solve, a modelling problem. For example, a *Norm* ↔ *Case* CODeP (Fig.3) facilitates the modelling of legal norms and cases (as well as their components and dependencies) in logical languages that require constraint reification. E.g. in OWL(DL), relations with an arity =2 are not allowed, therefore OWL(DL) modelling requires a reifi-

cation of those relations. A vocabulary for reification has been designed in the Descriptions and Situations ontology ("ExtendedDnS" in Fig. 4, see below), which is specialized in the $Norm \leftrightarrow Case$ pattern.

- A CODeP "extracts" a *connected fragment* of a reference ontology, which constitutes its "background". For example, a connected path of two relations and three classes ($A(x) \ni B(y) \ni C(z) \ni R(x,y) \ni S(y,z)$) can be extracted as a sub-theory of an ontology $O$ because of its relevance in a domain. Therefore, a CODeP lives in a reference ontology, which provides its taxonomic and axiomatic context. E.g. in the $Norm \leftrightarrow Case$ CODeP a foundational distinction is reused from DOLCE (Masolo et al., 2004), while the cardinalities for the relations are provided by the Core Legal Ontology (CLO, (Gangemi, Sagri and Tiscornia, 2004) (CLO, -)). DOLCE and CLO together form the reference ontology for the CODeP.

- Mapping and composition of patterns require a reference ontology, in order to check the consistency of the composition, or to compare the sets of axioms that are to be mapped. Operations on CODePs depend on operations on the reference ontologies. However, for a pattern user, these operations should be (almost) invisible.

- A CODeP can be represented in an ontology representation language whatsoever (depending on its reference ontology), but its intuitive and compact visualization is an essential requirement. It requires a critical size, so that its diagrammatical visulization is aesthetically acceptable and easily memorizable. For example, the $Norm \leftrightarrow Case$ CODeP only includes eight classes, with several, systematic relations between them: this makes it *dense*, but *manageable*.

- A CODeP can be an element in a partial order, where the ordering relation requires that at least one of the classes or relations in the pattern is specialized. A hierarchy of CODePs can be built by specializing or generalizing some of the elements (either classes or relations). For example, the *participation* pattern (of an object in an event) can be specialized to the *taking part in a public enterprise* pattern (of an agent in a social activity with public relevance).

- A CODeP should be intuitively exemplified, and should catch relevant, "core" notions of a domain. Independently of the generality at which a CODeP is singled out, it must contain the central notions and best practices that "make rational thinking move" for an expert in a given domain for a given task.

– A CODeP can be often built from informal or simplified schemata used by domain experts, together with the support of other reusable CODePs or reference ontologies, and a methodology for domain ontology analysis. Typically, experts spontaneously develop schemata to improve their business, and to store relevant know-how. These schemata can be reengineered with appropriate methods (e.g. (Gangemi et al., 2004)).

– A CODeP can be similar to a database schema, but a pattern is defined wrt to a reference ontology, and has a general character, independently of system design. In this sense, it is closer to so-called *data modelling patterns* (Hay, 1996), but a CODeP should be contextualized in a reference ontology, making it more interoperable than a data modelling pattern, at least in principle.

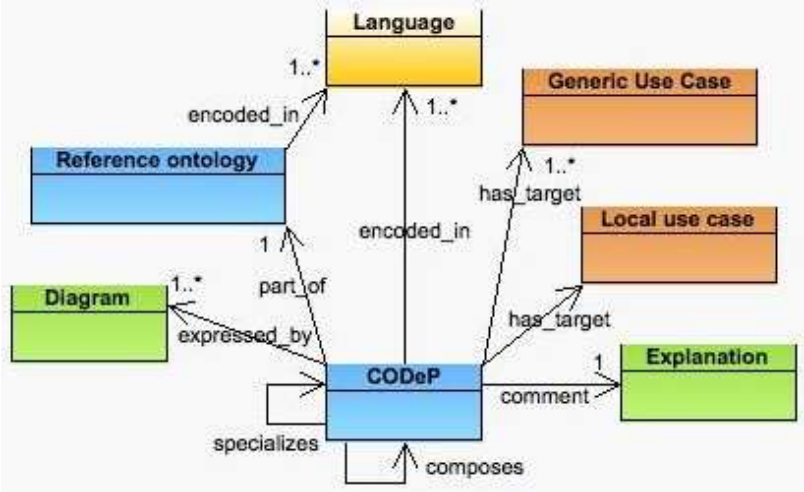



*Figure 2.* The CODeP annotation pattern

Conceptual Ontology Design Patterns (CODePs) are a resource and design method for engineering ontology content over the Semantic Web.

A template (Fig. 2, also available in OWL from (CODeP, )) can be used to annotate CODePs as sub-theories of reference ontologies, in order to share them in pre-formatted documents, as well as to describe, visualize, and make operations over them them appropriately.

The CODeP template consists of:

– Two slots for the *generic use case*, and the *local use cases*, which includes a description of context, problem, and constraints/requirements.

- Two slots for the addressed *logic*, and the *reference ontologies* used as a background for the pattern.

- Two slots for -if any- the *specialized* pattern and the *composed* patterns (by inheritance, and inverse inheritance, it's possible to obtain the closure of specialized and expanding patterns).

- Two slots for the *maximal relation* that encodes the case space, and its intended *axiomatization*: a full first-order logic with meta-level is assumed here, but the slot can be empty without affecting the functionality of a CODeP frame.

- Two slots for *explanation* of the approach, and its *encoding* in the logic of choice.

- A last slot for a *class diagram* that visually reproduces the approach.

The template can be easily encoded in XSD or in richer frameworks, like semantic web services (e.g. Motta et al. 2003) or knowledge content objects (Behrendt et al. 2005), for optimal exploitation within Semantic Web technologies. The high reusability of CODePs and their formal and pragmatic nature make them suitable not only for isolated ontology engineering practices, but ideally in distributed, collaborative environments like intranets, the Web or the Grid.

CODePs can also be used to generate intuitive, friendly UIs, which can present the user with only the relevant pattern diagram, avoiding the awkward, entangled graphs currently visualized for medium-to-large ontologies.

The advantages of CODePs for ontology lifecycle over the Semantic Web are straightforward: firstly, patterns make ontology design easier for both knowledge engineers and domain experts (imagine having a menu of pre-built, formally consistent components, pro-actively suggested to the modeller); secondly, patterned design makes it easier ontology mapping - perhaps the most difficult problem in ontology engineering. For example, the *time-indexed participation* presented in this paper requires non-trivial knowledge engineering ability to be optimally represented and adapted to a use case: a CODeP within an appropriate ontology management tool can greatly facilitate such representation.

The CODeP examples and the related frame and methods introduced in this paper have been applied for two years (some of them even before) in several administration, business and industrial projects, e.g. in fishery information systems (Gangemi et al., 2004), insurance CRM, biomedical

ontology integration (Gangemi et al., 2004), anti-money-laundering systems for banks (Gangemi et al., 2001), service-level agreements for information systems, biomolecular ontology learning (Ciaramita et al 2005), legal norms formalization (Gangemi, Sagri and Tiscornia, 2004)(Sagri et al., 2004).

Current work focuses on building a tool that assists development, discussion, retrieval, and interchange of CODePs over the Semantic Web, and towards establishing the model-theoretical and operational foundations of CODeP manipulation and reasoning. In particular, for CODePs to be a real advantage in ontology lifecycle, the following functionalities should be available:

- Categorization of CODePs, based either on the use cases they support, or on the concepts they encode.

- Pattern-matching algorithms for retrieving the CODeP that best fits a set of requirements, e.g. from a natural language specification, or from a draft ontology.

- Support for specialization and composition of CODePs. A CODeP $p_2$ *specializes* another $p_1$ when at least one of the classes or properties from $p_2$ is a sub-class or a sub-property of some class resp. property from $p_1$, while the remainder of the CODeP is identical. A CODeP $p_2$ *expands* $p_1$ when $p_2$ contains $p_1$, while adding some other class, property, or axiom. A CODeP $p_3$ *composes* $p_1$ and $p_2$ when $p_3$ contains both $p_1$ and $p_2$. The formal semantics of these operations is ensured by the underlying (reference) ontology for the patterns. Notice that CODePs –differently from "knowledge patterns" in (Clark et al., 2000), which are characterized as *invariant under signature transformation*– are intended to be *downward conservative under signature transformation*, meaning that a pattern semantics is structure-preserving when the pattern is *specialized, expanded,* or *composed*, but this conservativeness holds only in the downward taxonomical ordering. On the contrary, *logical* ontology design patterns are conservative under signature transformation both down- and up-wardly.

- Interfacing of CODePs for visualization, discussion, and knowledge-base creation.

- A rich set of metadata for CODeP manipulation and exploitation within applications.

## 4. The Core Legal Ontology and its Related Patterns

The need for an extended typology of legal entities is becoming a pressure, even from traditionally "bottom-up" approaches. For example, the need to pair case-based reasoning with an ontology of first-principles should be investigated in order to represent the two kinds of structures employed in reasoning: abstraction from cases, and satisfaction of constraint sets (e.g. norms) (Forbus et al., 2002).

The level of granularity is also a core issue in developing formal ontologies, specially because a decentralized architecture is emerging for ontologies as well: how to compare/integrate/transform two ontologies about a close domain, but with a different detail encoded in their vocabulary and axioms?

The Core Legal Ontology (CLO) (Gangemi, Sagri and Tiscornia, 2004) (CLO, -) is developed on top (Fig.3) of *DOLCE* (Masolo et al., 2004) and *Descriptions and Situations* (Gangemi and Mika, 2003) (Masolo et al., 2004) ontologies within the DOLCE+ library (DOLCE, -). CLO allows for the representation of first principles (by means of a rich axiomatization), and granularity (by means of its reification vocabulary and axioms) in the legal domain.

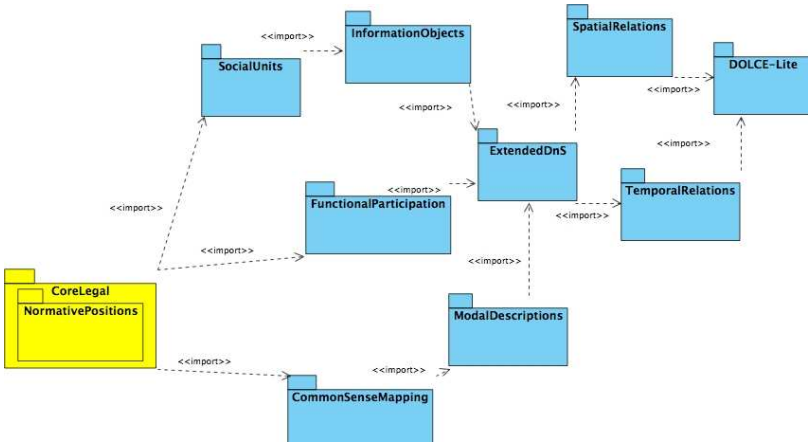



*Figure 3.* CLO depends on other ontologies: DOLCE, ExtendedDnS, InformationObjects, Temporal and Spatial relations, etc. The *Norm↔Case* CODeP is a component within the CoreLegal (CLO) module.

The two pillars of CLO as a plugin to DOLCE+ are: *stratification* and *reification*.

Based on the stratification principle, CLO provides types and relation for the heterogeneous entities from the legal domain, be it about the physical, cognitive, social, or properly legal worlds (cf. (Moore,

2002) (Gangemi, Sagri and Tiscornia, 2004)). According to stratification, entities from different layers can be *spatio-temporally co-located*, yet being completely different and (mutually or one-way) dependent. For example, a physical person pertains to the physical world as a biological organism, but the properties of the organism are not sufficient to characterize it as a social person. On its turn, the properties of a social person are not sufficient to characterize it as a legal person. Clearly, there are dependencies among those properties, but if the properties from each layer are simply summed up in a same entity, an ontology designer can get undesirable results, e.g. it would be possible to infer that an organism (physical layer) can be "acting" after its death because of the legal existence (legal layer) of a person until its legal effects disappear.

In DOLCE, the solution includes a very general pattern, expressed as a disjointness axiom between the class of physical vs. social objects, whereas legally-relevant entities are mostly in the social realm (see Fig.s 5,6,7,8,9, which also include some use of CLO for the JurWordNet lexical ontology (Sagri, 2003) (Gangemi, Sagri and Tiscornia, 2004)). Among social objects, *agentive* and *non-agentive* are disjoint, and among non-agentive ones, *legal descriptions, concepts, facts, collectives, persons*, and *information objects* are also disjoint. Among legal descriptions, *constitutive* vs. *regulative* descriptions are distinguished from *principles, rationales, modal descriptions* (e.g. *duties, powers, liabilities,* etc.), as well as mixed regulations such as contracts and bundles of norms. Among legal facts, *natural, human, cognitive,* and strictly *legal cases* are distinguished. Legally-relevant *circumstances* are further distinguished from legal facts as being ancillary to primary facts. Among legal persons, *organizations*, *natural persons*, and *legal subjects* are also distinguished.

Based on the reification principle, CLO enables an ontology engineer to quantify either on legal rules or relations (type reification) (Masolo et al., 2004) (Galton, 1991), or on legal facts (token reification) (Gangemi and Mika, 2003) (Galton, 1991). CLO extends the *Descriptions and Situations* vocabulary for reification. For example, *intensional specifications* like norms, contracts, subjects, and normative texts can be represented in the same domain as their *extensional realizations* like cases, contract executions, agents, physical documents.

A reification-based pattern models the structure of an intensional specification, called *Description* in (Masolo et al., 2004), as composed of its concepts and their internal dependencies. For example, the structure of a norm (a *legal description*) employs that pattern. Another pattern models the matching between a description and its extensional realization, called a *Situation* in (Gangemi and Mika, 2003), which can

be described as the configuration of a set of entities according to the structure of a description. A legal application of this pattern can be found in the dependencies among the *rules in a contract*, when they can be matched to a *legal case* (a legal situation or fact) or a *contract execution*. The matching is typically performed when checking if each entity in a legal fact is compliant to a concept in a legal description. CLO is currently used to support the definition of legal domain ontologies (Sagri et al., 2004), the definition of a juridical wordnet (Sagri, 2003), and the design of legal semantic web services.

In order to describe some of the ontological expressiveness of CLO, a complex CODeP is introduce here which has CLO as its reference ontology. This is is the $Norm \leftrightarrow Case$ CODeP (Fig. 4): it is used as a pattern for representing legal cases, is specialized for types of norms and cases, e.g. for crime investigation, and is composed with other patterns, e.g. for norm conflict checking.

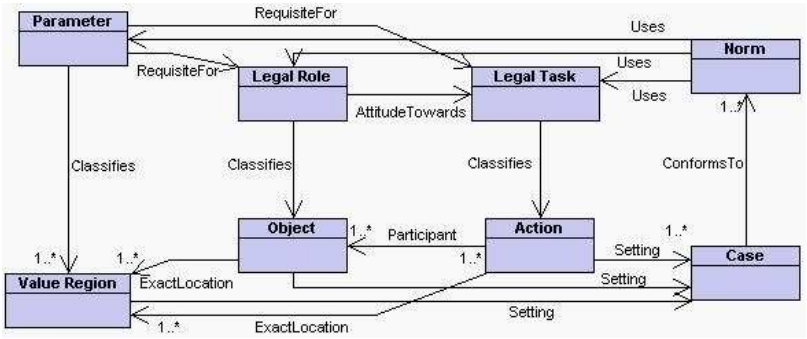
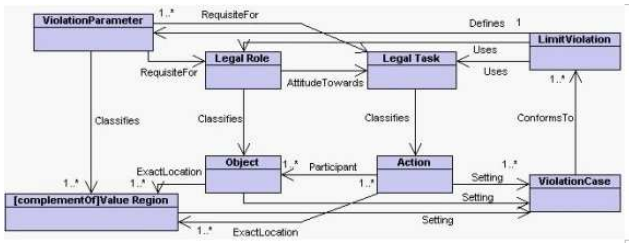



*Figure 4.* The $Norm \leftrightarrow Case$ CODeP: norms *use* tasks, roles, and parameters; legal cases *conform* to norms when actions, objects and values are *classified* by tasks, roles, and parameters respectively. Moreover, relations between legal roles, tasks and parameters correspond to relations between objects, actions and values. For example, an *obligation* for a role towards a task should correspond to a *participation* of an agent (object) in an action; a spatial parameter that is *requisite* for an object should correspond to an *exact location* of an object in a spatial value region that is *classified* by that parameter.

In the following template (Tab. I, see Fig. 2 for its datamodel), the $Limit \leftrightarrow ViolationCase$ CODeP is compactly introduced as a specialization of the generic $Norm \leftrightarrow Case$ CODeP:

Table I.

| **Slot** | **Value** |
|---|---|
| Generic use case | Legal situations to be checked for (non)conformity to existing norms that establish limits. |
| Local use cases | Legally-defined roles, functions, and parametric values exist to control the social life of a country. When talking about agents and social action, there is a network of senses implying a dependence on roles, functions (or tasks), and parametrized value ranges within a normative description. Intended meanings include the possible roles played by certain objects and agents, the actual actions occurring during social life, as well as parametric limits over value types, such as age limits, speed limits, etc. Therefore, both class- and instance-variables are present in the maximal relation for this pattern. |
| Logic addressed | OWL, DL species |
| Reference ontologies | DOLCE-Lite-Plus, Core Legal Ontology |
| Specialized CODeP | $Norm \leftrightarrow Case$ |
| Composed CODePs | Time-Indexed-Participation, Concept↔Description, Description↔Situation |
| Formal relation | CaseConformsToLimitViolation($\varphi$,$\psi$,$\chi$,$x$,$y$,$z$,$t$,$c_1$,$c_2$,$c_3$,$d$,$s$), where $\phi(x)$ is an agent class, $\psi(y)$ is a process class, $\chi(z)$ is a class of values within a value range, t is a time interval, $c_1$, $c_2$ and $c_3$ are three reified intensional concepts, d is a reified intensional relation, and s is a reified extensional relation. |
| Sensitive axioms | CaseConformsToLimitViolation(s,d) $=_{df}$ $\forall x, y, z(\varphi(x) \wedge \psi(y) \wedge \chi(z) \wedge$ participantIn($x$,$y$,$t$) $\wedge$ locationOf($z$,$y$,$t$) $\wedge$ (Object($x$) $\vee$ Agent($x$)) $\wedge$ Action($y$) $\wedge$ Speed($z$) $\wedge$ TimeInterval($t$)) $\leftrightarrow$ $\exists c_1,c_2,c_3$(CF($x$,$c_1$,$t$) $\wedge$ MT($c_1$,$c_2$) $\wedge$ CF($y$,$c_2$,$t$) $\wedge$ $\neg$CF($z$,$c_3$,$t$) $\wedge$ REQ($c_3$,$c_2$) $\wedge$ (DF($d$,$c_1$) $\wedge$ DF($d$,$c_2$) $\wedge$ DF($d$,$c_3$) $\wedge$ $\forall s$(SAT($s$,$d$) $\leftrightarrow$ (SETF($s$,$x$)$\wedge$ SETF($s$,$y$) $\wedge$ SETF($s$,$t$))) |
| Explanation | Since OWL(DL) does not support relations with >2 arity, reification is required. The Description↔Situation pattern provides typing for such reification. Since OWL(DL) does not support classes in variable position, we need reification for class-variables. The Concept Description pattern provides typing for such reification. Similarly, since participation is time-indexed, we need the time-indexed-participation pattern, which is here merged with the previous two patterns (time indexing appears in the setting of the general normative situation). |
| OWL(DL) encoding (abstract syntax) | Class(CaseConformsToLimitViolation complete<br>Description<br>restriction(defines someValuesFrom(Object))<br>restriction(defines someValuesFrom(Action))<br>restriction(defines someValuesFrom(TimeInterval)))<br>Class(LegalAgent complete<br>Role<br>restriction(used-by someValuesFrom(CaseConformsToLimitViolation))<br>restriction(classifies allValuesFrom(Object))<br>restriction(attitude-towards someValuesFrom(LegalTask)))<br>Class(LegalTask complete<br>Task<br>restriction(used-by someValuesFrom(CaseConformsToLimitViolation))<br>restriction(classifies allValuesFrom(Action))<br>restriction(attitude-target-of someValuesFrom(LegalAgent)))<br>Class(LimitViolation complete<br>Parameter<br>restriction(used-by someValuesFrom(CaseConformsToLimitViolation))<br>restriction(classifies allValuesFrom(complementOf LimitValueRegion))<br>restriction(requisite-for someValuesFrom(unionOf(LegalTask LegalAgent)))<br>Class(ViolationCase complete<br>Situation<br>restriction(satisfies someValuesFrom(CaseConformsToLimitViolation))<br>restriction(setting-for someValuesFrom(Object))<br>restriction(setting-for someValuesFrom(Action))<br>restriction(setting-for someValuesFrom(Limit))<br>restriction(setting-for someValuesFrom(Time-Interval))) |
| Class diagram |  |

## 5. Conclusions

An overview of the relevance of ontology patterns for legal knowledge engineering (LKE) has been presented.

For some generic LKE use cases (conflict checking, information extraction, etc.), some solutions and issues, both on the reasoning and (content) modelling sides, have been mentioned. While the reasoning side of LKE is a fast-moving target, with interesting solutions coming from e.g. hybridating different inference engines and classifiers, the modelling side is far less developed, despite the huge literature that focuses on legal and jurisprudential content, let alone the work in formal ontology and beyond. This is not surprising. Currently, very few ontologies are actually *reused*, against the great expectations that have been grown in the field of reusability of semantic components.

In the past, the need for structures that grant systematicity to content modelling have mainly focused on so-called *top-level* ontologies, but the power of a small set of categories is not enough for realistic ontology projects like those presented to LKE. A top-level can even be a problem when its categories are brittle with respect to the domain task.

A more sophisticated approach, which ensures a much higher level of cognitive interoperability, is constituted by *foundational* ontologies like DOLCE. Nonetheless, although their rich axiomatization makes foundational ontologies ideally suited for building a partial-order of reference ontologies, they require a substantial cognitive load to be accessed and then successfully reused.

A new dimension of reusability has been introduced in ontology engineering (and here extended to LKE) which revisits some good practices from AI (*knowledge patterns*) and databases (*data model patterns*), by providing a meta-model, some operations, and a generalized characterization to *conceptual ontology design patterns* (CODePs). A foundation to CODePs is supposed to enhance the construction of tools for ontology design, as e.g. envisaged in the NeOn project (NeOn, ), and to facilitate the collaborative and distributed negotiation of meaning across the members of a same or of sufficiently close communities.

For LKE, CODePs appear slightly more complex than in other domains, mostly because the use cases in LKE involve a layering of meaning (from the physical to the social, cognitive, and finally legal realms), which also require an extended reification of entities such as norms, contracts, cases, legal text corpora, etc. A complex network of dependencies between roles and tasks, agents, *normative positions*, validity parameters, assumed goals, cognitive attitudes, etc. makes the modelling task in LKE harder than elsewhere. A few steps toward the creation of a

repository of legal CODePs have been sketched, together with some possible directions for further development.

## References


Gangemi, A., Sagri M.T., Tiscornia D.: A Constructive Framework for Legal Ontologies. In R. Benjamins, P. Casanovas, J. Breuker, A. Gangemi (eds.): Law and the Semantic Web, Springer, Berlin (2004).

Gangemi, A., Borgo, S. (eds.): Proceedings of the EKAW*04 Workshop on Core Ontologies in Ontology Engineering. `http://sunsite.informatik.rwth-aachen.de/Publications/CEUR-WS//Vol-118/` ((2004)).

Gangemi, A., F. Fisseha, J. Keizer, J. Lehmann, A. Liang, I. Pettman, M. Sini, M. Taconet: A Core Ontology of Fishery and its Use in the FOS Project, in (Gangemi and Borgo, 2004) (2004).

Masolo, C., A. Gangemi, N. Guarino, A. Oltramari and L. Schneider: WonderWeb Deliverable D18: The WonderWeb Library of Foundational Ontologies (2004).

Sagri M.T., Tiscornia D., Gangemi A., An Ontology-based Approach for Representing "Bundle-of-rights", M. Jarrar & A. Gangemi (eds.), 2nd International Workshop on Regulatory Ontologies, in OTM Workshops, Springer, 2004.

Sagri M.T.: Progetto per lo sviluppo di una rete lessicale giuridica on line attraverso la specializzazione di ItalWordnet, in Informatica e Diritto, ESI, Napoli, 2003.

Motta, E., Domingue, J., Cabral, L., Gaspari, M. (2003) IRS-II: A Framework and Infrastructure for Semantic Web Services. 2nd International Semantic Web Conference (ISWC2003) 20-23 October 2003, Florida, USA (2003).

Gruninger, M., and Fox, M.S.: The Role of Competency Questions in Enterprise Engineering. Proceedings of the IFIP WG5.7 Workshop on Benchmarking - Theory and Practice, Trondheim, Norway (1994).

Masolo, C., L. Vieu, E. Bottazzi, C. Catenacci, R. Ferrario, A. Gangemi and N. Guarino: Social Roles and their Descriptions. In C. Welty (ed.): Proceedings of the Ninth International Conference on the Principles of Knowledge Representation and Reasoning, Whistler (2004).

Gangemi, A., Catenacci, C., Battaglia, M.: Inflammation Ontology Design Pattern: an Exercise in Building a Core Biomedical Ontology with Descriptions and Situations. D.M. Pisanelli (ed.) Ontologies in Medicine, IOS Press, Amsterdam (2004).

Galton, A.: Reified Temporal Theories and How To Unreify Them. Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI), 1991.

Clark, P., Thompson, J., Porter, B.: Knowledge Patterns. Proceedings of KR00 (2000)

Gangemi, A., Ontology Design Patterns for Semantic Web Content. Y. Gil, E. Motta, R. Banjamins, M. Musen, Proceedings of $4^{th}$ International Semantic Web Conference, ISWC05 (2005).

McGuinness D.L. and van Harmelen F. (editors), OWL Web Ontology Language Overview, W3C Recommendation, 10 February 2004, `http://www.w3.org/TR/2004/REC-owl-features-20040210/` (2004).

Moore M.S. Legal Reality: A Naturalist Approach to Legal Ontology. *Law and Philosophy* 21: 619–705, 2002.

Forbus, K., Mostek. T., Ferguson, R., An analogy ontology for integrating analogical processing and first principles reasoning, in Proceedings of AAAI02, 2002.

DOLCE + library, available in OWL from: http://dolce.semanticweb.org Directly loadable from: http://www.loa-cnr.it/ontologies/DLP.owl.

Core Legal Ontology, loadable from: http://www.loa-cnr.it/ontologies/CLO/CoreLegal.owl

Gangemi, A., Pisanelli, D.M., Steve G. A Formal Ontology Framework to represent Norm Dynamics. Proc. Of Second International Workshop on Legal Ontologies, Amsterdam, 2001.

Gil, R., Garcia, R., Delgado, J. An interoperable framework for IPR using web ontologies, J. Lehmann, E. Biasiotti, E. Francesconi, M.T. Sagri (eds.), Proceedings of the First LOAIT Workshop, Bologna, Italy (2005).

Evren Sirin and Bijan Parsia and Bernardo Cuenca Grau and Aditya Kalyanpur and Yarden Katz Pellet: a practical owl-dl reasoner, Available at http://www.mindswap.org/papers/PelletJWS.pdf

Gilardoni, L., Biasiuzzi C., Ferraro, M, Fonti, R., Slavazza, P. LKMS – A Legal Knowledge Management System Exploiting Semantic Web Technologies. Y. Gil, E. Motta, R. Banjamins, M. Musen (eds.), Proceedings of $4^{th}$ International Semantic Web Conference, ISWC05 (2005).

Prud'hommeaux E., Seaborne A. SPARQL Query Language for RDF, http://www.w3.org/TR/rdf-sparql-query/

Ian Horrocks, Peter F. Patel-Schneider, Harold Boley, Said Tabet, Benjamin Grosof, Mike Dean SWRL: A semantic web rule language combining owl and ruleml, W3C Member Submission, 21 May 2004. Available at http://www.w3.org/Submission/SWRL/.

Basili, R., Pennacchiotti, M., Zanzotto, F.M. Language Learning and Ontology Engineering: an integrated model for the Semantic Web. B. Magnini (ed.), Proceedings of Meaning Workshop (2005).

Horrocks, I. Fact++ web site. http://owl.man.ac.uk/factplusplus/

Hustadt, U., Motik, B., Sattler, U. Reducing SHIQ Description Logic to Disjunctive Datalog Programs. Proc. of the 9th International Conference on Knowledge Representation and Reasoning (KR2004), June 2004, Whistler, Canada, pp. 152-162.

Domingos, P., Richardson, M. Markov Logic: A Unifying Framework for Statistical Relational Learning. Proceedings of the ICML-2004 Workshop on Statistical Relational Learning and its Connections to Other Fields (pp. 49-54), 2004. Banff, Canada: IMLS.

Peters W., Sagri M.T., Tiscornia D. and Castagnoli S. The LOIS Project. LREC 2006, Genova, 2006.

Hay, D.C. Data Model Patterns. Conventions of Thought. Dorset House, NY (1996).

CODeP Repository, available from http://www.loa-cnr.it/codeps/index.html

Gangemi A., Mika P. Understanding the Semantic Web through Descriptions and Situation, Meersman R, et al. (eds.), Proceedings of ODBASE03, Springer, Berlin, 2003.

NeOn Project EU-IST 27595. http://www.neon-project.org.

A. Gangemi, V. Presutti, C. Catenacci, J. Lehmann, M. Nissim, C-ODO: an OWL meta-model for collaborative ontology design. Noy et al. (eds.): Proceeding of First CKC workshop, WWW2007 Conference, Banff, 2007.

## Appendix: some excerpts from the Core Legal Ontology

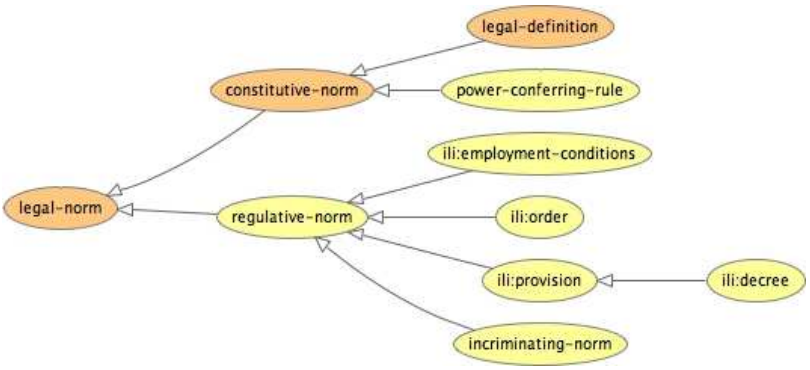



*Figure 5.* An excerpt of CLO taxonomy: legal norm types. "ili" classes are from the JurWordNet ontology (Sagri, 2003)(Peters et al, 2006). The $Norm \leftrightarrow Case$ CODeP is the generically related CODeP for legal descriptions

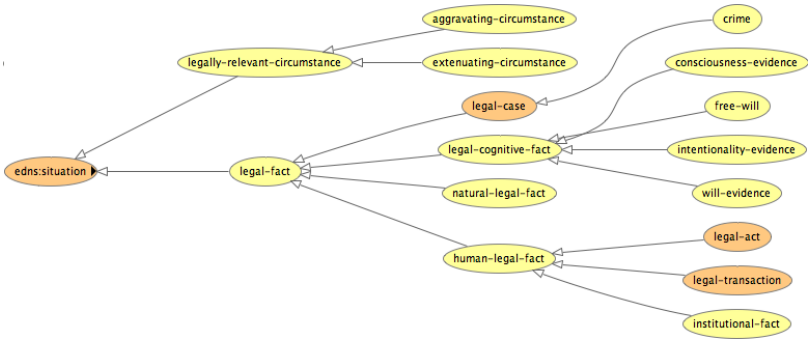



*Figure 6.* An excerpt of CLO taxonomy: legal fact and circumstance types as situations. The $Norm \leftrightarrow Case$ CODeP is the generically related CODeP for legal facts.

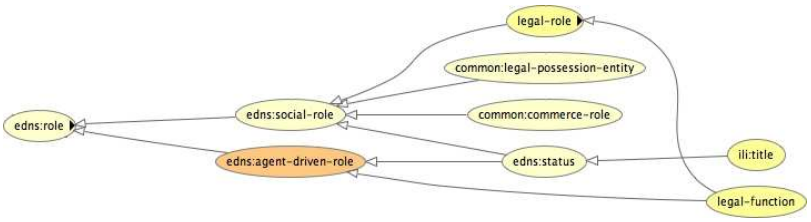



*Figure 7.* An excerpt of CLO taxonomy: legal roles. The $Norm \leftrightarrow Case$ CODeP is the generically related CODeP for legal roles.

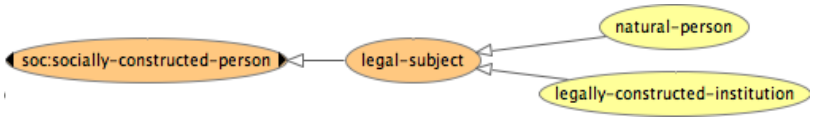



*Figure 8.* An excerpt of CLO taxonomy: legal subjects (or persons). In Fig. 9 a related CODeP for socially-constructed persons is provided.

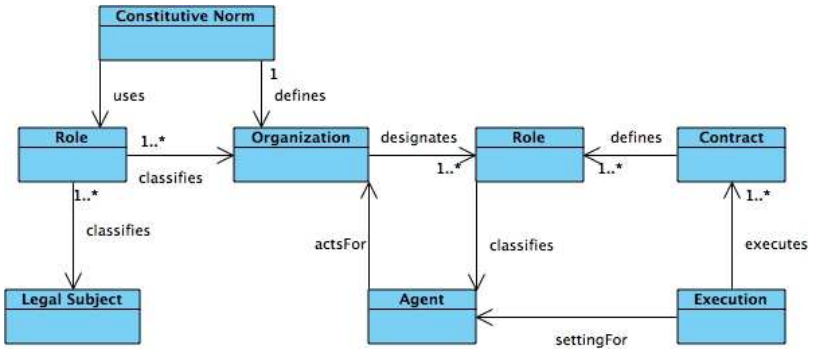



*Figure 9.* The *SociallyConstructedPerson* CODeP. Persons can be legal subjects (either natural persons or not), including organizations, or legal entities deputed to law enforcement. For example, an organization is defined by means of a constitutive norm that also uses concepts that either classify the organization or other legal subjects. An organization designates at least one role that can classify agents. For classified agents, we can say that they *actFor* the organization when they are in the *setting* of a contract execution that defines the organization's role.

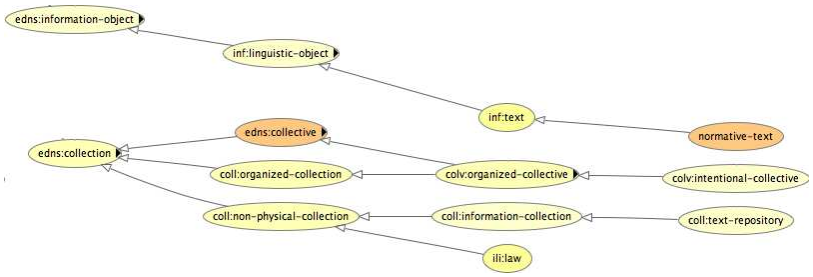



*Figure 10.* An excerpt of CLO taxonomy: legal informations and collections. In Fig. 11 a related CODeP is presented for information realizations and collections.

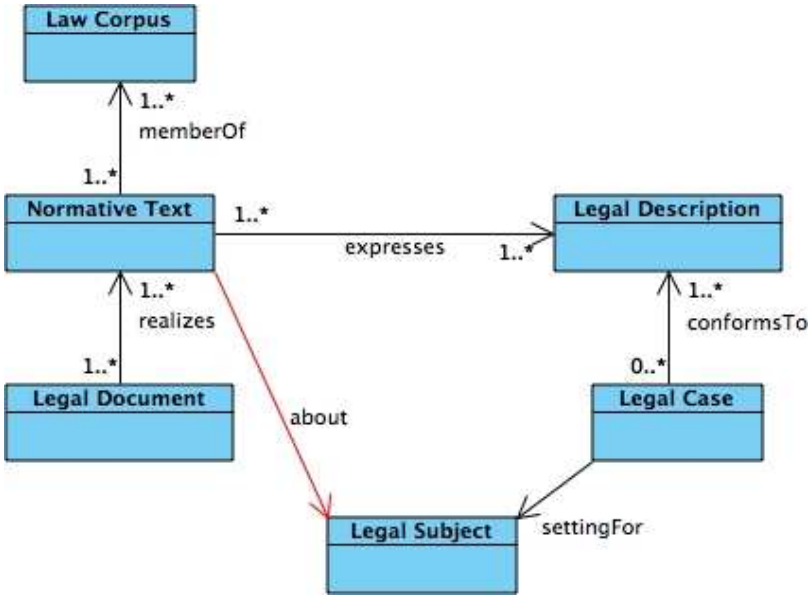



*Figure 11.* The *LegalInformationRealization&Collection* CODeP. A normative text is an information object, *member* of a corpus of laws, and *realized* by at least one legal document (its support). A text *expresses* a legal description (the public meaning of a law). For any legal case that *conforms* to the description, the text is *about* the legal subjects in the *setting* of the legal case.

# Developing ontologies for legal multimedia applications

Xavier Binefa[1], Ciro Gracia[1], Marius Monton[2], Jordi Carrabina[2], Carlos Montero[2], Javier Serrano[2], Mercedes Blázquez[3], Richard Benjamins[3], Emma Teodoro[3], Marta Poblet[4], Pompeu Casanovas[3]
[1] *Digital Video Understanding Group, UAB*
[2] *Laboratory for HW/SW Prototypes and Solutions (CEPHIS) UAB*
[3] *Institute of Law and Technology, Law Dpt., UAB*
[4] *ICREA Researcher at the Institute of Law and Technology, Law Dpt., UAB*

**Keywords:** Semantic Search, Ontology, HW/SW Acceleration Platforms, Reconfigurable Devices, Speaker Diarization, Video Segmentation.

## 1. Introduction

Search, retrieval, and management of multimedia contents are challenging tasks for users and researchers alike. The development of efficient systems to navigate through content has recently become an important research topic. Since domains as parliaments, courts, ministries, or security and military forces are producing enormous masses of video, audio and text files, the requirement of a specific content management solution have arisen naturally.

The aim of E-Sentencias is to develop a software-hardware system for the global management of the multimedia contents produced by Spanish civil courts. The Civil Procedure Act of January 7th, 2000 (1/2000) introduces the video recording of oral hearings. As a result, Spanish civil courts are currently producing a massive number of DVDs which have become part of the judicial file, together with suits, indictments, injunctions, judgments and pieces of evidence. This audiovisual material is used by lawyers, prosecutors and judges to prepare, if necessary, appeals to superior courts. Nevertheless, there is no available system at present to automatically annotate audiovisual contents within the judicial domain. E-Sentencias proposes a meta-search engine to manage text (legislation, jurisprudence, procedural documents, etc.), images, graph materials, and audiovisual contents in a dynamic way that combines algorithmic techniques with legal ontologies. Both automatic and semiautomatic processes facilitate the exploitation of the stored information by the users' website. In this regard, e-Sentencias involves technologies such as the Semantic Web, ontologies, NLP techniques, audio-video segmentation, and IR. The ultimate goal is to obtain an automatic classification of images and segments of the audiovisual records

that, coupled with textual semantics, allows the efficient navigation and retrieval of judicial documents and additional legal sources.

Section 2 below describes the current situation concerning the audiovisual recording of civil cases in Spain. In Section 3 we offer an overview of the steps followed towards the construction of a conceptual structure to classify video segments and the development of legal ontology applications. Sections 4 and 5 depict respectively the structure and architecture of the video system prototype at the present stage of research and, finally, we conclude by offering some expected results and conclusions in sections 5 and 6.

## 2. Video Recording of Civil Procedures in Spain

The provisions made by the 1/2000 Civil Procedure Act for the video recording of civil proceedings in Spain do not include a homogeneous protocol establishing how to obtain audiovisual records. Rather, and since an ever growing number of Autonomous Governments in Spain hold competencies on the organization of the judicial system there is a plurality of standards, formats, and methods to produce audiovisual records. As a result, analogical and digital standards coexist with different recording formats. The support in which copies are provided to legal professionals (i.e. to prepare an appeal) may also consist of either VHS videotapes or CDs. And, finally, the procedures to store, classify, and retrieve audiovisual records may vary even from court to court.

As regards the basic typology of civil proceedings, the 1/2000 Act sets two declarative processes: the ordinary proceeding and the verbal proceeding. The main differences between the two lie in the value of the case – more or less than Ä3000, respectively – and the legal object at dispute.

The steps of the process also vary depending on the specific proceeding. On the one hand, the ordinary proceeding starts with a separate, independent oral hearing called "*audiencia previa*" to resolve pre-judiciary issues (documents, evidences to be accepted, etc.), while verbal proceedings take place in the same judicial event. On the other hand, in the ordinary proceeding the claim of the plaintiff is contested in written terms, while in the verbal proceeding is replied orally in the same act.

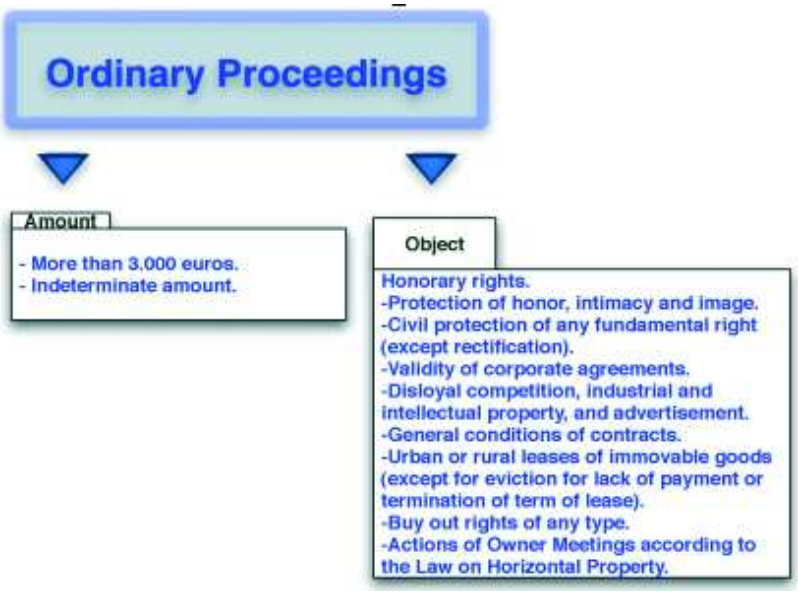



*Figure 1.* The ordinary proceeding: amount and content

## 3. Conceptual structure and ontology legal applications

One of the core objectives of e-Sentencias is to develop automatic classification strategies to classify video segments. To do so, we have started from scratch by transcribing a small set of oral hearings (corresponding to fifteen civil cases). Textual transcriptions also mark the different steps of the oral hearing and include a manual coding of legal concepts (i.e. judgment, injunction, cause of necessity, deed, etc.) and legal expressions (i.e. "with the permission of your Honor". In addition, they facilitate the coding of practical rules of procedure that are implicit in the video sequences, such as the following piece of transcription shows:

This is only a first level of textual and visual annotation of judicial hearings, but it is also the basis to create specific annotation templates at different levels (concepts, legal formulae, practical rules of interaction, etc.) that facilitate the construction of different types of ontologies.

In practice the use of ontologies for different tasks and purposes requires to consider the particular task as context for the ontology. The reason is that ontologies are often not really designed independent of the task at hand (Haase et al. 2006). In general, the context of use has an impact on the way concepts are interpreted to support certain functionalities. As some aspects of a domain are important in one con-

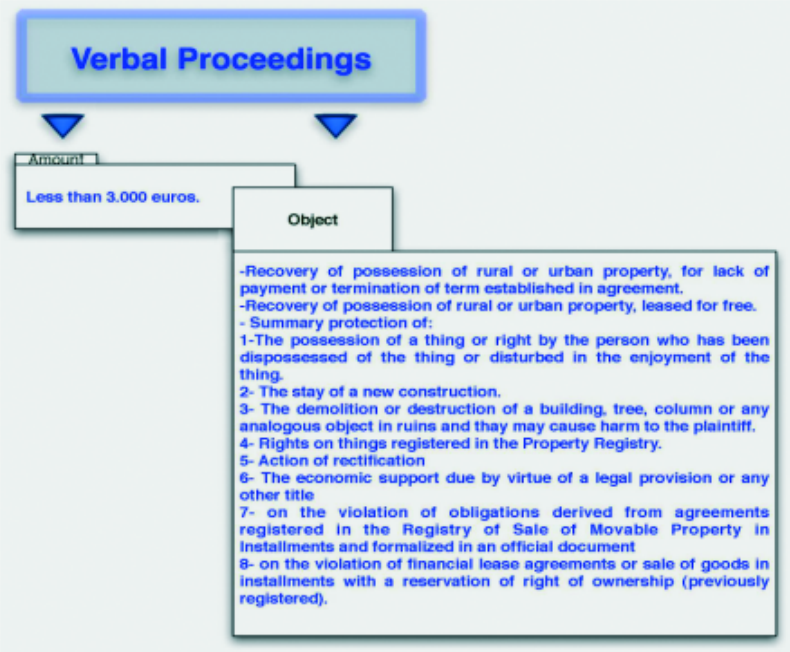



*Figure 2.* The verbal proceeding: amount and content

text but do not matter in another one, an uncontextualized ontology does not necessarily represent the features needed for a particular use. In order to solve this problem, we have to find ways to enable the representation of different viewpoints that better reflect the actual needs of the application at hand.
When talking about viewpoints, we can distinguish two basic use cases: In the first case, the aim is to provide means for maintaining and integrating different existing viewpoints. In the second use case, one may want to extract a certain viewpoint from an existing model that best fits the requirements of an application.

In many application domains (such as law) it is acknowledged that the creation of a single universal ontology is neither possible nor beneficial, because different tasks and viewpoints require different, often incompatible conceptual choices. As a result, we need to support situations where different parties commit to different viewpoints that cannot be integrated by imposing a global ontology. This situation demands for a weak notion of integration, in order to be able to exchange information between the viewpoints (Stuckenschmidt, 2006). Stuckenschmidt describes one of such examples from oncology. Oncology is a complex domain where several specialties, e.g. chemotherapy, surgery,

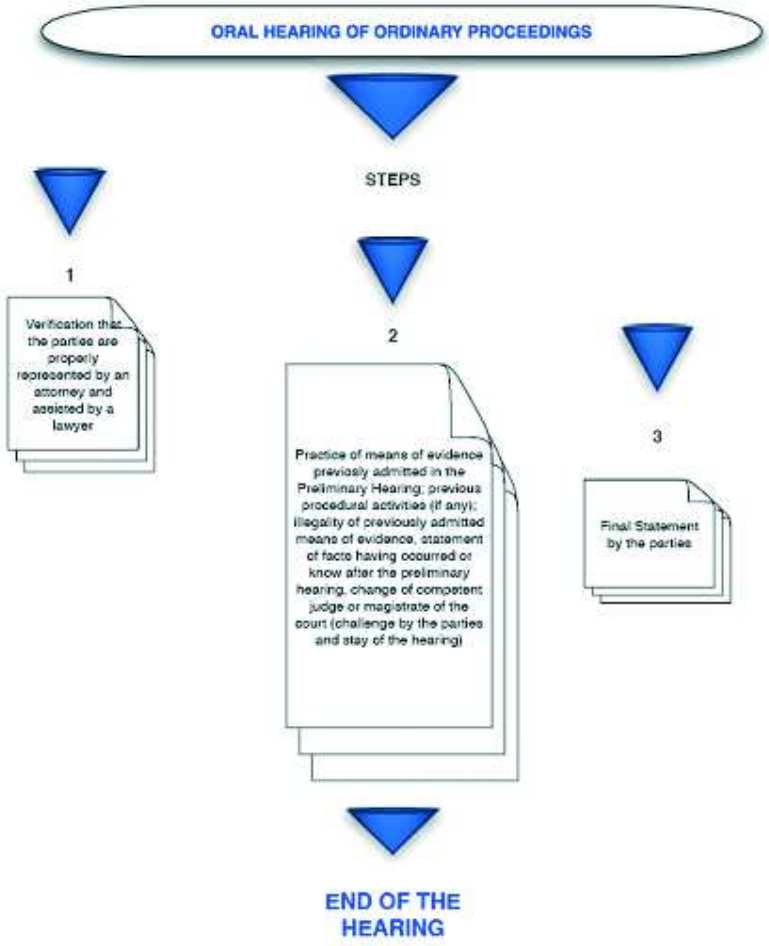



*Figure 3.* Steps of the process in ordinary proceedings.

and radiotherapy are involved in a sequence of treatment phases, each representing a particular viewpoint.

Law is also a complex domain, where several roles are involved (judge, prosecutor, defendant ...). They must be represented from different points of view, thinking of the possible use of the images of the hearings for multiple (and adversarial) purposes.

We find in the recent literature several approaches to this perspective problem and the so-called 'semantic gap': (i) multi-context ontologies vs. mono-context ontologies (Bensliman et al. 2006 ; Arara and Laurini, 2005 ; Dong and Li, 2006); (ii) low-level descriptors [pixel color, motion vectors, spatio-temporal relationships] vs. semantic descriptors [person,

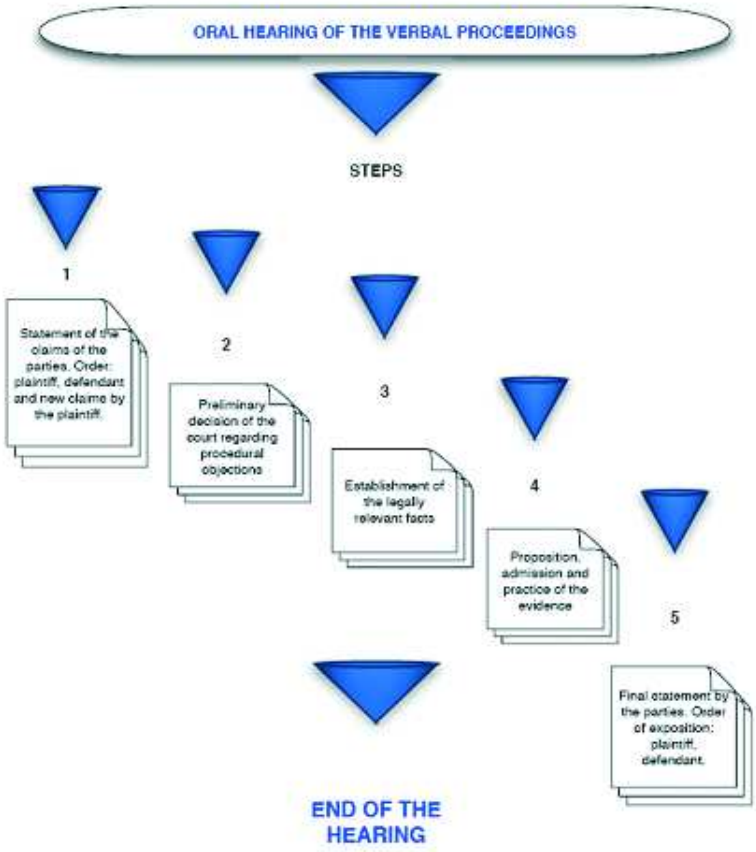



*Figure 4.* Steps of the process in verbal proceedings.

vehicle. . . ] (Petrides et al. 2005, Athanasiadis et al. 2005, Boehorn et al 2005) ; modal keywords of perceptual concepts [aural, visual, olfactory tactile, taste] vs. content topics (Jaimes et al. 2003a; Jaimes et al. 2003b); (iii) cross-media annotation (Deschachts and Moens 2007).
From a legal multimedia user-centered perspective there are two problems related to these proposals that have to be addressed (i) the definition of context in merging and aligning legal and multi-media ontologies; (ii) the specific exophoric nature of the legal videorecording.
Researchers on contextual ontologies use to define 'context' as *local* (not shared with other ontologies) and opposed to content *ontologies* themselves (shared models of a domain) (Bouquet et a. 2004; Haase et

<actor name="judge" tc="00.01.30">
Let us see mr. *** DEFENDANT STANDS UP AND APPROACHES TO THE MICROPHONE come to the microphone [PROCEDURAL FORMULA, EXCLUSIVE USE BY THE JUDGE]
</actor>
<actor name="defendant" tc="00.01.31">
yes
</actor>
<actor name="judge" tc="00.01.38">
and answer the questions that both attorneys are going to formulate, starting by the attorney of the plaintiff [GENERAL RULE: IF BOTH PARTIES HAVE REQUESTED EXAMINATION, THE PLAINTIFF'S ATTORNEY ALWAYS COMES FIRST IN EXAMINATING THE DEFENDANT, AND THEN CONTINUES THE DEFENDANT'S ATTORNEY].
</actor>
<actor name="plaintiff's attorney" tc="00.01.38">
With the permission of your honor [PROCEDURAL FORMULA, EXCLUSIVE USE BY THE ATTORNEYS] eh do you know whether mrs. **** is being living with her grandmother mrs ** since january 2001
</actor>

*Figure 4.*

al. 2006).[1] Therefore, to cope with the directionality of information flow, the local domains and the context mapping, which cannot be represented with the current syntax and semantics of OWL, C-OWL is being developed.[2]

From the multimedia researchers point of view, context is defined currently as 'the set of interrelated conditions in which visual entities (e.g. objects, scenes) exist' (Jaimes et al. 2003a,b). This grounds the strategy of the direct vs. indirect exploitation of the knowledge base to annotate the content of the videos, using *visual* and *content* descriptors alike (Bloedhorn et al. 2005).[3] But, most important, this definition of context entails a theoretical approach in which 'actions and events in

[1] 'It can be argued that the strengths of ontologies are the weakness of contexts and vice-versa' (Bouquet et al., Haase et al. ibid.).

[2] *Directionality of information flow*: keeping track of the source and the target ontology a specific piece of information; *local domains*: giving up the hypothesis that all legal ontologies are interpreted in a single global domain; *context mapping*: stating that two elements (concepts, roles, individuals) of two ontologies, though extensionally different, are contextually related, e.g., because they both refer to the same object in the word (Bouquet at al. 2004).

[3] 'The main idea of our approach lies in a way to associate concepts with instances that are deemed to be prototypical by their annotators with regard to their visual characteristics' (ibid. 2005: 593).

time and space convey stories, so, a video program (raw video data) must be viewed as a document, not a non-structured sequence of frames' (Song et al. 2005, 2006). In such an approach, visual low level features, object recognition and audio speaker diarization (process of partitioning the audio stream in homogenous segments and clustered according to speaker identity) are crucial to analyze e.g. a sport or movies' sequences.

However, the audiovisual documents that are recorded in Spanish courtrooms do not convey actions, but *legal narratives*. Motion and colour are generally uniform, since they are not considered the relevant aspect of those documents. Thus, court records are technically very poor (see fig. 5), filmed using a one-shot perspective (the camera is situated above and behind the judge, who never appears on the screen). Rather than *telling a story*, the video structures a single framework in which a story is referred, conveyed and constructed by the procedural actors (judge, counsels, testimonies, secretary, and court clerks).

Here lies the *layered exophoricity* of the legal discourse. Actions, events and stories are referred into a contextually embedded discourse, procedurally-driven, and hierarchically conducted by the judge (judge-centered). Therefore, a strong *décalage* is produced between audio and video as sources of information. A legal court video record would be completely useless without the audio, because we may only infer procedural (but not substantial) items from the motion. What is important is what is *said* in court, not what is *done*. Visual images are only ancillary related to the audio stream. This is an important feature of the records, which has to be taken into account in the tasks of extracting, merging and aligning ontologies, because what the different users require (judges, lawyers, citizens) is the combination of different functionalities focused on the legal information content (legislation quoted, previous cases and judgements –precedent-, personal professional records, and so on). This is the reason for a hybrid user-centred approach that is the kernel of our theoretical approach.

## 4. Structure of the Video Prototype

The development of an intuitive user interface constitutes a central requirement of the system. While preserving the simplicity of use, the application allows: a) access to the legally significant contents of the video file; b) integration of all procedural documents related to the oral hearing; c) management of sequential observations, and d) semantic queries on the contextual procedural aspects.

The structure of the application is based on two intuitive and semantically powerful metaphors: the *oral hearing line* and the *oral hearing*

*Figure 5.* Image quality.

*axe*. The *oral hearing line* presents a timeline divided into segments. Each segment represents a different speech, produced by one of participants in the process: judge, secretary, attorneys, witnesses, etc. Each participant is represented by a different color to obtain an identification at first glance of their interventions. Therefore, it is possible to visualize specific contents of the video by merely clicking on a particular colored sequence. Moreover, it is possible to add textual information to any instant of the intervention.
The *oral hearing axe* consists of a column representing the different phases of the event as defined by procedural legislation. Different phases (as opening statements, presentation of evidences, concluding statements, etc) are represented by different colors, allowing a quick access. It is also possible to access to legal documents related to each phase (i. e. pieces of evidence such as contracts, invoices, etc.) as well as to jurisprudence quoted in the oral hearing and detected through phonetic analysis. This legal information is also structured in directories and folders.

As Figure 6 shows, the user interface is divided into two main parts: the upper part contains the video player, the *oral hearing axe* and the *oral hearing line*. The lower part is devoted to external information layers (i.e. references to articles, documents annexed, manual annotations, links to jurisprudence, etc.). This part is divided into two tabs. The first

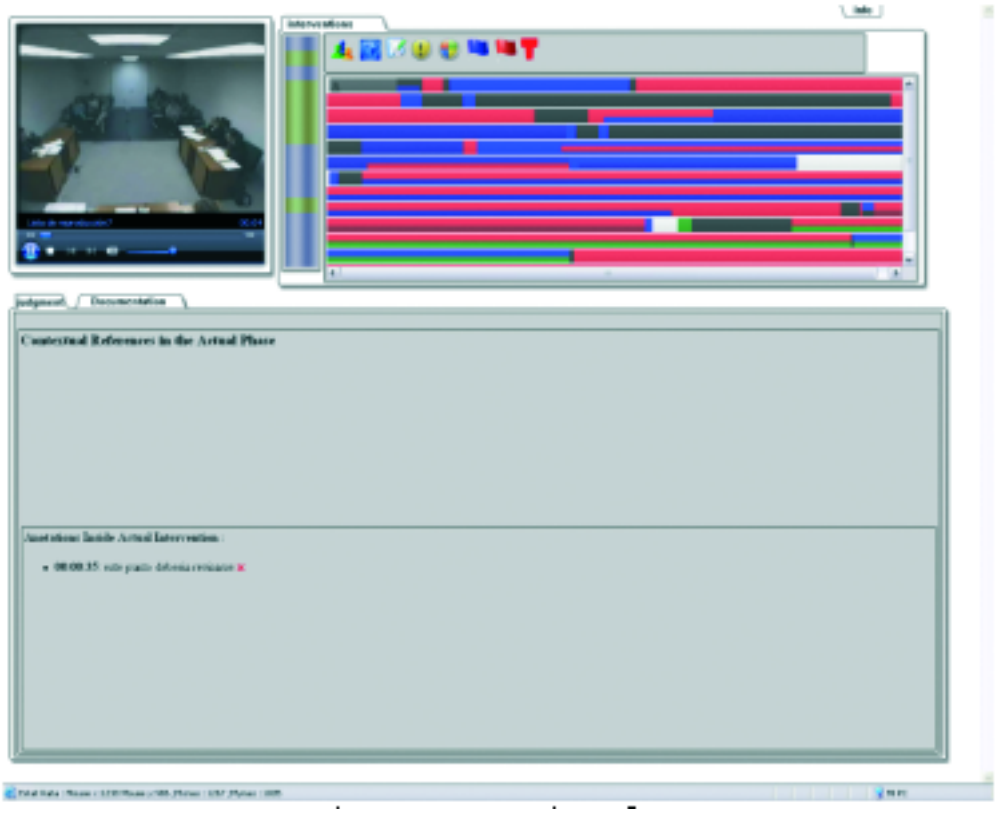


*Figure 6.* User interface.

one contains important information of the selected phase, allowing the addition of the different documents presented during the phase. The second tab contains historical information of the process and all the related information available in advance.

The main functionalities offered in the upper part of the user interface are:

1) The information tab: this is a scrollable tab containing the most relevant data of the process.

2) The *oral hearing line*: the timeline of sequences and interventions assigned to the different actors of the process. One single sequence of the video may contain interventions of different actors. Therefore, sequences may be either mono-colored (intervention of one single part) or multi-colored (more than one part intervening in the same sequence). The horizontal length of each segment of the

timeline is proportional to its length in seconds. The application includes two modes of playing video, apart of the usual one. It is possible to select either the visualization of all the interventions by a single participant or, in turn, all the interventions on a given phase.

3) The list of intervening parties: Each actor intervening in the process is represented by an icon. As in the case of the *oral hearing line*, we may choose to visualize only those sequences appearing one specific participant (i.e. the judge or de defense attorney).

4) The *oral hearing axe*: this is the vertical line representing the procedural phases of the process. The judicial process is therefore divided in procedural phases which can, as well, be subdivided in interventions. The vertical axe has the advantage of providing quick access to interventions belonging to a given phase.

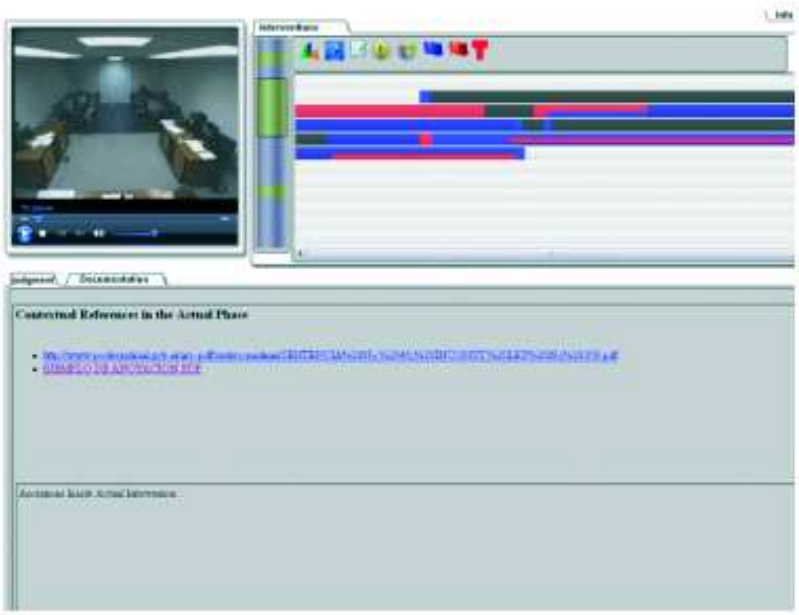



*Figure 7.* Interventions of one procedural phase and related information.

In addition to these functionalities, it is possible make a manual annotation of the sequence. Double-clicking with the right bottom of the mouse over a sequence running on the video screen opens a pop-up with a manual annotation tool.

As regards the lower part of the user interface, this area contains all the relevant information and documents of the process, but also enables the user to add and organize the information appearing during the different phases. This part is divided into two different sections:

1) An area enabling the visualization of all the references related to each phase of the process. References consist of data (i.e. Civil Code articles, judgments, Internet links, etc.) automatically introduced through semantic annotation.

2) An area including all manual annotations of the sequences made by the user.

## 5. Architecture of the video prototype

The architecture of the system is based on a web system including the following components:

1) Video server WMS: a server based on Windows 2003 Enterprise server with a streaming Windows Media services which allows video broadcast of audiovisual content of the judicial processes under demand. Application server TOMCAT: the application serves web contents and provides the required interaction with the database by means of Java Server Pages;

2) Mysql Database: the Mysql database contains the information related to all processes and their respective annotations;

3) Client browser IE 7.0: It allows the management of the user interface and the management of the user interaction with the embedded Windows Media Player 11 that streams the video.

## 6. Conclusions and expected results

In the E-Sentencias project we expect to obtain two different types of results. On the one hand, a fully annotated legal corpus of multimedia oral hearings classified in 15 procedural classes, as regulated by the 1/2000 Act. On the other hand, an operational system with a human-computer interface as described in this paper. Using the system prototype, the automatic capabilities of speaker interventions and phases detection will be tested against manually annotated corpus. It

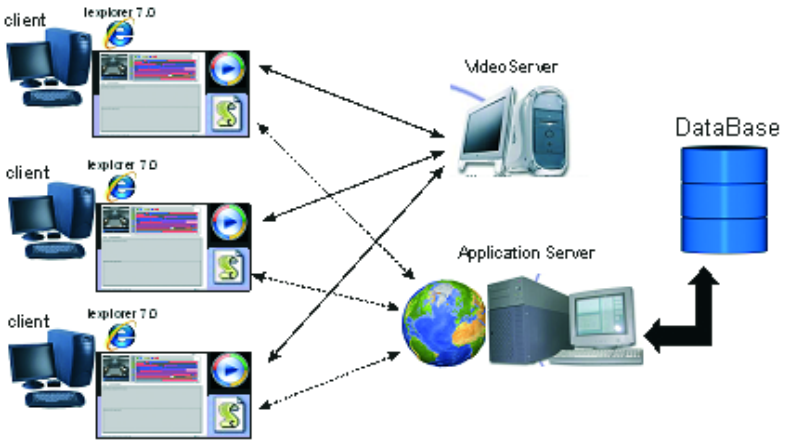



*Figure 8.* Architecture components and interactions between them.

will also be evaluated in cross-oral hearings retrieval based on hardware accelerated and specifically implemented multimedia ontologies.

## Acknowledgements

E-Sentencias (E-Sentencias. *Plataforma hardware-software de aceleración del proceso de generación y gestión de conocimiento e imágenes para la justicia*) is a Project funded by the Ministerio de Industria, Turismo y Comercio (FIT-350101-2006-26). A consortium of: Intelligent Software Components (iSOCO), Wolters Kluwer España, IUAB Institute of Law and Technology (IDT-UAB), Centro de Prototipos y Soluciones Hardware - Software (CHEPIS - UAB) y Digital Video Semantics (Dpt. Computer Science UAB).

## References

Arara, A.A., Laurini, R. Formal Contextual Ontologies for Intelligent Information Systems, 2005. *Enformatika* 5: 303-306.

Athanasiadis, T., Tzouvaras, V., Petridis, K., Precioso, F., Avrithis, Y., Kompatsiaris, Y. Using a Multimedia Ontology Infrastructure for Semantic Annotation of Multimedia Content, 2005. Proc. of 5th International Workshop on Knowledge Markup and Semantic Annotation ( '05), Galway, Ireland, November 2005

Benjamins, V.R., Casanovas, P., Gangemi, A. and Breuker, J. (ed.). 2005. *Law and the Semantic Web: Legal Ontologies, Methodologies, Legal Information Retrieval, and Applications.* Lecture Notes in Computer Science. Berlin, Springer Verlag.

Bensilimane, D., Arara, A., Falquet, G., Maamar, Z., Thiran, P., Gargouri, F. 2006. Contextual Ontologies. Motivations, Challenges and Solutions. Fourth Biennial International Conference on Advances in Information Systems ADVIS, October $18^{th}$-$20^{th}$, Ankara.

Bloedhorn, S., Petridis, K., Saathoff, C., Simou, N., Tzouvaras, V., Avrithis, Y., Handschuh, S., Kompatsiaris, Y., Staab, Y., Strintzis, M.G. Semantic Annotation of Images and Videos for Multimedia Analysis. A.Gómez Pérez and J. Euzenat (eds.) ESWC 2005, Lecture Notes in Computer Science 3532, 592-607.

Bouquet, P.; Giunchiglia, F., van Harmelen, F., Serafín, L., Stuckenschmidt, H. C-OWL: Contextualizing Ontologies. In D. Fensel et al. ISWC 2003, Lecture Notes in Computer Science 2870: 164-179.

Bouquet, P., Giunchiglia, F., van Harmelen, F., Serafín, L., Stuckenschmidt, H. 2004. Contextualizing ontologies. *Journal of Web Semantics* 26): 1-19.

Breuker, J., Elhag, A., Petkov, E. and Winkels, R. 2002. Ontologies for legal information serving and knowledge management. In *Legal Knowledge and Information Systems, Jurix 2002: The Fifteenth Annual Conference.* IOS Press.

Casanovas, P., Poblet, M., Casellas, N., Vallbé, J.-J., Ramos, F., Benjamins, V.R., Blázquez, M., Rodrigo, L., Contreras, J. and Gorroñogoitia, 2004. J. *D10.2.1 Legal Case Study: Legal Scenario.* Technical Report SEKT, EU-IST Project IST-2003-506826.

Casanovas, P., Casellas, N., Vallbé, J.-J., Poblet, M., Ramos, R., Gorroñogoitia, J., Contreras, J., Blázquez, M. and Benjamins, V.R. 2005. Iuriservice II: Ontology Development and Architectural Design. In *Proceedings of the Tenth International Conference on Artificial Intelligence and Law (ICAIL 2005)*. Alma Mater Studiorum-University of Bologna, CIRSFID.

Casanovas, P., Casellas, N., Vallbé, J.-J., Poblet, M., Benjamins, V.R. Blázquez, M., Peña-Ortiz, Rl and Contreras, J. 2006. Semantic Web: A Legal Case Study. In J. Davies, R. Studer and P. Warren, editors, *Semantic Web Technologies: Trends and Research in Ontology-based Systems.* John Wiley & Sons.

Casellas, N., Jakulin, A., Vallbé, J.-J. and Casanovas, P. 2006. Acquiring an ontology from the text. In M. Ali and R. Dapoigny, editors, *Advances in Applied Artificial Intelligence, $19^{th}$ Internatoinal Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems (IEA/AIE 2006).* Annecy, France, June 27-30 2006, Lecture Notes in Computer Science 4031, Springer, 1000-1013.

Deschacht, K. and Moens, MF. 2007. Text Analysis for Automatic Image Annotation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics, Prague, Czech Republic, June 23rd–30th, 2007.*

Dong, A., Li, H. 2006. Multi-ontology Based Multimedia Annotation for Domain-specific Information Retrieval. Proc. Of the IEEE International Conference on Sensor Networks, Ubiquitous and Trusworthy Computing (SUTC' 06).

Haase, Peter; Hitzler, Pascal; Rudolph, Sebastian; Oi, Guilin; Grobelnik, Marko; Mozeti, Igor; Damjan, Bojad Ziev; Euzenat, Jerome; d'Aquin, Mathieu; Gangemi, Aldo; Catenacci, Carola. 2006. *D3.11 Context Languages-State of the Art.* NeOn Project EU-IST Integrated project (IP) IST-2006.

Jaimes, A., Smith, J.R. 2003a. Semi-automatic, Data-driven Construction of Multimedia Ontologies. ICME 203, IEEE.

Jaimes, A., Tseng, B., Smith, J.R. 2003b. Modal Keywords, Ontologies, and Reasoning for Video Understanding. CIVR 2003, E.M.Bakker et al. (eds.) Lecture Notes on Computer Science 2728: 248-259.

Petridis, K., Precioso, F., Athanasiadis, T., Avrithis, Y., Kompatsiaris, Y. 2005. Combined Domain Specific and Multimedia Ontologies for Image Undertanding.

28th German Conference on Artificial Intelligence, Koblenz, Germany, September 11-14.

Song, D., Liu, H.T., Cho, M., Kim, H., Kim, P. 2005. Domain Knowledge Ontology Building for Semantic Video Event Description. W.K. Leow et al. (Eds.) CIVR-05, Lecture Notes in Computer Science 3568, 267-275.

Song, D., Cho, M., Choi, C., Shin, J., Park, J., Kim, P. A. Hoffmann et al. (Eds.) PKAW -06, Lecture Notes in Artificial Intelligence 4303, 144-155.

Stuckenschmidt, Heiner. 2006. Toward Multi-viewpoint Reasoning with Owl Ontologies. In ESWC, pages 259.272.

# A linguistic-ontological support for multilingual legislative drafting: the DALOS Project

Enrico Francesconi, Pierluigi Spinosa, Daniela Tiscornia
*Institute of Legal Information Theory and Techniques, Italian National Research Council (ITTIG-CNR), Italy*
*{francesconi,spinosa,tiscornia}@ittig.cnr.it*

**Abstract.** Coherence and alignment of the legislative language highly contribute to the quality of legislative processes, to the clarity of legislative texts and to their accessibility. DALOS aims at ensuring that legal drafters and decision-makers have control over the multilingual language of European legislation, and over the linguistic and conceptual issues involved in its transposition at national levels. The project will contribute to this goal by providing law-makers with linguistic and knowledge management tools to support the legislative drafting activity.

**Keywords:** Legislative drafting, multiligualism, domain ontology, lexical taxonomy

## 1. Introduction

Coherence, interoperability and harmonization in the legislative knowledge of, and control over, the legal lexicon is a precondition for improving the quality of legislative language and for facilitating access to legislation by legal experts and citizens. In a multilingual environment, and in particular, in EU regulations, only the awareness of the subtleties of legal lexicon, in the different languages, can enable drafters to maintain coherence among the different linguistic version of the same text. This is as much important for the EU Member State legal orders, strongly influenced by the obligation to implement EU directives.

To face this problem recently the DALOS[1] project has been launched within the "eParticipation" framework, the EU Commission initiative aimed at promoting the development and use of Information and Communication Technologies in the legislative decision-making processes, with the aim to foster the quality of the legislative production, to enhance accessibility and alignment of legislation at European level, as well as to promote awareness and democratic participation of citizens to the legislative process.

In particular DALOS aims at ensuring that legal drafters and decision-makers have control over the legal language at national and European level, by providing law-makers with linguistic and knowledge manage-

[1] DrAfting Legislation with Ontology-based Support

ment tools to be used in the legislative processes, in particular within the phase of legislative drafting.

Nowadays the key approach for dealing with lexical complexity is the ontological one, by which we mean a characterisation (understood both by people and processed by machines) of the conceptual meaning of the lexical units and of their connection with other terms. On the basis of an ontological characterisation of legal language DALOS wants to provide law-makers with linguistic and knowledge management tools to support legislative drafting in a multilingual environment.

In this paper an overview of the DALOS project is given. In particular in Section 2 the complexity of the multilingual legal scenario is addressed; in Section 3 the characteristic of the DALOS linguistic-ontological approach is discussed; in Section 4 the specification of the DALOS Knowledge Organization System (KOS) is presented; in Section 5 the methodologies to implement the DALOS ontological-linguistic resource are shown; finally in Section 6some conclusions are reported.

## 2. Interfacing multilingual legal terminologies

In legal language every term collection belonging to a language system, and any vocabulary originated by a law system, is an autonomous vocabulary resource and should be mapped through relationships of equivalence with the others. Based on the assumption that in a legal domain one cannot transfer the conceptual structure from one legal system to another, it is obvious that the best approach consists in developing parallel alignment with the same methodology and the same conceptual model. Different methods may be applied, depending on the characteristic of the domain, the data structure and on the result to achieve.

As regards the data structure, the first consideration is that unstructured list of terms (as for instance traditional flat terminologies) cannot be mapped in a consistent way, but only connected by a one-to-one correspondence among terms, which is an invalid approach for a context dependent technical terminology, such as law vocabulary. Among structured data different degrees of formalization can be distinguished:

– controlled vocabularies (such as thesauri, classification trees, directories, key-words lists): terms are organized in taxonomic trees, linked by generic associative relations, and concepts are implicitly expressed by lists of preferred and variant terms (descriptors/non-descriptors);

- semantic lexicons, also called computational lexicons or lightweight ontologies are based on commonly accepted semantic definitions and on a limited formal modeling;
- foundational, core, and domain ontologies are formal models (logical theories) of a conceptualization of a given domain, often based on axiomatic definitions.

The integration of lexical resources (heterogeneous because belonging to different law systems, or expressed in different languages, or pertaining to different domains) leads to different final results depending on the desired results:

- generate a single resources covering both (merging);
- compare and define correspondences and differences (mapping);
- combine different levels of knowledge representation, basically interfacing lexical resources and ontologies.

Of the three strategies, the methodological approach for DALOS requires the definition of mapping procedures among semantic lexicons, driven by the reference to an ontological level where the basic entities which populate the legal domain are described. In the next section the semantic structure of the lexical component is outlined.

### 2.1. A legal semantic lexicon: the LOIS database

Semantic lexicons are a means for content management which can provide a rich semantic repository. Compared to formal ontologies, semantic lexicons are lightweight ontologies as they are based on a weak abstraction model, with limited formal modeling, since constraints over relations are based on the grammatical distinctions of language (noun, verbs, adjectives, adverbs), for instance the agent-role relation holds between a noun (agent) and a verb or event denoting nouns (action) ((Castagnoli et al., 2006)) In the legal field, one of the wider semantic lexicons currently available is the LOIS database[2] composed by about 35.000 concepts in five European languages (English, German, Portuguese, Czech, and Italian, linked by English).

In LOIS a concept is expressed by a synset, the atomic unit of the semantic net. A synset is a set of one or more uninflected word forms (lemmas) with the same part-of-speech (noun, verb, adjective,

[2] created within the European project LOIS (Legal Ontologies for Knowledge Sharing, EDC 22161, 2003-2006)

and adverb) that can be interchanged in a certain context. For example *action, trial, proceedings, law suit* form a noun synset because they can be used to refer to the same concept. More precisely each synset is a set of wordsenses, since polysemous terms are distinct in different wordsenses. A synset is often further described by a gloss, explaining the meaning of the concept. English glosses drive cross-lingual linking.

In monolingual lexicons terms are linked by lexical relations: synonymy (included in the notion of synset), near-synonym, antonym, derivation. Synsets are linked by semantic relations of which the most important are hypernymy/hyponymy (between specific and more general concepts), meronymy (between parts or wholes), thematic roles, instance-of.

Cross-lingual linking is based on equivalence relations of each synsets with an English synset: these relations indicate complete equivalence, near equivalence, or equivalence as a hyponym or hyperonym. The network of equivalence relations, the Inter-Lingual-Index (ILI), determines the interconnectivity of the indigenous wordnets. Language-specific synsets from different languages linked to the same ILI-record by means of a synonym relation are considered conceptually equivalent. The LOIS approach are not completely language-independent, since the equivalence setting passes throughout the English wordnet and the English translation of glossas support the localization process.

The lesson learned from the LOIS experience is that a limited language independence could be enough for cross-lingual retrieval tasks, but that it could be a weak point when considering re-using, extending, updating the semantic connections or when integrating external lexical resources (for instance multilingual thesauri) within the framework. What is needed is "the distinction between conceptual modeling at a language-independent level and a language and culture specific analysis and description of discourse-related units of understanding" (Kerremans and Temmerman, 2004).

These consideratons led us to make clear distinction, when designing the overall model of DALOS and the system architecture, among:

- types of knowledge
- layers of knowledge representation
- classes of semantic relationships between knowledge elements.

## 3. Which knowledge for the DALOS service?

DALOS aims at providing a knowledge resource on the basis of the LOIS experience.

The two projects however address two different scenarios: while the LOIS knowledge resource is addressed to multiligual legal information retrieval, the DALOS knowledge resource is expected to support legislative drafting.

This distinction of the addressed scenario is particularly important because it contributes to identify the type of knowledge to be described within the DALOS service, so to avoid the so called *epistemological promiscuity* addressed by Breuker and Hoekstra (Breuker and Hoekstra, 2004), namely the common attitude to "indiscriminately mixing epistemological knowledge and domain knowledge in ontologies" which prevents knowledge representations from being automatically reusable outside the specific context for which the knowledge representation was originally developed.

As underlined by (Boer et al., 2004) the "*norm* is an epistemological concept identified by its role in a type of reasoning and not something that exclusively belongs to the vocabulary of the legal domain". As argued, "knowledge about reasoning – *epistemology* – and knowledge about the problem domain – *domain ontology* – are to be separated if the knowledge representation is to be reusable" (Boer et al., 2004).

The DALOS case addresses the legislative drafting process, namely a process that creates norms on specific domains to be regulated. What is needed therefore is a knowledge and linguistic support giving a description of concepts, as well as their lexical manifestations in different languages, in specific domains *before* they are regulated.

In particular, for the DALOS knowledge resource, avoiding *epistemological promiscuity* means to avoid that the knowledge to be used as support for legislative drafting (*domain knowledge*) is mixed with the knowledge on the general process of drafting (*epistemological knowledge*) which, obviously, pertains to different domains (see also (Biagioli and Francesconi, 2005)).

According to previous works (Biagioli, 1997) the epistemological knowledge related to the legislative drafting process can be modelled by the *Model of Provisions* which establishes a taxonomy of provision types (rules as *definition*, *obligation*, *prohibition*, *sanction*) and amendments (*insertion*, *repeal*, *substitution*) which describe legislative texts irrespective to the domain addressed, and pertain to the process of legislative drafting. Such kind of knowledge therefore will not be described by the DALOS resource, which, on the contrary, will contain knowledge on a

domain of interest. In particular for the aim of developing a project pilot, the "consumer protection" domain has been chosen.

## 4. KOS of the linguistic-ontological resource

In this phase of the project the most part of the activities are addressed to provide the specification for the DALOS resource. Chosen the domain of interest ("consumer protection") currently the activities for domain knowledge specification are oriented to:

- the standards to be used for knowledge representation;
- the Knowledge Organization System (KOS).

As regards the standards, the RDF/OWL standard conversion of WordNet approved by the W3C standards will be used for the linguistic resource (), thus guaranteeing interoperability as well as scalability of the solution.

As regards KOS, on the basis of the arguments expressed in Section 2.1, the DALOS resource is expected to be organized in two layers of abstraction (see Fig. 1):

- the *ontological layer* containing the conceptual modeling at a language-independent level;
- the *lexical layer* containing the lexical manifestations in different languages of the concepts at the ontological layer.

Basically the ontological layer acts as a knowledge layer where to align concepts at European level independently from the language and the legal order, according to the EU Commission recommendations for Member State legislations. Moreover the ontological layer allows to reduce the computational complexity of the problem of multilingual term mapping (N-to-N mapping). Concepts at the ontological layer act a "pivot" meta-language in a N-language environment, allowing the reduction of the number of bilingual mapping relationships from a factor $N^2$ to a factor $2N$. Concepts at the ontological layer are linked by taxonomical (`is_a`) as well as object property relationships.

On the contrary the lexical layer aims at describing language-dependent lexical manifestations of the concepts of the ontological layer. At this level terms will be linked by linguistic relationships as those ones used for the LOIS database (`hyperonymy`, `hyponymy`, `meronymy`, etc.). In particular, to implement the lexical layer, the subset of the LOIS

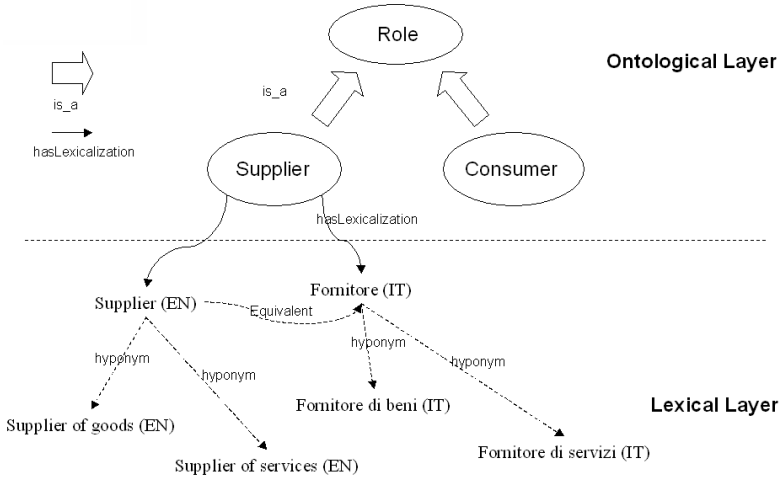



*Figure 1.* Knowledge Organization System (KOS) of the DALOS resource.

database pertaining to the "consumer protection" lexicon will be used. Moreover this database will be upgraded by using further texts where to extract pertaining terms from.

The connection between these two layers is aimed at representing the relationship between concepts and their lexical manifestations:

- within a single-language context (different lexical variations (lemmas) of the same meaning (concept));
- in a cross-language context (multilingual variations of the same concept).

In the DALOS KOS such link is represented by the `hasLexicalization` (and its inverse `hasConceptualization`) relationship.

## 5. Implementation of the DALOS resource

In order to implement the DALOS linguistic-ontological resource three main activites are foreseen:

1. Extracting terms of the domain of "consumer protection" law from a set of chosen texts by using NLP tools; this activity is aimed at upgrading the LOIS database (Lexical layer);
2. Construction of a Domain Ontology on the "consumer protection" domain (Ontological layer);

3. Semi-automatic connection between the LOIS database selection and the Domain Ontology by the `hasLexicalization` property implementation and its inverse `hasConceptualization` [Lexical layer ↔ Ontological layer]). This activity will be supported by automatic tools and validated by humans.

The first activity (implementation of the Lexical Layer) will be carried out using different NLP tools specifically addressed to process Italian texts (T2K) as well as English and other EU language texts (GATE).
T2K[3] is a terminology extractor and ontology learning tool jointly developed by CNR-ILC[4] and University of Pisa which combines linguistic and statistical techniques. It performs the following tasks: a) acquisition of domain terminology, both simple and multi-word terms, from a document collection; b) organisation and structuring of the set of acquired terms into taxonomical chains and clusters of semantically related terms. It works on Italian document collections; in principle it could be applied to document collections in languages other than Italian provided than NLP resources and tools for those languages exist (i.e. taggers, chunkers, dependency parsers).
GATE[5] is a tool to support advanced language analysis, data visualisation, and information sharing in many languages, owned/provided and maintained by the Department of Computer Science of the University of Sheffield.

The second activity (construction of a Domain Ontology) will be an intellectual one which aims at describing the scenario to be regulated. In this context the use of an ontology is of primary importance. Laws in fact usually contain provisions (Biagioli, 1997) which deal with entities (arguments) but they do not provide any general information on them: for example the Italian privacy law regulates the behaviour of the entity "Data controller" who is the owner of a set of personal data, but such law does not give any additional information on this role in the real domain-life (Biagioli and Francesconi, 2005). Therefore a formalized description in terms of an ontology of the domain to be regulated will allow the possibility to obtain such additional general information on the entities a new act will deal with. Moreover, the use of an ontology, and particularly of the associated lexicon, allows to obtain a normalized form of the terms with which entities are expressed, enhanching the quality and the accessibility of legislative texts.

[3] Text-to-Knowledge
[4] Institute of Computational Linguistic of the Italian National Research Council
[5] General Architecture for Text Engineering

The third activity will deal with the connection between the two level of abstractions (the *ontological layer* and the *lexical layer*). This activity is expected to be particularly time consuming, since it will implement the legal concept alignment on the basis of their lexical manifestations in a multilingual environment. A tool to support such semi-automatic mapping is expected to be implemented within the project.

## 6. Conclusions

In this paper an overview of the DALOS project has been presented. The main purpose of the project is to provide law-makers with linguistic and knowledge management tools to be used in the legislative processes, in particular within the phase of legislative drafting. The aim is to keep control over the legal language, especially in a multilingual environment, as the EU legislation one, enhancing the quality of the legislative production, the accessibility and alignment of legislation at European level, as well as to promote awareness and democratic participation of citizens. The ontological approach designed for the project has been presented.

## References

Breuker J. and R. Hoekstra, *Epistemology and ontology in core ontologies: FOLaw and LRICore, two core ontologies for law.* In Proceedings of EKAW Workshop on Core ontologies. CEUR, 2004.

Boer A., T. van Engers, and R. Winkels, *Using Ontologies for Comparing and Harmonizing Legislation*, In Proceedings of the International Conference on Artificial Intelligence and Law, Edinburgh (UK), 2003. ACM Press.

Boer A., T. van Engers, and R. Winkels, *Mixing Legal and Non-legal Norms*, In Moens, M.-F. and Spyns, P., editors, *Jurix 2005: The Eighteenth Annual Conference.*, Legal Knowledge and Information Systems, pages 25–36, Amsterdam. IOS Press.

Biagioli C. and E. Francesconi, *A Visual Framework for Planning a New Bill*, In Quaderni CNIPA (Proceedings of the $3^{rd}$ Workshop on Legislative XML), n. 18, p.83-95, 2005.

Biagioli C., *Towards a legal rules functional micro-ontology*, Proceedings of workshop LEGONT '97.

Castagnoli S., W. Peters , M. T. Sagri, D. Tiscornia, *The LOIS Project*, in Proceedings of the LREC 2006 Conference, Genova, May 2006.

Kerremans K. and Temmerman R., *Towards Multilingual, Termontological Support in Ontology Engineering*, in Proceeding of Termino 2004, Workshop on Terminology, (2004).

# NLP-based ontology learning from legal texts. A case study.

Alessandro Lenci[1], Simonetta Montemagni[2], Vito Pirrelli[2], Giulia Venturi[2]
[1] *Dipartimento di Linguistica Ũ Università di Pisa, Italy*
[2] *Istituto di Linguistica Computazionale - CNR, Italy*

**Abstract.** The paper reports on the methodology and preliminary results of a case study in automatically extracting ontological knowledge from Italian legislative texts in the environmental domain. We use a fully–implemented ontology learning system (T2K) that includes a battery of tools for Natural Language Processing (NLP), statistical text analysis and machine language learning. Tools are dynamically integrated to provide an incremental representation of the content of vast repositories of unstructured documents. Evaluated results, however preliminary, are very encouraging, showing the great potential of NLP–powered incremental systems like T2K for accurate large–scale semi-automatic extraction of legal ontologies.

**Keywords:** ontology learning, document management, knowledge extraction from texts, Natural Language Processing

## 1. Introduction

Ontology building is nowadays a very active research field, as witnessed by the fast growing literature on the topic and the increasing number of Knowledge Management applications based on automated routines for ontology navigation and update. The enterprise, however, requires harvesting domain–specific knowledge on an unprecedented scale, by tapping and harmonizing knowledge sources of highly heterogeneous conception, format and coverage, ranging from foundational ontologies and structured databases to electronic text documents. As electronic texts still represent the most accessible and natural repositories of specialised information worldwide, we seem to have reached a stage where an unlimited demand for ontologically–interpreted knowledge disproportionally exceeds the availability of automatically–interpreted textual information.

To bridge such a critical gap, different methodologies have been proposed to automatically extract information from texts and provide a structured organisation of extracted knowledge in as diverse domains/sectors as bio–informatics, health–care, public administration and company document bases. The situation in the legal domain is in line with this general trend and probably made even more critical by the fact that laws are invariably conveyed through natural language.

The last few years have seen a growing body of research and practice in constructing legal ontologies and applying them to the law domain. A number of legal ontologies have been proposed in different research projects: yet, most of them focus on a upper level of concepts and were mostly hand-crafted by domain experts (for a survey of legal ontologies, see Valente 2005). It goes without saying that realistically large knowledge-based applications in the legal domain need more comprehensive ontologies incorporating up-to-date knowledge: ontology-learning from texts could be of some help in this direction.

To our knowledge, however, relatively few attempts have been made so far to automatically induce legal domain ontologies from texts: this is the case, for instance, of Lame (2005), Saias and Quaresma (2005) and Walter and Pinkal (2006). The work illustrated in this paper represents another attempt in this direction. It reports the results of a case study carried out in the legal domain to automatically induce ontological knowledge from texts with an ontology learning system, hereafter referred to as T2K (*Text–to–Knowledge*), jointly designed and developed by the Institute of Computational Linguistics (CNR) and the Department of Linguistics of the University of Pisa. The system offers a battery of tools for Natural Language Processing (NLP), statistical text analysis and machine language learning, which are dynamically integrated to provide an accurate representation of the content of vast repositories of unstructured documents in technical domains (Dell'Orletta *et al.*, 2006). Text interpretation ranges from acquisition of lexical and terminological resources, to advanced syntax and ontological/conceptual mapping. Interpretation results are annotated as XML metadata, thus offering the further bonus of a growing interoperability with automated content management systems for personalised knowledge profiling. Prototype versions of T2K are currently running on public administration portals and have been used for indexing E–learning and E–commerce materials. In what follows, we report some ontology learning experiments carried out with T2K on Italian legislative texts.

## 2. From text to knowledge: the role of NLP tools

Technologies in the area of knowledge management and information access are confronted with a typical acquisition paradox. As knowledge is mostly conveyed through text, content access requires *understanding the linguistic structures* representing content in text at a level of considerable detail. In turn, processing linguistic structures at the depth needed for content understanding presupposes that a considerable amount of *domain knowledge is already in place*. Structural ambigui-

ties, long–range dependency chains, complex domain–specific terms and the ubiquitous surface variability of phraseological expressions require the operation of a battery of disambiguating constraints, i.e. a set of interface rules mapping the underlying conceptual organization of a domain onto surface language. With no such constraints in place, text becomes a slippery ground of unstructured, strongly perspectivised and combinatorially ambiguous information bits.

In our view, there is no simple way out of this paradox. Pattern matching techniques allow for fragments of knowledge to be tracked down only in very limited text windows, while foundational ontologies are too general to be able to make successful contact with language variability at large. The only effective solution, we believe, is to understand and face the paradox in its full complexity. An incremental interleaving of robust parsing technology and machine learning techniques can go a long way towards meeting this objective. Language technology offers the jumping-off point for segmenting texts into grammatically meaningful syntagmatic units and organizing them into non-recursive phrasal "chunks" that do not seem to require domain-specific knwoledge. In turn, chunked texts can sensibly be accessed and compared for statistically–significant patterns of domain–specific terms to be tracked down. Surely, this level of paradigmatic categorization is still very rudimentary: at this stage we do not yet know how chunked units are mutually related in context (i.e. what grammatical relations link them in texts) or how similar they are semantically. To go beyond this stage, we suggest getting back to the syntagmatic organization of texts. Current parsing technologies allow for local dependency relations among chunks to be identified reliably. If a sufficiently large amount of parsed text is provided, local dependencies can be used to acquire a first level of domain–specific conceptual organization. We can then use this preliminary conceptual map for harder and longer dependency chains to be parsed and for larger and deeper conceptual networks to be acquired. To sum up, facing the bootstrapping paradox requires an incremental process of annotation–acquisition–annotation, whereby domain–specific knowledge is acquired from linguistically–annotated texts and then projected back onto texts for extra linguistic information to be annotated and further knowledge layers to be extracted.

To implement this scenario, a few NLP ingredients are required. Preliminary term extraction presupposes pos–tagged texts, where each word form is assigned the contextually appropriate part–of–speech and a set of morpho–syntactic features plus an indication of lemma. Whenever more information about the local syntactic context is to be exploited, it is advisable that basic syntactic structures are identified. As we shall see in more detail below, we use chunking technology to attain this

level of basic syntactic structuring. NLP requirements become more demanding when identified terms need be organised into larger conceptual structures and connected through long–distance relational information. For this purpose syntactic information must include identification of dependencies among lexical heads.

The approach to ontology learning adopted by T2K differentially exploits all these levels of linguistic annotation of texts in an incremental fashion. Term extraction operates on texts annotated with basic syntactic structures (so-called "chunks", see below). Identification of conceptual structures, on the other hand, is carried out against a dependency–annotated text. In what follows, the general architecture of the Italian parsing system underlying T2K (henceforth referred to as AnIta, Bartolini *et al.*, 2004) is briefly illustrated.

### 2.1. An outline of AnIta

The AnIta system consists of a suite of linguistic tools in charge of:

1. tokenisation of the input text;
2. morphological analysis (including lemmatisation) of the text;
3. parsing, articulated in two different steps:
   a) "chunking", carried out simultaneously with morpho-syntactic disambiguation;
   b) dependency analysis.

In what follows we will focus on the syntactic parsing components in charge of the linguistic pre-processing of texts for the different ontology learning tasks of T2K.

Text chunking is carried out through a battery of finite state automata (CHUG–IT, Federici *et al.*, 1996), which takes as input a morphologically analysed and lemmatised text and segments it into an unstructured (non-recursive) sequence of syntactically organized text units called "chunks" (e.g. nominal, verbal, prepositional chunks). Chunking requires a minimum of linguistic knowledge; its lexicon contains no other information than the entry's lemma, part of speech and morpho–syntactic features. A chunk is a textual unit of adjacent word tokens sharing the property of being related through dependency relations (es. pre–modifier, auxiliary, determiner, etc.). A chunked sentence, however, does not give information about the nature and scope of inter–chunk dependencies which are identified during the phase of dependency analysis (see below). Morpho–syntactic disambiguation is performed simultaneously to the chunking process.

*Il presente decreto stabilisce le norme per la prevenzione ed il contenimento dell'inquinamento da rumore [...]*
'this decree establishes the rules for prevention and control of noise pollution [...]'

```
[[CC:N_C][DET:IL#RD@MS][PREMOD:PRESENTE#A@MS][POTGOV:DECRETO#S@MS]]
[[CC:FV_C][POTGOV:STABILIRE#V@S3IP]]
[[CC:N_C][DET:LO#RD@FP][POTGOV:NORMA#S@FP]]
[[CC:P_C][PREP:PER#E][DET:LO#RD@FS][POTGOV:PREVENZIONE#S@FS]]
[[CC:COORD_C][CONJTYPE:E#CC]]
[[CC:N_C][DET:IL#RD@MS][POTGOV:CONTENIMENTO#S@MS]]
[[CC:di_C][DET:LO#RD@MS][POTGOV:INQUINAMENTO#S@MS]]
[[CC:P_C][PREP:DA#E][POTGOV:RUMORE#S@MS]]
```

*Figure 1.* A sample of chunked text

To be more concrete, the sentence fragment reported in Figure 1 is segmented into eight chunks, each including a sequence of adjacent word tokens mutually related through dependency links of some kind. For example, the first nominal chunk (N_C) covers three word tokens, *il presente decreto*: the noun head *decreto*, the adjectival premodifier *presente* and an introducing definite article. Although the representation is silent about the relationship between *stabilire* 'establish' and *le norme* 'the rules', this is not to entail that such a relationship cannot possibly hold: simply, the lexical knowledge available to this parsing component makes it impossible to state unambiguously how chunks relate to each other and the nature of this relationship. This is the task for further analysis steps.

Dependency parsing is aimed at identifying the full range of syntactic relations (e.g. subject, object, modifier, complement, etc.) within each sentence: syntactic relations are represented as dependency pairs between lexical heads. It is carried out by IDEAL (Bartolini *et al.*, 2002), a finite state compiler for dependency grammars. The IDEAL general grammar of Italian is formed by ca. 100 rules covering the major syntactic phenomena. The grammar rules are regular expressions (implemented as finite state automata) defined over chunk sequences, augmented with tests on chunk and lexical attributes. A "confidence value" (PLAUS) is associated with identified dependency relations, to determine a plausibility ranking among competing analyses. Figure 2 reports the dependency representation of the same sentence.

The output consists of binary relations between content words, typically a head and a dependent. There may be features associated with both participants in the relation conveying other types of information such as the semantic type of a dependent (ROLE) or the preposition

```
MODIF(DECRETO[34544.1],PRESENTE[34544.1]<role=RESTR>)plaus=100
SUBJ(STABILIRE[34544.2],DECRETO[34544.1])plaus=50
OBJD(STABILIRE[34544.2],NORMA[34544.3])plaus=50
COMP(NORMA[34544.3],PREVENZIONE[34544.4]<intro=PER>)plaus=50
COORD(PREVENZIONE[34544.4],CONTENIMENTO[34544.6]<role=CONJ>)plaus=50
ARG(CONTENIMENTO[34544.6],INQUINAMENTO[34544.7]<intro=DI>)plaus=60
COMP(INQUINAMENTO[34544.7],RUMORE[34544.8]<intro=DA>)plaus=50
```

*Figure 2.* A sample of dependency–parsed text

introducing a certain relation (INTRO). The sentence fragment is described by 7 dependency relations including subject, object as well as other modification relations: for instance, *decreto* has been identified as the subject of the verb *stabilire* and *norme* as its direct object.

There are some reasons to believe that chunked texts are a suitable starting point for term extraction from a continuously expanding document base. First, thanks to its knowledge–poor lexicon, chunking is fairly domain–independent. Moreover, its finite–state technology makes chunking very robust and flexible in the face of parse failures: unparsed sequences are tagged as unknown chunks and parsing can resume from the first ensuing word–form which is part of a parsable chunk. Thirdly, chunking provides a first level of syntactic grouping which, however crude, paves the way to reliable and wide–coverage identification of candidate domain terminology, including both single and multi–word terms. As chunks standardise a considerable amount of grammatical information, searching for candidate terms in a chunked text can be done at a considerable level of abstraction from language nitty–gritty. On the other hand, identification of clusters of semantically related terms or acquisition of relations between terms constitute more demanding tasks requiring deeper levels of linguistic analysis such as dependency parsing.

## 3. T2K architecture

T2K is a hybrid ontology learning system combining linguistic technologies and statistical techniques. T2K does its job into two basic steps:

1. extraction of domain terminology, both single and multi–word terms, from a document base;

2. organization and structuring of the set of acquired terms into proto-conceptual structures, namely a) fragments of taxonomical chains, and b) clusters of semantically related terms.

Figure 3 illustrates the functional architecture of T2K:

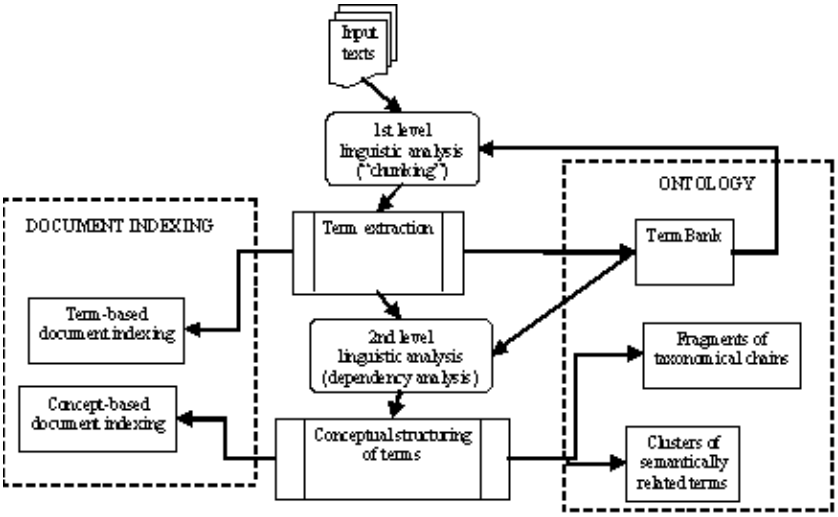



*Figure 3.* T2K architecture

The two basic steps take the central pillar of the portrayed architecture, showing the interleaving of NLP and statistical tools. Acquired results are structured in the ontology box on the right–hand–side of the diagram, whose stratified organization is reminiscent of the hierarchical cascade of knowledge layers in the "Ontology Learning Layer Cake" by (Buitelaar *et al.*, 2005), going from terminological information to proto–conceptual structures corresponding to taxonomical and non-hierarchical relationships among terms. Acquired knowledge is also used for document indexing, on the basis of extracted terms and acquired conceptual structures. In what follows we focus on the ontology learning process.

### 3.1. Term extraction

Term extraction is the first and most–established step in ontology learning from texts. Terms are surface realisations of domain–specific concepts and represent, for this reason, a basic prerequisite for more advanced ontology learning tasks. In principle, they need be recognized whatever the surface form they show in context, irrespectively of morpho–syntactic and syntactic variants. For our present purposes, a term can be a common noun as well as a complex nominal structure with modifiers

(typically, adjectival and prepositional modifiers). Term extraction thus requires some level of linguistic pre-processing of texts.

T2K looks for terms in syntactically chunked texts such as those illustrated in Section 2.1 (Figure 1). Candidate terms may be one word terms ("single terms") or multi–word terms ("complex terms"). The acquisition strategy differs in the two cases.

Single terms are identified on the basis of frequency counts in the chunked source texts, after discounting stop–words. The acquisition of multi–word terms, on the other hand, follows a two–stage strategy. First, the chunked text is searched for on the basis of a set of chunk patterns. Chunk patterns encode syntactic templates of candidate complex terms, interpreted and applied by the IDEAL compiler. The set of chunk patterns covers the main types of modification observed in complex nominal terms: i.e. adjectival modification (e.g. *organizzazione internazionale* 'international organisation'), prepositional modification (e.g. *tutela del territorio* 'protection of the territory'), including more complex cases where different modification types are compounded (e.g. *incenerimento dei rifiuti pericolosi* 'incineration of dangerous waste'). Secondly, the list of acquired potential complex terms is ranked according to their log–likelihood ratio (Dunning, 1993), an association measure that quantifies how likely the constituents of a complex term are to occur together in a corpus if they were (in)dipendently distributed, where the (in)dependence hypothesis is estimated with the binomial distribution of their joint and disjoint frequencies. We tested the log–likelihood ratio against other association measures such as mutual information, chi–square etc., log-likelihood faring consistently better than the others. Moreover this measure is known to be less prone to assigning high scores to very sparse pairs. It should be recalled that the log–likelihood ratio is commonly used for discovering collocations. Hence, we are treating complex terms as though they belonged to the more general class of collocations. However, T2K uses the log–likelihood ratio in a somewhat atypical way: instead of measuring the association strength between adjacent words, T2K measures it between the lexico–semantic heads of adjacent chunks. The main and often underestimated advantage of defining co–occurrence patterns over syntactic structures is that we can broaden our search space (the text window) in a controlled way, by making it sure that there is a syntactic pattern linking two adjacent lexical heads.

So far, acquisition of potential complex terms has involved chunk pairs only (bigrams). In T2K recognition of longer terms is carried out by iterating the extraction process on the results of the previous acquisition step. This means that acquired complex terms are projected back onto the original text and the acquisition procedure is iterated on

the newly annotated text. The method proves helpful in reducing the number of false positives consisting of more than two chunks (Bartolini *et al.*, 2005). Interestingly, the chunk patterns used for recognition of multi–word terms need not necessarily be the same across different iteration stages. In fact, it is advisable to introduce potentially noisy patterns only at later stages. This is the case, for instance, of coordination patterns.

The iterative process of term acquisition yields a list of candidate single terms ranked by decreasing frequencies, and a list of candidate complex terms ranked by decreasing scores of association strength. The selection of a final set of terms to be included in the TermBank requires some threshold tuning, depending on the size of the document collection and the typology and reliability of expected results. Thresholds define *a)* the minimum frequency for a candidate term to enter the lexicon, and *b)* the overall percentage of terms that are promoted from the ranked lists. Typical values for a corpus of about one million tokens are as follows: minimum frequency threshold equal to 7 for both single and complex terms; selected single terms are the topmost 10% in the ranked list; selected multi-word terms are the topmost 70% in the ranked list of potential complex terms.

### 3.2. Term organization and structuring

In the second extraction step, proto–conceptual structures involving acquired terms are identified. The basic source of information is no longer a chunked text, but rather the dependency–based analysis exemplified in Figure 2, with the original text containing an explicit indication of the multi–word terminology acquired at the previous extraction stage.

We envisage two levels of conceptual organization. Terms in the TermBank are first organized into fragments of head–sharing taxonomical chains, whereby *ambiente urbano* 'urban environment' and *ambiente marino* 'marine environment' are classified as co–hyponyms of the general single term *ambiente* 'environment'.

Moreover, T2K clusters semantically–related terms by using CLASS, a distributionally–based algorithm for building lexico–semantic classes (Allegrini *et al.*, 2003). According to CLASS, two terms are semantically related if they can be used interchangeably in a statistically significant number of syntactic contexts. The starting point for the CLASS algorithm is provided by a dataset of dependency triples – $< T, C, s >$ –, where $T$ is a target linguistic expression, $C$ is a linguistic context for $T$, and $s$ is the particular syntagmatic dependency relation between $T$ and $C$. For our present concerns, variables are interpreted as follows:

1. $T$ corresponds to an acquired term in the TermBank;

2. $s$ stands for either a subject or a direct object dependency relation;

3. $C$ corresponds to a verb with which $T$ is attested to co–occur as a subject or a direct object. In fact, of all verb–term pairs attested in the corpus only a subset of highly salient such pairs is considered for clustering by CLASS. Light verbs such as *take* or *make* are likely to give very little information about the semantic space of the terms they select in context. Hereafter we shall refer to the set of highly salient verbs keeping company with subject/object $T$ as the *best verbs for* $T$, or $BVT$. For each term $T$, $BVT$ contains those verbs only whose strength of association with subject/object $T$ (measured by the log–likelihood ratio) exceeds a fixed threshold.

For all terms (both single and complex) in the TermBank, we extract from the dependency–annotated text the best verb/subject and verb/object pairs. CLASS then computes the degree of semantic relatedness between two terms $T_1$ and $T_2$ by measuring the degree of overlapping between $BVT_1$ and $BVT_2$, according to the metric described in Allegrini *et al.*, (2003). This corresponds to the assumption that the semantic similarity between two terms is a function of the possibility for the entities denoted by the terms to be involved in similar events, where the latter are expressed by the term best verbs. The cluster of terms semantically related to a target term $T$ is finally ordered by decreasing similarity scores with respect to $T$. For each term, the user can define the maximum number of related terms to be returned by the system; this parameter can be set on the basis of the user's needs (it should be kept in mind that going down in the ranked list of related terms the semantic distance from $T$ increases; therefore, it becomes more likely to find spurious associations).

## 4. Ontology learning from legislative texts: a case study

In this section we summarise the results of a case study carried out on a corpus of legal texts belonging to the environmental domain (Venturi, 2006).

### 4.1. Corpus description and preprocessing

The corpus consists of 824 legislative, institutional and administrative acts concerning the environmental domain, for a total of 1.399.617 word tokens, coming from the BGA (*Bollettino Giuridico Ambientale*) database edited by the Piedmont local authority for environment.[1] The

[1] http://extranet.regione.piemonte.it/ambiente/bga/

Table I. An excerpt of the automatically acquired TermBank

| ID | Term | Freq | Lemmatised headwords |
|---|---|---|---|
| 2192 | acqua calda | 11 | acqua caldo |
| 974 | acqua potabile | 36 | acqua potabile |
| 501 | acqua pubblica | 121 | acqua pubblico |
| 47 | acque | 1655 | acqua |
| 2280 | acque costiere | 10 | acqua costiero |
| 2891 | acque di lavaggio | 6 | acqua lavaggio |
| 2648 | acque di prima pioggia | 8 | acqua pioggia |
| 3479 | acque di transizione | 5 | acqua transizione |
| 1984 | acque meteoriche | 12 | acqua meteorico |
| 1690 | acque minerali | 16 | acqua minerale |
| 400 | acque reflue | 231 | acqua refluo |
| 505 | acque sotterranee | 120 | acqua sotterraneo |
| 486 | acque superficiali | 131 | acqua superficiale |
| 2692 | acque utilizzate | 8 | acqua utilizzato |

corpus includes acts released by three different agencies, i.e. the European Union, the Italian state and the Piedmont region, which cover a nine years period (from 1997 to 2005). It is a heterogeneous document collection including legal acts such as national and regional laws, european directives, legislative decrees, etc. as well as administrative acts such as ministerial circulars, decisions, etc.

### 4.2. The legal–environmental TermBank

Table I contains a fragment of the automatically acquired TermBank. For each selected term, the TermBank reports its prototypical form (in the column headed "Term"), its frequency of occurrence in the whole document collection, and the lemma of the lexical head of the chunk covering the term (see column "Lemmatised headwords"). The choice of representing a domain term through its prototypical form rather than the lemma (as typically done in ordinary dictionaries) follows from the assumption that a bootstrapped glossary should reflect the actual usage of terms in texts. In fact, domain-specific meanings are often associated with a particular morphological form of a given term (e.g. the plural form). This is well exemplified in Table I where the acquired terms headed by *acqua* 'water' can be parted into two groups according to their prototypical form: either singular (e.g. *acqua potabile* 'drinkable water') or plural (e.g. *acque superficiali* 'surface runoff'). It

should be noted, however, that reported frequencies are not limited to the prototypical form, but refer to all occurrences of the abstract term.

As expected from the peculiar nature of processed documents, the acquired TermBank includes both legal and environmental terms. Since the two classes of terms show quite different frequency distributions, different acquisition experiments were carried out by setting different thresholds (see Section 3.1). By using standard thresholds with respect to corpus size, we obtained a TermBank of 4.685 terms (both single and multi–word terms): the selected minimum frequency threshold for both single and multi-word terms was 7, the percentage of selected terms from the ranked lists was 10% in the case of single terms and 70% for multi–word terms. Yet, in this TermBank, environmental terms were scarcely represented due to their high rank (and low frequency) according Zipf's law. Since the focus of our interest was on both types of terminology, we carried out new acquisition experiments by reducing the minimum frequency thresholds to 5 and 3. In both experiments, the number of acquired environmental terms increased, unfortunately together with noisy terms. For instance, with the minimum frequency threshold set to 3, the numer of extracted terms is more than doubled, i.e. it is equal to 11.103.

Evaluation of acquired results was carried out with respect to the TermBank of 4.685 terms (i.e. the one obtained by setting the minimum frequency threshold equal to 7). Due to the heterogeneous nature of the terms in the glossary, belonging to both the legal–administrative and the environmental domains, two different resources were taken as a gold standard: the *Dizionario giuridico* (Edizioni Simone) available online[2] was used as a reference resource for what concerns the legal domain (henceforth referred to as Legal_RR), and the *Glossary of the Osservatorio Nazionale sui Rifiuti* (Ministero dell'Ambiente) available online[3] for the environmental domain (henceforth referred to as Env_RR), which contain respectively 6.041 and 1.090 terminological entries recorded in their prototypical form. For evaluation purposes, different types of matches were taken into account. Besides the full match between the T2K term and the term in the reference resource, different types of partial matches were also considered, i.e.:

1. the same term appears both in the T2K TermBank and in the gold standard resource but under different prototypical forms: this is the case, for instance, of the term *accordi di programma* 'programmatic agreement' which appears in the plural form in T2K and in the singular form in Legal_RR. At this level, two terms may also differ

[2] http://www.simone.it/cgi-local/Dizionari/newdiz.cgi?index,5,A
[3] http://www.osservatorionazionalerifiuti.it/ShowGlossario.asp?L=Z

for the prepositions linking the nominal headwords of a complex term, as in the case of *acquisizione dati* vs *acquisizione* ***di*** *dati* 'acquisition of data' or *abbandono* ***di*** *rifiuti* vs *abbandono* ***dei*** *rifiuti* 'waste abandon';

2. the gold reference resource contains a more general term whereas T2K acquired one of its hyponyms: this is the case of the T2K term *abrogazione di norme* 'repeal of rules', which in Legal_RR occurs in its more general form *abrogazione* 'repeal';

3. the reverse case with respect to 2 above, i.e. the gold reference resource contains a more specific term with respect to T2K which extracted a more general term, typically its hyperonym: e.g. *agente di polizia* 'policeman' (T2K) vs *agente di polizia giudiziaria* 'prison guard' (Legal_RR).

In the cases described in 2 and 3 above, a distinction is made – again – between matches concerning the prototypical form and matches at the level of stemmed words.

The results of the evaluation carried out on the basis of the criteria described above can be summarised as follows: in 51% of the cases a match, either full or partial, was found between the T2K glossary and the references resources; in particular, 89% of identified matches was concerned with legal terms and 34,5% with environmental ones, with a 23,5% of terms occurring in both reference resources. The question arising at this point is whether the remaining 49% of terms for which no match was found was represented by errors and noisy terms or were domain–specific terms not appearing in the selected reference resources. In order to answer this question, we selected two additional resources available on the Web: the list of keywords used for the online query of the *Archivio DoGi (Dottrina Giuridica)*[4] for the legal domain, and the thesaurus *EARTh (Environmental Applications Reference Thesaurus)*[5] for the environmental domain, against which a manual evaluation was carried out for 25% of the automatically acquired T2K glossary. The results are quite encouraging: by including these two richer reference resources, the percentage of matching terms increased to 75,4%. This percentage grows up to 83,7% if we also include terms which, in spite of their absence in the selected reference resources, were manually evaluated as domain–relevant terms: this is the case, for instance, of the terms *anidride carbonica* 'carbon dioxide' for what concerns the environmental domain or *beneficiari* 'beneficiary' for the legal one. The percentage of

[4] http://nir.ittig.cnr.it/dogiswish/dogiConsultazioneClassificazioneKWOC.php
[5] http://uta.iia.cnr.it/earth.htm#EARTh%202002

manually detected errors is 21,1%, also concerning some of the terms for which a partial match was detected. Whereas on the basis of these results it can be claimed that the accuracy of T2K for what concerns term extraction is quite high, nothing can be said as far as recall is concerned. As a matter of facts, the selected reference resources could not be used for this specific purpose due to their wider coverage, not circumscribed to the environmental domain.

### 4.3. Proto–conceptual organisation of terms

A first step towards the conceptual organization of terms in the TermBank consists in building taxonomical chains. This is to say that single and multi–word terms are structured in vertical relationships providing fragments of taxonomical chains such as the one reported below:

```
applicazione
   applicazione dei paragrafi
   applicazione dell' articolo
   applicazione della direttiva
   applicazione della legge
   applicazione della tariffa
   applicazione delle disposizioni
   applicazione delle sanzioni
      applicazione delle sanzioni amministrative
      applicazione delle sanzioni previste
   applicazione del presente decreto
   applicazione del regolamento
   applicazioni di quarantena
```

where the acquired direct and indirect hyponyms of the term *applicazione* 'enforcement' are reported. In this example, it can be noticed that terms sharing the head only are the direct hyponyms of the root term. Further hyponymy levels can be detected when two or more multi–word terms share not only the head but also modifiers, as in the case of the *applicazione delle sanzioni amministrative* 'enforcement of administrative sanctions' with respect to the more general term *applicazione delle sanzioni* 'enforcement of sanctions'.

With minimum frequency threshold set to 7, the numer of extracted hyponymic relations is 2.181 referring to 272 hyperonym terms; with the threshold set to 3, identified hyponymic relations increase to 6.635 regarding 454 hyperonym terms.

The second structuring step performed by T2K consists in the identification of clusters of semantically related terms which is carried out on

the basis of distributionally–based similarity measures (see Section 3.2). In what follows, clusters of semantically related terms are exemplified for both domains:

```
disposizioni 'provision'
   norme, disposizioni relative, decisione, atto, prescrizioni
legge 'law'
   regolamento, protocollo, accordo, statuto, amministrazioni comunali
inquinamento 'pollution'
   danno ambientale, inquinamento marino, effetti nocivi, conseguenza,
   inquinamento atmosferico
impatto ambientale 'environmental impact'
   esposizione, danno, esigenze, conseguenza, pericolo
```

For each target term, the set of the first 5 most similar terms is returned, ranked for decreasing values of semantic similarity. With the minimum frequency threshold set to 7, the number of identified related terms is 3.448 referring to 665 terminological headwords.

As illustrated in Section 3.2, these clusters of related terms were computed with respect to the most salient verbs associated with each target term: for instance, for *disposizione* 'provision' the most strongly associated verbs included *applicare* 'enforce', *adottare* 'pass', *abrogare* 'repeal', *decorrere* 'to have effect from' etc., whereas for *inquinamento* 'pollution' they range from *combattere* 'fight against', *ridurre* 'reduce', *prevenire* 'prevent', *eliminare* 'eliminate' to *causare* 'cause', *provocare* 'bring about' and *controllare* 'watch'. The terms similarity chains resulting from context-sensitive similarity measures are then merged and ranked according to decreasing similarity weights. It should be appreciated that in these clusters of semantically related words different classificatory dimensions are inevitably collapsed; they include not only quasi–synonyms (as in the case of *disposizioni* 'provision' and *norme* 'regulations' or *inquinamento* 'pollution' and *danno ambientale* 'environmental damage'), hyperonyms and hyponyms (e.g. *inquinamento* 'pollution' and *inquinamento atmosferico* 'atmospheric pollution'), but also looser word associations. As an example of the latter we mention the relation holding between *legge* 'law' and *amministrazione comunale* 'municipal administration', or between *pericolo* 'danger' and *conseguenza* 'consequences' and the environmental term *impatto ambientale* 'environmental impact'.

## 5. Conclusions and further directions of research

We reported preliminary but extremely encouraging results of the application of an automatic ontology learning system, T2K, on a corpus of Italian legislative texts in the environmental domain. Our work shows that the incremental interleaving of robust NLP and machine–learning technologies is the key to any attempt to successfully face what we termed the acquisition paradox. By bootstrapping base domain–specific knowledge from texts through knowledge–poor language tools we can incrementally develop more and more sophisticated levels of content representation. In the end the purported dividing line between language–knowledge and domain–specific knowledge proves to be untenable in language use, where language structures and bits of world–knowledge are inextricably intertwined.

There is an enormous potential for this bootstrapping technology. Acquired TermBanks can be transformed into semantic networks linking identified legal and environmental entities. Current lines of research in this direction include a) semi–automatic induction and labelling of ontological classes from the proto–conceptual structures identified by T2K, and b) the extension of the acquired ontology with concept–linking relations (first steps in this direction are reported in Venturi, 2006).

Our experiments also hightlighted some interesting open issues which need to be tackled in the near future. As pointed out in Section 4.2, running T2K on a corpus of legislative and administrative acts results in a two–faced terminological glossary, which includes terms belonging to both the legal–administrative and environmental domains. Estabilishing the domain relevance of each acquired term represents a central issue when dealing with legal–administrative texts. Some preliminary experiments have already been carried out in order to semi–automatically identify the domain–relevance of each acquired term. In particular, terminology acquisition was carried out with T2K on thematically different legislative corpora. By comparing the TermBanks automatically extracted from different corpora, we could classify the terms belonging to their intersection as belonging to the legal–administrative lexicon. This is in line with the contrastive approach to term extraction proposed by Basili *et al.* (2001). Similarly, the relevance of environmental terms will be validated by running terminology extraction on the environmental literature.

## References

Allegrini, P., Montemagni, S. and V. Pirrelli. Example-Based Automatic Induction Of Semantic Classes Through Entropic Scores. *Linguistica Computazionale*, 1-43: 2003.

Bartolini, R., Lenci, A., Montemagni, S. and V. Pirrelli. Grammar and Lexicon in the Robust Parsing of Italian. Towards a Non-Naïve Interplay. In *Proceedings of the International COLING-2002 Workshop "Grammar Engineering and Evaluation"*, Taiwan 2004.

Bartolini, R., Lenci, A., Montemagni, S. and V. Pirrelli. Hybrid Constrains for Robust Parsing: First Experiments and Evaluation. In *Proceedings of LREC 2004*, Lisbon 2004.

Bartolini, R., Giorgetti, D., Lenci, A., Montemagni, S. and V. Pirrelli. Automatic Incremental Term Acquisition from Domain Corpora. In *Proceedings of the 7th International conference on "Terminology and Knowledge Engineering" (TKE2005)*, Copenhagen 2005.

Basili, R., Moschitti, A., Pazienza, M.T. and Zanzotto, F.M. A contrastive approach to term extraction. In *Proceedings of the 4th Conference on Terminology and Artificial Intelligence (TIA2001)*, Nancy, France, 2001.

Buitelaar, P., Cimiano, P., and B. Magnini. Ontology Learning from Text: an Overview. In Buitelaar et al. (eds.), *Ontology Learning from Text: Methods, Evaluation and Applications* (Volume 123 Frontiers in Artificial Intelligence and Applications): 3–12, 2005.

Dell'Orletta, F., Lenci, A., Marchi, S., Montemagni, S. and V. Pirrelli. Text-2-Knowledge: una piattaforma linguistico-computazionale per l'estrazione di conoscenza da testi. In *Proceedings of the SLI-2006 Conference*: 20–28, Vercelli 2006.

Dunning, T. Accurate Methods for the Statistics of Surprise and Coincidence. *Computational Linguistics*: 19(1), 1993.

Federici, S., Montemagni, S. and V. Pirrelli. Shallow Parsing and Text Chunking: a View on Underspecification in Syntax. In *Proceedings of the Workshop On Robust Parsing*, in the framework of the European Summer School on Language, Logic and Information (ESSLLI-96), Prague 1996.

Lame, G. Using NLP techniques to identify legal ontology components: concepts and relations. *Lecture Notes in Computer Science*, Volume 3369: 169–184, 2005.

Sais, J. and P. Quaresma. A Methodology to Create Legal Ontologies in a Logic Programming Based Web Information Retrieval System. *Lecture Notes in Computer Science*, Volume 3369: 185–200, 2005.

Valente, A. Types and Roles of Legal Ontologies. *Lecture Notes in Computer Science*, Volume 3369: 65–76, 2005.

Venturi, G. L'ambiente, le norme, il computer. Studio linguistico–computazionale per la creazione di ontologie giuridiche in materia ambientale. Degree Thesis, Manuscript, December 2006.

Walter, S. and M. Pinkal. Automatic extraction of definitions from german court decisions. In *Proceedings of the COLING-2006 Workshop on Information Extraction Beyond The Document*: 20–28, Sidney 2006.

# Semantic Spaces and Multilingualism in the Law: The Challenge of Legal Knowledge Management

Doris Liebwald
*Vienna Center for Computers and Law, Austria*
*d@liebwald.com*

**Abstract.** It is the concern of the author to arrange cogitations and experiences she gained by collaborating in relevant international project works, by conducting scientific studies regarding legal knowledge representation and by teaching legal information retrieval. The main focus is the demonstration of problems of communication within and between humans and legal information systems, which are often hidden, overlooked or ignored. The author uses the concept "semantic spaces" to describe and explain semantic related difficulties detected in legal knowledge bases and data retrieval. Realization of these various semantic spaces might help further work in this area. Emphasis is also placed on the problem of multilingualism and diversity of legal cultures in EU legislation. The practical examples of the EU tools N-Lex and EUROVOC are used to illustrate the various situations, the current limits and the specific requirements and information needs in multilingual and cross-national legal information retrieval.

**Keywords:** Semantics, linguistics, information retrieval, knowledge presentation, legal language, multilingualism, cross-national IR, diversity of legal cultures and traditions, ontology, thesaurus, European Union law, national law, EUROVOC, N-Lex.

## 1. Introduction

This paper deals with machine processible semantics. Of particular concern are the following questions. Firstly, if it is possible, how can the normative and the real world be represented in machine executable language? Secondly, what problems must be resolved in order to accomplish this task? Numerous previous attempts to represent the meaning of legal concepts and the knowledge related to those concepts, especially in regards to coping with the step from simple string matching to an interpretation and comprehension of semantics, have proven tedious and labor-intensive.

What are the goals for these efforts? The goals can be divided into two categories: on one side is the user-oriented editing of legal information to provide experts as well as laymen easier access to legal documents; on the other is the implementation of machine-processible representation of legal norms to create systems which are capable of

applying legal rules or supporting humans in the application of the rules[1].

Communication takes places within and between a wide range of different semantic spaces[2]. In the area of law it is important to consider the various perceptions and information needs of the large array of people involved in the process. Examples are that of a judge, whose focus is on case-solving, the application-oriented approach of an administrative officer, the systematical point of view of a legislative drafter, or that of the persons subject to the law, i.e. the layperson. Even in the domain of "*legal informatics*", different semantic spaces exist and cause communication errors between legal and computer experts. While the computer scientist uses the syntax and semantics of a programming language, the lawyer considers the treatment of legal conceptualities, which are not easy for the legal expert to formulate in a computer sensitive way.

Another important element is intelligibility. Intelligibility is not inherent in the text; it is rather a process of understanding – a constructive, mental activity. Background knowledge and the intention of the reader as well as the design, composition and characteristics of the text play an important role. When reading text, common sense knowledge, factual knowledge, and the individual semantic spaces of the reader are activated. Therefore reading comprehension, which is a knowledge dependant mental representation, goes beyond what is explicitly communicated by the text. Is it possible to represent textual knowledge and implicit human knowledge with machines? It is indeed possible, however, only partially.

Ideally, semantic editing begins at the origin, such as in the preparation phase of a document, e.g. the draft bill. In this way knowledge about the realization of a law (e.g. explanatory notes, expert reports, opinions, etc.) and other metadata may be correctly related at source. Further applications may reuse and exploit this knowledge base to support further legislation (also amendments, impact assessments, follow up costs, etc.), execution of the enacted law, and decision making processes as well as information retrieval and document and knowledge management in general.

---

[1] Prominent examples are automatic contracting in e-commerce and rule-based systems for public administration or large insurance companies.

[2] The author introduces the concept "semantic space" to point up the different interpretations and spaces of meanings attached to a specific phenomenon or concept. The more similar different semantic spaces are, the easier communication will take place. People and/or machines not sharing a common or at least very similar semantic space run the risk of more or less obvious communication errors. See the in-depth analysis in Liebwald (2007): Semantische Räume als Strukturhintergrund der Rechtsetzung *("Semantic Spaces as Structural Patterns of Legislation")*.

### 1.1. Terms and Semantics

A term obtains conceptual content by the semantics assigned to it. Semantics[3] refers to the meanings attached to words, expressions and sentences, and are not part of the syntax; semantics comes from the "outside" and is constructed by individual mental models. Semantics are needed to turn terms into contextual concepts[4]. Term relates to the exterior, concept to the semantic content.

Concepts are used to characterize and to distinguish phenomena, whereby, dependent from the observer, each phenomenon may feature different semantic spaces. Individuals use different notions and different pictures of reality. Therefore, the intention of the author of this paper may differ to the interpretation of the paper by the reader. A certain notion of reality is not necessarily true or false, rather, it may be considered as more or less appropriate or functional. But even one particular observer may, dependent on the respective context, interpret one particular phenomena in different ways. For instance, the mother may relate the term *warmth* to love and security. The physicist, however, may offer a definition on the transfer of thermal energy. In the case that the mother and the physicist meet in a cold room, they will attach the same or at least a very similar meaning to the term *warmth*. Can the same be said about the term *warming*?

The meaning assigned to a phrase or sentence and therefore the interpretation and understanding of (technical) language may not only significantly differ between diverse organizations or expert circles, but also between the individuals participating. Furthermore, semantics are dynamic: they may change, e.g. due to new experiences or knowledge, changes in reality, progress.

In a joint semantic space, a space of mutual understanding, the semantics must follow a logic, which is shared by all members of this space.[5] Semantic spaces can be defined as networks of concepts that are used to describe the world as well as for behavioral orientation of the individuals acting in these spaces. Therefore semantic interoperability

---

[3] From *semantikos* (significant meaning), Greek; derives from *sema, semeion* (sign).

[4] According to ISO 1087 a concept is defined as *"a unit of thought constituted through abstraction on the basis of properties common to a set of objects"*; this definition is accompanied by the note *"concepts are not bound to particular languages; they are, however, influenced by the social or cultural background."*

[5] Compare Uschold's definition of ontology: *"An ontology is a shared understanding of some domain of interest."* Uschold/Gruninger (1996): Ontologies: Principles, Methods and Application (1996).

is of particular importance. Semantic interoperability[6] exists where the accurate meaning of information is understood and interpreted in the same way by all individuals and applications involved. All the actors must share the same model of what the data represents. The necessary linkage of several semantic networks of concepts necessitates a network of semantic spaces.

### 1.2. Legal Concepts[7]

A legal practitioner applies conceptual thinking and legal structural knowledge that she or he gained over long-term training. The complexity of law demands an abstract, differentiating, economical, and functional technical language ("legal language") which is able to represent the structures and meanings in law. The law is not just a collection of mechanical if/then-rules; based on the same facts and on the same legal rules, legal experts may indicate contradictory solutions. A correct syllogism may be overruled by social conventions, principles or extraordinary circumstances. Although where explicit knowledge exits, some legal problems may not be resolved simply and legal decisions will not always be predictable; in other cases the legal expert may be confronted with controversial facts.

Law is based on text and language and language is dependent on interpretation. Even if the lawmaker is anxious to reach maximal precision in legal texts and concepts, she or he will never reach absolute precision, because language itself is often ambiguous. Similarly, vague legal concepts can be considered an answer to missing accuracy of reality. Furthermore there exists some deliberate vagueness of legal concepts (or perhaps even deliberate incomprehensibility of legal texts). Reasons for this could be to cover future, not yet predictable circumstances or to cover at least all typical cases, to leave space for more specific rules, judicial discretion and interpretation, or just because more "accurate" political consent is missing.

Where legal rules are implemented in informatics systems, classical logic of jurisprudence and symbolic logic of informatics encounter one another. Open legal concepts, inherent dynamics of law, system models and syntactic ambiguities prove to be extremely problematic, whereby vague concepts seem to be the largest obstacle to overcome.

---

[6] Galinski follows the semiotic triad and cuts more accurately into a syntactic, pragmatic and conceptual level of semantic interoperability. Galinski (2006): Wozu Normen? Wozu semantische Interoperabilität? (*"Why Norms? Why Semantic Interoperability?"*).

[7] For a competent and comprehensive scholarly piece see Bydlinski (1991): Juristische Methodenlehre und Rechtsbegriff[2] (*"Legal Methodology and Nomen Juris"*).

### 1.3. Semantic Spaces in Law

The legal language cannot be considered as one semantic space, but rather a network of semantic spaces. Therefore it is not sufficient to only differentiate between legal experts and laypersons, since even between and within various groups of legal experts, the concepts, document types, styles of writing and parlance may vary.[8] Each field of law forms its own specific concepts and structures, which all show significant differences in their semantics. This is also true within legislation, administration, justice and doctrine. In some cases, when a draft bill, the enacted law and subsequent amendments are compared, there is a substantial shift in semantics; in other cases, judges' interpretations of a constant legal rule may change[9]. Where legal experts interact with other experts, the differences in the semantics of jargon may also have an effect, e.g. in reports, opinions, studies, comments. In such groups hidden misunderstandings are "pre-programmed". Divergent semantic spaces of different national legal systems or of national legal languages in comparison to the EU legal language are, however, more obvious. Nevertheless, the identification and expression of the subtle differences of similar concepts that arise from various national legal traditions is a sophisticated process.

### 1.4. The Problem of Multilingualism and Cultural Diversity in EU-Law

The EU currently embraces 27 Member States and has 23 official languages[10]. Legislation and documents of major public importance or interest are produced in all official languages, but most of the institutions' work is available in French and/or English only. Communication with the EU and its institutions by governments, civil servants, businesses and citizens may take place in any of the official languages.

Especially in regards to legal texts, multilingualism and diversity in legal culture pose intractable situations. Of course, EU legislation is translated into 23 languages, but the EU legal language and the specific

[8] Consider also e.g. the different semantic spaces of a public appointed/sworn expert, an eye-witness, the victim, the offender, the attorneys, the judge, the jury, the media, a person who has the power of pardon, etc.

[9] See e.g. Warta (2005): Zauberworte – Verwandlungen des Gleichheitsgrundsatzes in der Judikatur des österreichischen Verfassungsgerichtshofes (*"Magic Words – Metamorphoses of the Principle of Equality in the Legal Practice of the Austrian Constitutional Court"*).

[10] Some languages spoken in Member States (e.g. Catalan, Welsh, Basque, Breton, Sardinian) don't have the official EU language status. English, French and German are the three strongest languages within the EU.

concepts chosen do not correspond with the national legal language and concepts of the respective Member State to a very high degree.[11] 27 Member States interpret the same legal text, each influenced by its own political system, legal tradition, legal language and concepts, and overall legal view. Member States are required to implement EU legislation into their existing framework of national legislation, and these frameworks are not congruent with one another to varying degrees. Within the EU most countries belong to the civil law tradition, with the exceptions of Ireland and the United Kingdom. In some countries, the *"Länder", or states*, have minor legislative importance, but this is not true in all countries, e.g. Germany, Austria, and Belgium. Even where the same language is used (e.g. Austria, Germany), the legal systems, its structures, hierarchies and legal terminology differ. Therefore, e.g. one particular EU Directive[12] may be implemented in more than 27[13] different ways. Furthermore the national law of the Member States is not translated into the official languages of the EU. Thus, it is very difficult for the EU institutions to watch, compare and correct implementation measures, and it is also very difficult for governments, businesses and citizens to locate relevant cross-national legal information.[14]

[11] *Lesmo* et al. give a descriptive example by using the concept "*in clear and comprehensible manner*" taken from the Directive on Distance Contracts 97/7/EC. The authors compare the conditions a distance seller has to fulfill to provide a distance contract in clear and comprehensible manner under the U.K. (*"clear and comprehensible*"), German (*"klar und verständlich"*) and the Italian (*"chiaro e comprensible"*) legal system. Finally they point out that the main foci (form or the writing of the information must be clear and legible; information must be intelligible by the consumer; language of the information must be that of the consumer) set to identify a *"clear and comprehensible manner"* vary in all cases. See Lesmo et al. (2005): The next EUR-Lex: What should be done for the needs of lawyers belonging to different national legal systems?

[12] Most of EU legislation is made in the form of Directives. Contrary to EU Regulations, Directives are only binding on the Member States (not directly applicable to citizens) and usually leave some leeway as to the exact rules to be adopted.

[13] On the federal and the state level.

[14] The problem is not reduced to legislation. Schacherreiter analyzed two written statements on a decision of the European Court of Justice, one of a German, one of an English expert. Their conclusions are absolutely contrary: while the German expert (civil law) considers the findings of the ECJ indicatory and general applicable, the English expert (common law) cannot detect a new general rule, he rather considers the ruling of the ECJ an exception of the general rule, justified by very specific circumstances and facts. Schacherreiter (2006): Legal culture und europäische Harmonisierung (*"Legal Culture and European Harmonization*").

## 2. "Up to Date" Approaches: XML and Ontologies

Considering all of the semantic spaces, the relationships between semantic spaces and between concepts, and the inconsistency of natural language itself, is it now possible to put the legal and the corresponding real world knowledge into the machine? It is perhaps impossible or at least infeasible to make the machine automatically determine the exact meaning of legal text, but it is feasible to create machine-processible specifications of the semantics, at least to some extent. An overview of current approaches addressing these problems reveals two predominant keywords: XML and ontologies, most frequently connected to the concepts "Semantic-Web" or "Web 2.0"[15].

### 2.1. The Extendable Markup Language XML

The Markup Language XML has proven to be very helpful to structure legal texts and to allocate meta-data. With regards to further automatic processing it is a significant advantage to acquire the main features of a document already in its preparatory phase. Moreover, XML allows for logic notation, automated linkage and simplified visualization. Yet, it is primarily tied to syntax and proves less suitable to represent semantics. The level of semantics assigned to a document depends on how XML is applied. XML is normally used to tag the implicit semantics of the document structure only, and the tags are freely interchangeable and do not carry the actual meaning of the document's content. Often, errors are cause because legal texts are drafted in complex MS word templates incorporating many macros, and then converted into XML files. Therefore each new element, e.g. the marking of legal definitions, the representation of relations between different level instruments' or the denotation of roles would complicate the drafting of a document and inevitably go beyond the scope of the drafter. Furthermore law is dynamic – hence standards must enable subsequent changes.

The full potential XML offers has surely not yet been exploited, but there are other, perhaps more appropriate technologies available. It seems to be more useful to take XML as an ideal basis, on which

[15] *"The Semantic Web is an extension of the current Web in which information is given well-defined meaning, better enabling computers and people to work in cooperation. It is based on the idea of having data on the Web defined and linked such that it can be used for more effective discovery, automation, integration, and reuse across various applications."* Hendler et al. (2002): Integrating Applications on the Semantic Web, p. 676.

further layers of semantic technologies may be established, similar to the semantic web vision of Tim Berners-Lee[16].

Currently there are also some ambitious efforts to draft common European legislative XML standards to be shared by all EU Member States.[17] However, those attempts will face similar problems that arose within the N-Lex project described below. Existing legislative drafting standards of the Member States correlate to national legislative procedures. Differences in document structure, legal hierarchies etc. reflect specific national needs and legal systems. A country-independent data format will therefore either have to be restricted to a very simple common level or will otherwise not satisfy national requirements or even constrain to national process models. Therefore a common standard must be flexible enough to cover different national needs. Nevertheless, unification on a low level will facilitate document and information exchange, and in areas with a high degree of European harmonization such applications or shared tools may prove very useful. In the words of Michael Uschold, *"The more agreement there is, the less it is necessary to have machine processable semantics."*[18]

### 2.2. Legal Ontologies

Ontologies are knowledge models used to describe the meaning and context of information. They allow an accurate definition of relevant concepts and the representation of concept coherences, higher-level relations as well as logic structures, and are used to specify semantics in machine executable form (formal semantics). Their possible fields of application in law are manifold and range from information retrieval

[16] The source graphic of the oft-quoted layer model originates from Koivunen/Miller (2002): W3C Semantic Web Activity.

[17] See in particular the ONE-LEX project (ONtologies for European Laws in EXecutable format, Prof. Sartor, European University Institute/Florence, `http://castor.iue.it/`), and Sartor (2005): The ONE-LEX project and the informational unification of the laws of Europe. A broader overview on the state of the art is given by Biagioli et al. (eds.) (2007): Proceedings of the V Legislative XML Workshop (Florence 2006). See also the MetaLex Project (`http://www.metalex.nl/`) and the LEXML (`http://www.lexml.de/`) initiative. A different approach takes the new ESTRELLA project (European project for Standardized Transparent Representations in order to Extend Legal Accessibility, University of Amsterdam et al., `http://www.estrellaproject.org/`). The main objective of ESTRELLA is to develop a legal knowledge interchange format and to facilitate a market of interoperable components for legal knowledge-based systems, allowing public administrations and other users to freely choose among competing development environments, inference engines, and other tools. In the pilot applications, European and national tax legislation of two European countries will be modeled.

[18] Uschold (2003): Where are the Semantics in the Semantic Web?

across decision support systems to expert systems. Ontologies offer the advantage of rendering semantics more precisely; concurrently the nuances of relations allow for certain representation of ambiguity. The use of ontologies overcomes linear hierarchical structures, allows the integration of heterogeneous data sources and enables the step from text documentation to content documentation.

There exist three main techniques for ontology engineering: statistical approaches which are less laborious but entail a certain ambiguity, linguistic approaches whose reliability heavily depends on the application area and which are not sufficient in the field of law, and manual/intellectual methods, which – provided that there is a high degree of enthusiasm and motivation in their engineering – offer the best results, but are the most costly and time consuming. For most reasons it is advisable to combine statistical, linguistic and manual methods. Since statistical methods are more mature, subsequent manual adjustment is less laborious. Linguistic tools may solve well-known problems like synonymy, morphological changes of the word stem, compound words, etc. Such tools exist in varying levels of quality, but are costly.[19]

Additionally, there exist several types of ontologies, which can be roughly divided into meta-data ontologies, general ontologies to represent the world knowledge, specific domain ontologies, method- and task-oriented ontologies, and finally representative ontologies, which define only the frames of representation.

Methods are also diversified and range from WordNet-methods[20], which define concepts in natural language and go without a formal language for the definition of semantics, to rule-based systems with a high degree of formalization, e.g. Cyc[21], which uses millions of logic axioms, rules and other assertions to specify constraints on the individual objects and classes. Linguistically motivated ontologies like WordNet or in the legal field LOIS (Lexical Ontologies for legal Information Sharing)[22] are still primarily made for humans. The semantics are made explicit in an informal manner, in natural language definitions. Direct use of informally expressed semantics by machines is limited. For this

[19] In the English language a simple word stemming may already resolve a large part of the morphological problems. However, this does not apply to other languages.

[20] See `http://wordnet.princeton.edu/` and in particular Fellbaum (ed.) (1998): WordNet: An Electronic Lexical Database.

[21] See `http://www.cyc.com/` (there is also a list of publications at `http://www.cyc.com/cyc/technology/pubs`).

[22] LOIS is a multilingual legal thesaurus with natural language definition of legal terms based on the WordNet and the EuroWordNet (`http://www.illc.uva.nl/EuroWordNet/`) technology. See the LOIS homepage at `http://www.loisproject.org/` and Schweighofer/Liebwald (2007): Advanced lexical ontologies and hybrid knowledge based systems.

reason the semantics must be hardwired into application software to make the ontology usable for machines. But even in applications like Cyc, automated inference to process the semantics at runtime is limited. Cyc does not dynamically discern what content means; the meanings of terms and how to use them are hard coded by humans.

The use of ontologies for the formalization of the law is, however, not a new approach.[23] Today there are new implementation technologies available, which have given rise to numerous proposals and projects in this area[24]. The chosen approaches and methods are manifold, but a "universal valid code of practice" on how to engineer a legal ontology does not exist. Nevertheless, some critical points can be isolated.

The law is dynamic and consists of dissimilar, variable semantic spaces. Therefore ontologies need to be flexible and dynamic and must describe processes instead of static models. The formalization of implicit knowledge proved to be especially difficult. Application-oriented, specific domain ontologies (networks of meanings) are feasible at this stage. However, the cross-linking of different domain models and the interconnection of the concept spaces of world knowledge (the world model)[25] and legal knowledge (the domain models) are still substantial problems. Within a network of semantic spaces, overlapping, conflicting or even contradictory conceptualizations must be resolvable. Findings of related research, especially in the areas of forensic linguistics, comparative law and automatic text analysis, which could bridge the gap between conceptualization und stored information, seem to have been put aside in the fervor of ontology engineering but should be incorporated as much as possible.

Furthermore it can be clearly stated that ontologies are extremely cost and labor-intensive; they demand expert knowledge and a high level of consistency. The quality of the conceptualizations and their relation-

[23] See e.g. Hohfeld (1917): Fundamental Legal Conceptions as Applied in Judicial Reasoning; but also the findings of *Hart*, *Kelsen, etc.*

[24] An overview is given by Schweighofer/Liebwald (2007): Advanced lexical ontologies and hybrid knowledge based systems. More detailed is Benjamins' et al. (eds.) (2005): Law and the Semantic Web.

[25] The underlying problem may already exist in the modeling of the "necessary" world knowledge, in the "facts" (e.g. consideration of evidence, reconstruction/finding of facts, etc.). It cannot be ignored that models are always abstract – part of reality is lost in models. A nice example can be taken from the TRACS project. The TRACS prototype was developed about 1990 to check the consistency and completeness of a new (Dutch) traffic regulation. It drew inter alia the conclusion that a tram running on the tram-lane commits a traffic violation. This was due to the fact that Art. 10.1 stated that all vehicles except those mentioned in Articles 5 to 8 should use the drive-lanes. However, the tram is not mentioned in these articles, and the tram-lane is not a drive-lane. See Breuker et al. (2005): Use and Reuse of Legal Ontologies in knowledge engineering and Information Management.

ships are of utmost importance and cannot merely be replaced by a high number of low-quality concepts. The undertaking will be more worthwhile if ontologies allow for reuse. Again, this requires high quality, common design and compatible technologies. Nevertheless, ontology developers should always consider the specific needs of the intended application area(s) and user group(s).
Finally one must consider that a determination, standardization, or terminology normalization which is too strong or too stringent may also emerge as sort of "semantic shackle" which compromises diversity of language(s) and constrains further development. When dealing with European legal texts, merely reducing languages, legal systems and legal traditions to the highest common denominator will not contribute to a better mutual understanding. It is of highest importance to factor in national differences in legal language, concepts and structure. Contrary to e.g. a biological taxonomy, a legal ontology is not language and country independent.

## 3. Two Practical Examples: N-Lex and EUROVOC

Two practical examples, the new and experimental search-engine N-Lex and the traditional EUROVOC-Thesaurus, shall demonstrate the current problems within the EU caused by the diversity of legal traditions and semantic spaces in the law.

### 3.1. The Experimental N-Lex Project[26]

N-Lex is an attempt of the European Publications Office to provide a common gateway to national law of the EU Member States.[27] It is an experimental system, put online for test-use in April 2006. N-Lex allows users to search the national legal databases of 22 Member States using a single, uniform N-Lex search mask.

A user may choose the source country and then fill in one or more input fields. This query put to N-Lex is forwarded unaltered to the search form of the respective freely available national online-database. In the next step, N-Lex presents the original result set or result document in its main frame. For display of the search form and of some basic information on the respective national database the user may choose

[26] A more in-depth analysis is given by Liebwald (2007): Einheitsschnittstellen zu Rechtssystemen am Beispiel von N-Lex (*"Unified Interfaces to Legal Systems Using the Example of N-Lex"*).

[27] N-Lex is maintained by the Office for Official Publications of the EC; the application is publicly available at `https://europa.eu.int/celexdev/natlex/`.

her/his preferred language. However, a user lacking sufficient language abilities will not be able to formulate an adequate query or to assess the retrieved documents.[28]
Information on the respective national databases ("country information") is poor, and the user is left with many questions regarding completeness, authenticity and timeliness of the content, the document hierarchies and relationship of documents, or the technical functioning of the corresponding national systems – all of which influence the appropriateness of a search.

The search mask offers only a few input fields,[29] some of which may or may not function depending on the country selected. The input field for document numbers is not available for many countries, and where it is active, the necessary input format is not clear. In the advice section the N-Lex help-entry recommends against using the date/time span field, *"as it is . . . liable to produce zero responses"*. In fact, the results retrieved by using the date-field are not comprehensible, at least regarding searches for Austrian legal documents. Document numbers and date/time-span are, however, very important and in many cases even essential criteria for the identification of legal documents.
The original search forms of the national legal databases offer more sophisticated search fields and search functions. Their search masks are not only adapted to the national legal systems, the national document structure, and the national language(s), but also to the features and abilities of the respective technical system. Even search functions most typically used for legal information retrieval may have been implemented in very different ways. In the national systems digital information is available in different formats and comes along with country specific meta-data. A simple, unified search mask covering all of these different legal and technical systems nullifies many of the country specific functionalities, meta-data, and textual information (e.g. *"Länderrecht"*). Regarding the N-Lex system, the general principle that authenticity is directly related to the closeness to the original source is true.

In its current state, N-Lex displays many deficiencies and considers differences in the national legal and technical systems inadequately.[30] However, it is still in the experimental phase and some points of criticisms might become obsolete at a later time. At the very least, it offers

[28] Regarding the implementation of the EUROVOC thesaurus see the next section.
[29] Full text search, search in document titles, document type, document number, date of document.
[30] The EULEGIS (European Legal Information in a Structured Form, 1999-2001) research project already identified those problems. An overview on the EULEGIS reports is available at `http://www.it.jyu.fi/raske/publications.html`.

a single access point and has provided a first test that can be studied and improved upon.

### 3.2. The EUROVOC-Thesaurus

The multilingual und multidisciplinary EUROVOC-Thesaurus was originally built for processing the documentary information of the EU institutions.[31] It offers a controlled set of vocabulary covering 21 wide-ranging fields and more than 20 languages.[32] EUROVOC, however, contains "European" concepts, with a certain emphasis on the European legal language and parliamentary activities. It is an effective tool to index (European) documentary resources and to retrieve documents indexed by this means, but its general usability in information retrieval, especially in full text retrieval, is limited. The following two examples shall illustrate those restrictions.

The EUROVOC-Thesaurus, which is used in various applications, is also used in EUR-Lex, the gateway to EU law[33]. Legislative documents in EUR-Lex[34] are indexed according to EUROVOC, and the simple-search form allows a keyword search restricted to those EUROVOC descriptors. However, EUR-Lex contains a huge amount of documents[35] and only the upper levels of EUROVOC may be selected from the classification schema provided. Therefore, the use of EUROVOC descriptors usually results in a set of a few hundred, sometimes even of a few thousand documents. Of course, the system allows the user to refine the search by adding additional keywords, by selecting the document type or the date/time span. Alternatively the user can use the trial and error method and enter various EUROVOC descriptors from a deeper level. Admittedly, most of these choices assume additional knowledge about the document or about EUROVOC, which might not be available at this stage, or simply do not reduce the amount of result documents to a manageable number.

The implementation of the EUROVOC-Thesaurus in the experimental N-Lex application demonstrates the limits of such thesauri more obvi-

[31] EUROVOC is maintained by the Office for Official Publications of the EC and available at `http://eurovoc.europa.eu`.

[32] EUROVOC consists of more than 6000 concepts with a maximum depth of 8 levels. The 21 fields of the first level split up into some 130 micro-thesauri.

[33] EUR-Lex is also maintained by the Publications Office and available at `http://eur-lex.europa.eu/`.

[34] The ECJ case law is, however, not indexed by EUROVOC descriptors but by the case law directory code.

[35] According to the EUR-Lex FAQ (point 2.2.) *"it includes some 400000 references in several languages, 1400000 texts in total. An average of 15000 documents are added each year."*

ously. Within N-Lex, EUROVOC is intended to support full text search capabilities. National legal documents (legal documents produced by the Member States) are generally not linked to EUROVOC descriptors. Therefore a corresponding keyword search cannot be established. Users may either search and select a suitable descriptor in the target language or search for a descriptor in her/his preferred language and ask for the translation into the target language.
Due to the fact that EUROVOC uses European terminology, it is not convenient to search or to index national legal documents, even if those texts are partially based on European input requirements. Each Member State has its own legal tradition, legal system and legal terminology. A national indexer would in many cases choose different descriptors based on national legal traditions and interpret EUROVOC descriptors in a different way. Additionally, EUROVOC descriptors do not necessarily appear in the relevant national legal texts. On the contrary, more specific concepts, variants of concepts and specific national legal language terms are used within national codes and case law. Additionally, EUROVOC appears to be based to some extent on literal translations not indicating the exact implied meaning. European as well as literally translated concepts usually don't correspond to the terms and phrases a national user of a legal database would use naturally.[36] and the German translation *"persönliche Daten"*. Even though all German-speaking lawyers will understand this translation, the Austrian legal language uses the concept "*pesonenbezogene Daten*". Searching the Austrian law with the search term "*persönliche Daten*" will retrieve result documents, but not the relevant ones. It will mainly retrieve those texts containing the terms *"persönliche"* and *"Daten"* beyond the meaning of *"peronsenbezogene Daten"*.

Once the N-Lex user has chosen a fitting EUROVOC descriptor, the system sends the search question to the corresponding national databases. Most of these national databases execute a simple string search. Specific technical parlance, morphological changes, derivations, compounds, synonyms, polysems, etc. are therefore not considered.[37] Only those documents containing exactly the same character string are

[36] E.g. the German concept *"Datenschutz"* is a prominent concept in German and Austrian law, but is not covered by EUROVOC. Austrian and German lawyers will connect a specific concept regarding the protection of personal data processed by electronic means to the term *"Datenschutz"*. On the other hand, EUROVOC offers the English concept *"data-processing law"* with the German translation *"Datenrecht"*. What is *"Datenrecht"*? A test search concluded that *"Datenrecht"* is never used in Austrian or German legislation. EUROVOC also offers the concept "personal data"

[37] There are of course language and provider dependent differences (additional functions offered by the corresponding national provider will influence/better the result set).

sent back to the user.[38]
This accumulation of shortcomings produces incomplete result sets with low-recall, low-precision or empty result sets. This is contrary to the use of EUROVOC within EUR-Lex. Using EUROVOC in the way in which it is implemented in N-Lex wrongly assumes that all agents use exactly the same wording to state the same thing and that the same terms always have the same meaning. The correct conclusion is not that EUROVOC is generally a bad thesaurus of low quality but that it is being used for purposes other than originally intended and has not been adapted to such uses.

### 3.3. Excursus: The Semantic Spaces of the Persons Subject to the Law

To make the law more easily accessible for the persons subject to the law is an ambitious goal. The descriptions of the LOIS and the N-Lex projects stress the target to enable easier access to legal information for professional users as well as for laypersons. Both use the example of a family migrating to another EU Member State and searching for information regarding taxes, social insurance, childcare, etc. In fact neither LOIS nor N-Lex solve or even support such questions, at least at their current state. The semantic spaces of laypersons significantly differ from the semantic spaces of the lawmaker or legal expert. Citizens will use other concepts, other questions, and will have other information needs.[39] Usually they will not be able to retrieve relevant bills from legal databases or be able to identify the relevant articles therein. They will not understand the original text of a bill or the legal language, and they will need some complementary explanations in their common language. It is even more unlikely that citizens will understand the concepts, structure and language of a foreign legal system. It is not sufficient to enable easier access to the law by offering a choice of life situations from which a citizen may select, or by semantic translation of the common language information request, and then the presentation of the original legal texts to the citizen. This shall not, however, prevent establishing links to the

[38] EUROVOC offers e.g. the English concept *"protection of communications"* and the corresponding German concept *"Brief-, Post- und Fernmeldegeheimnis"*. This string is never used in Austrian legislation, even though the concept does exist. Searching the German law brings up 22 hits.

[39] Significantly there is a "*plain language guide to Eurojargon*" available in 20 languages at the Europe-server (`http://europa.eu/abc/eurojargon/index\_en.htm`). According to this site the guide was developed because euro-jargon can be very confusing to the general public. The language guide and the attached glossary contain in sum about 300 concepts and short descriptions, but the concepts are not linked to further information and the descriptions do not solve real life questions.

original legal sources where appropriate, but citizens primarily need citizen-tailored texts and issue-related information. In addition, citizens will appreciate supplementary information such as the responsible departments, contact data or references to further appropriate services. The approach to develop one combined system that serves experts and citizens is perhaps too ambitious and idealistic; such a system runs the risk of being a confusing compromise instead of serving both in the best possible manner.

## 4. Conclusions

It is an interesting matter that since the classic "Handbook of Legal Information Retrieval" edited by *Jon Bing* was published in 1984[40], improvement in legal information retrieval has not seen any major advancement. Quite to the contrary, information overload and increased demand for cross-national and cross-lingual legal information has amplified the basic problems. The handbook already points out many of the shortcomings a lawyer typically has to struggle with when searching for relevant legal documents. About 20 years later, authors such as Luuk Mathijssen, Peter Wahlgren and Doris Liebwald[41] as well as the common user still struggle with the very same problems. Legal information retrieval systems still do not represent legal structural knowledge, user friendliness regarding search strategies and input formats is lacking, and information about system functions and information content (completeness) is often not sufficient. Also, continuity, representation of time layers and consolidated versions are inadequate and different user situations and information needs are disregarded. Finally, finding the correct search terms is a game of chance, language approximation is minimal and even simple linguistic tools are missing.

Nevertheless, current developments in new technologies supporting communication in human/human, human/machine and machine/machine relations are promising. A shift from simple full text and keyword search to more sophisticated semantic querying appears to be within reach. Hopefully, these technologies will be used to serve the fundamental principles of accessibility and intelligibility of the law.

---

[40] A revised version is freely available at `http://www.lovdata.no/litt/hand/hand-1991-0.html`.

[41] See Matthijssen (1999): Interfacing between Lawyers and Computers; Wahlgren (1999): The Quest for Law; Liebwald (2003): Evaluierung juristischer Datenbanken (*"Evaluation of Legal Databases"*) and Liebwald (2005): An Evaluation of "New EUR-Lex": All Tasks Achieved and All Problems Solved?

## References


Benjamins, R. et al. (eds.): Law and the Semantic Web: Legal Ontologies, Methodologies, Legal Information Retrieval, and Applications. Lecture Notes in Computer Science Vol. 3369, Springer, Berlin et al. 2005

Biagioli, C. et al. (eds.): Proceedings of the V Legislative XML Workshop (Florence 2006). European Press Academic Publishing, Florence 2007

Bing, J. (ed.): Handbook of Legal Information Retrieval. Elsevier, New York 1984

Breuker, J.A. et al.: Use and Reuse of Legal Ontologies in knowledge engineering and Information Management. In: Benjamins (eds.): Law and the Semantic Web. Lecture Notes in Computer Science Vol. 3369, Springer, Berlin et al. 2005, 36-64

Bydlinski, F.: Juristische Methodenlehre und Rechtsbegriff[2] *("Legal Methodology and Nomen Juris")*. Springer, Wien 1991

Fellbaum, C. (eds.): WordNet: An Electronic Lexical Database. MIT Press, MA 1998

Galinski, C.: Wozu Normen? Wozu semantische Interoperabilität? (*"Why Norms? Why Semantic Interoperability?"*). In: Pellegrini/Blumauer (eds.): Semantic Web: Wege zur vernetzten Wissensgesellschaft. Reihe X.media.press, Springer, Berlin et al. 2006, 47-72

Hendler, J. et al.: Integrating Applications on the Semantic Web. Journal of the Institute of Electrical Engineers of Japan Vol. 122(10), October 2002, 676-680

Hohfeld, W.N.: Fundamental Legal Conceptions as Applied in Judicial Reasoning. The Yale Law Journal Vol. 26/8 (June 1917), 710-770

Koivunen, M.-R./Miller, E.: W3C Semantic Web Activity. In: Proc. of the Semantic Web Kick-off Seminar (Helsinki 2001). HIIT Publications, Helsinki 2002/1, 27-43 (available at `http://www.w3.org/2001/12/semweb-fin/w3csw`)

Lesmo, L. et al.: The next EUR-Lex: What should be done for the needs of lawyers belonging to different national legal systems? In: Proc. of the JURIX EU-Info Workshop. Brussels 2005 (available at `http://www.di.unito.it/\ensuremath{\sim}guido/PS/jurixWorkshopPaper.pdf`)

Liebwald, D.: An Evaluation of "New EUR-Lex": All Tasks Achieved and All Problems Solved? MR-Int 3/2005 (European Media, IP & IT Law Review), Verlag Medien & Recht, Vienna, 156-160

Liebwald, D.: Einheitsschnittstellen zu Rechtssystemen am Beispiel von N-Lex (*"Unified Interfaces to Legal Systems Using the Example of N-Lex"*). In: Schweighofer et al. (eds.): Aktuelle Fragen der Rechtsinformatik 2007. Boorberg, Stuttgart et al. 2007 (in print).

Liebwald, D.: Evaluierung juristischer Datenbanken *("Evaluation of Legal Databases")*. Verlag Österreich, Vienna 2003.

Liebwald, D.: Semantische Räume als Strukturhintergrund der Rechtsetzung ("*Semantic Spaces as Structural Patterns of Legislation*"). In: Bildungsprotokolle der Kärtner Verwaltungsakademie zu den Klagenfurter Legistik-Gesprächen 2006, Klagenfurt 2007 (in print).

Matthijssen, L.: Interfacing between Lawyers and Computers: An Architecture for Knowledge-based Interfaces to Legal Databases. Kluwer Law International, The Hague et al. 1999

Sartor, G.: The ONE-LEX project and the informational unification of the laws of Europe. In: Proc. of the Klagenfurter Legistikgespräche 2005. Bildungsprotokolle Vol. 12, Kärntner Verwaltungsakademie, Klagenfurt 2006, 193-202.

Schacherreiter, J.: Legal culture und europäische Harmonisierung *("Legal Culture and European Harmonization")*. Juridikum 2006/1, Verlag Österreich, Vienna, 17-21

Schweighofer, E./Liebwald, D.: Advanced lexical ontologies and hybrid knowledge based systems: First steps to a dynamic legal electronic commentary. AI&Law Journal Vol. 15 (February 2007), Springer International, NL

Uschold, M.: Where Are the Semantics in the Semantic Web? AI Magazine 24(3): Fall 2003, AAAI Press, Menlo Park, 25-36

Uschold, M./Gruninger, M.: Ontologies: Principles, Methods and Application. The Knowledge Engineering Review 11(2)/1996, Cambridge University Press, 93-115

Wahlgren, P.: The Quest for Law. Jure AB, Stockholm 1999.

Warta, P.: Zauberworte – Verwandlungen des Gleichheitsgrundsatzes in der Judikatur des österreichischen Verfassungsgerichtshofes (*"Magic Words – Metamorphoses of the Principle of Equality in the Legal Practice of the Austrian Constitutional Court"*). In: Schweighofer et al. (eds.): Aktuelle Fragen der Rechtsinformatik 2005. Boorberg, Stuttgart et al. 2005, 576-583.

# Legal Query Expansion using Ontologies and Relevance Feedback

Erich Schweighofer, Anton Geist
*Centre for Computers and Law*
*Section for International Law and International Relations*
*University of Vienna, Austria*

**Abstract.** The aim of our research is the improvement of Boolean search with query expansion using lexical ontologies and user feedback. User studies strongly suggest that standard search techniques have to be improved in order to meet legal particularities. Query expansion can exploit the potential of linguistic knowledge and successful user behaviour. First tentative results show the feasibility of our approach. A first search prototype has been built and tested in the area of European state aid law.

**Keywords:** Legal Information Retrieval, Ontologies, Query Expansion, Relevance Feedback

## 1. Introduction

Lawyers are knowledge workers and have to cope with a tremendous load of information of at least 1 GB of data (500 000 pages). In the legal domain, almost all available information is stored as text, most of the time in relatively unstructured forms (Stranieri and Zeleznikow, 2005). As work consists of solving legal problems, consultation of various texts is a prerequisite of legal work. This legal research can be outsourced to paralegals but in the very end good lawyers have to refine the quantity of relevant legal texts themselves.

Information and Communication Technology has dramatically altered legal research. Starting in the seventies with Boolean legal information systems, profiting from the internet revolution concerning on-line access, user interfaces and data handling, a very powerful and easygoing way of handling the mass of legal information was offered as the main ICT tool for legal knowledge management.

Boolean search has many advantages like rapidity, accuracy, and updating, but also one serious disadvantage. Users have to be very intelligent and highly trained in order to cope with the linguistic challenge of successful search. In order to get sufficiently good results users must know the appropriate terms and at least all synonyms, homonyms and polysems in a text corpus with more than 50 000 words. This more than Shakespearian endeavour (Shakespeare used about 20 000 words

in his works) usually ends in some failure, followed by iterative steps and frequent references to text books and commentaries.

Legal vocabularies contain open-textured terms, they are inherently dynamic. To a certain degree, this is necessary, because legal terms have to be flexible to be able to adapt to new life circumstances. Thus, legal concepts are ambiguous, their definitions vary depending on many factors like source and context. This allows for contradictions to arise from judicial problem solving. A "legal language", consisting of a complex structure of concepts, forms an abstraction from the text corpus as represented in legal databases. Such legal structural knowledge does not only contain interpretations of the meaning of legal terms, but also shows the (supposed) logical and conceptual structure. Bridging the gap between legal text archives and legal structural knowledge is a principal task of studying the law, and the key challenge in legal information retrieval.

Term frequencies do not help as much in law as in other domains. No redundancy exists in legal norms, but a lot of information is irrelevant in case law. Relevant texts parts may consist only of a short paragraph or even only of a single sentence in a very long legal document.

## 2. The Idea

The aim of our research is to improve the retrieval results of legal information systems.

On the one hand, we support the user with additional linguistic knowledge. In the last years, powerful legal ontologies have been developed that can be used for supporting querying as shown in the LOIS project (Dini et al., 2005). Legal text analysis has developed many methods that support the creation of ontologies.

One the other hand, we use search contexts to improve search queries. Legal information system providers have already stored information on search practices, and using query logs to improve search engine performance would be easy to implement.

Query expansion is a quite old technique for advanced search (Salton and McGill, 1986). But - unlike weighting citations ("Google's PageRank") - it never really took off, but remained in research labs.

The goal of our research project "Google the Law: Modern Text Retrieval in the Legal World" is the development of a methodology, a prototype and test applications for improved information retrieval using query expansion. This ambitious endeavour has reached the status of test applications although many improvements of our prototype are still

waiting for implementation. As a first test environment, we have chosen European State aid law.

The remainder of the paper is organized as follows: section 3 deals with related work, section 4 describes our methodology, section 5 the prototype and section 6 our test results. Last but not least section 7 draws conclusions and outlines future work.

## 3. Related work

Information retrieval (Salton and McGill, 1986; Frakes and Baeza-Yates, 1992) deals with the storage of documents in databases and their retrieval according to their relevancy to a query. This query is, at least in classical information retrieval systems, composed of key terms and subsequently matched with the index terms of all documents that are stored in the database. As a result to the query, the system returns those documents whose index terms match the query. It is important to note that only hints for relevant information are given.

Lawyers were eager to use information retrieval in working with the huge amounts of electronically available legal texts. It is no surprise that automated retrieval from large electronic legal document collections was one of the earliest applications of computer science to law (Moens, 2001). The limitations of information retrieval (only hints to information) and in particular of Boolean retrieval (need for exact terms and logical structure for queries) were never really liked. Single term searches seem to remain popular whereas theory considers them as quite unproductive, as they return many irrelevant hits and miss relevant ones.

Matthijssen developed a special interface for addressing four theoretical limitations in present legal information retrieval (Matthijssen, 1999): (1) the fact that the index of a database only partially describes its information contents, (2) the imperfect description of an information need by the query formulation, (3) the rough heuristics and tight closed world assumption of the matching function, and (4) the presence of the conceptual gap: the discrepancy between users' views of the subject matter of the stored documents in the context of their professional setting and the reduced formal view on these subjects as presented by information retrieval systems. Legal practitioners have to translate their information need - which they have in mind in the form of legal concepts - into a query, which must be put in technical database terms.

For the Norwegian jurisdiction (here two versions of the same language are used, Bokmål and Nynorsk), a special method called "conceptual text retrieval" was developed and is still successfully used. Queries

are described by a term class called "conceptor" consisting of a class of words representing the same idea (Bing, 1984). This idea derives from the NRCCL - Norwegian Research Center for Computers and Law - that has followed information retrieval research in law for more than 35 years and published famous books on that subject (Harvold and Bing, 1977; Bing, 1984) - and numerous articles (e.g. (Bing, 1987; Bing, 1995)).

The essential assumption of the so-called inference model is that the best retrieval quality is achieved with a ranking according to probability of relevance of the documents (Turtle, 1995). Bayesian inference nets are an elegant means of representing probabilistic dependencies and thus linguistic relations. The query representing information needs is extended via defined and computed dependencies.

Similar representations could also be achieved by a connectionist network containing nodes of terms, documents and authors. Synonym relations are represented in the nodes of terms (Rose, 1994). It may be also noted that a connectionist network seems to take most advantage of relevance feedback that may be used at a later stage of our project.

Legal publishers tried to cope with the linguistic problem by adding meta data (classification, thesauri, summaries etc.) to documents stored in legal information systems. European systems, in particular CELEX, are prominent for this approach that, however, did not get sufficient user support (Schweighofer, 2000). Users were simply not willing to learn all knowledge to use meta data. Hypertext (Bing, 1998) slightly improved the situation as browsing allowed easier use and learning in using meta data. The EUR-Lex (formerly CELEX) database still contains much meta data but it remains open if costs meet gains. Synonym lists are also partly added (e.g. in the Austrian LexisNexis system). Westlaw's WIN seems to have found the best and only solution: offer this support without interference by the user and at the highest quality available.

Ontologies (Gruber, 1993) constitute an explicit formal specification of a common conceptualization with term hierarchies, relations and attributes that makes it possible to reuse this knowledge for automated applications. The formalization must be on the one hand sufficiently powerful with regard to the knowledge representation, on the other hand it must offer functionalities for automation as well as tools to be produced automatically (see for lexically based ontologies (Hirst, 2003)).

Ontologies in law have some particularities. The motivations for the creation of legal ontologies are evident: common use of knowledge, examination of a knowledge base, knowledge acquisition, representation and reuse of knowledge up to the needs of software engineering (Bench-Capon and Visser, 1997).

After important preliminary work (e.g. (McCarty, 1989), (Hafner, 1981), (Stamper, 1991)), the frame-based ontology FBO of (Van Kralingen, 1995) and (Visser, 1995) as well as the functional ontology FOLaw of (Valente, 1995) achieved some prominence. Both were formalized with the description language ONTOLINGUA (Gruber, 1992; Gruber, 1993) and represent a rather epistemic approach. FOLaw has been used in the follow-up projects like ON-LINE, an architecture for artificial case solving, and CLIME/MILE with the test applications of classification of ships and maritime law (Winkels et al., 2002). The central difficulty of the FOLaw proved to be the modelling of the "world knowledge". The knowledge gained from FOLaw was used in the project E-Court and in the development of a core legal ontology, LRI-Core. Within the framework of this project, a flexible, multilingual information retrieval system using heterogeneous sources (audio, video, text) has been developed in the field of criminal procedure. The LRI-Core also finds experimental use in the projects E-Power (Van Engers et al., 2001) and DIRECT (Breuker and Hoekstra, 2004).

The main task of the EU-funded e-Content project LOIS (Lexical Ontologies for legal Information Sharing) was building a multi-lingual legal WordNet with concepts in six European languages for the purpose of facilitating legal information retrieval. Thus, the LOIS project focus was limited to one piece of the "cake of problems", the thesaurus problem. Up-to-date thesaurus and lexical ontologies research was used to develop a cross-lingual ontology with 5000 thesaurus entries in 6 languages in order to improve legal information retrieval (Dini et al., 2005).

In the very end, legal information systems should develop into dynamic electronic commentaries (Schweighofer, 2006) summarizing, structuring and indexing relevant legal information as required by users. Standard text books comply with this aim but are not sufficiently dynamic. Quite often, they are only updated every few years. The same methodology as described in the next section may be used for developing such electronic commentaries but for the time being ontologies and text analysis methods are not sufficiently developed for an implementation in practice.

## 4. Supplementing Boolean search: Query Expansion and Relevance Feedback

Our model should not replace but supplement current legal information retrieval systems. As the quality of the query is the main problem query improvement is the first logical step for improving retrieval performance.

Two methods have been developed and tested so far: query expansion using ontologies, and using relevance feedback.

### 4.1. Query Expansion using Ontologies

Improving the user's query with additional terms is called query expansion. For quite some time, query expansion has been seen as an effective way to improve retrieval performance (Salton and McGill, 1986). New words and phrases are added to the existing search term(s) to generate an expanded query.

In the LOIS project, some sort of query expansion was used for searching with appropriate terms in other jurisdictions. Our approach is similar but more focused on the terminology of the same legal jurisdiction. A lexical ontology was built for providing the knowledge base containing about 5500 terms, definitions and relations between concepts. Most of the terms were reused from the LOIS database; the extensions concern mostly competition law, European law and international law. It has to be noted that 3 types of relevant lexical information are stored in the database: terms, definitions and relations that could be weighted differently. The ILI concept of LOIS was also reused.

The one or two (or more) words provided in a query are searched in the knowledge base and weighted: The easy case concerns the search for a synonym. If the term exists and a synonym relation is established, a weight of 1 is given. More difficult is the case if several subterms exist. These terms are given a weight of 0.5. All meaningful terms in a definition are selected and given a weight of 0.25. All these assigned weights for terms are added. It would be fine if these weights could be reused but Boolean retrieval does not allow that. So weights greater than 1 are reduced to 1, weights greater than 0.5 are enlarged to 1 and the rest is simply not taken into account. No linguistic pre-processing besides automatic use of truncation exists at the moment.

Example: Knowledge base entry for term "animal welfare"
Animal welfare:
ILI: Tierschutz=Tierwohlfahrt (DE), le bien-être des animaux (FR), el bienestar de los animales (ES) etc.
Sub-terms: Artenschutz (DE), Tierhaltung (DE), Tiertransporte (DE), Schlachtung (DE), Tierversuche (DE)
Definition: payments for additional costs and income foregone for treatment of animals beyond the relevant mandatory standards established pursuant to Art. 4 of and Annex III of Regulation 1782/2003 (Directives

91/629, 91/630 and 98/58) and other mandatory requirements

### 4.2. Relevance Feedback: Using Search Context Information

In classic relevance feedback, relevance information is collected from the documents retrieved using an initial query, in order to form a second query. We think, however, that relevance feedback potential lies within the search context of the different users.

Legal information systems store - for billing purposes - accumulated information on user interactions consisting of query, results and downloaded documents. As a start, we - in our system - only consider the quantitative most important queries and documents. Even quite irrelevant terms are taken into account in order to support those with some "erroneous imagination" (e.g. the term subvention takes into account also Community support that is technically not State aid).

In the near future, this approach of relevance feedback will be tested in a sub-domain of Austrian law, tax law.

## 5. Prototype

The prototype consists of a database of about 1770 Commission decision on State aid in the agriculture sector covering the period of 2000 to 2006 but also the relevant guidelines and case law. 22 Community languages should be covered, however, still with strong focus on English, French, German and Spanish. This text corpus simulates an index covering all relevant sources on State aid (websites EUR-Lex, Directorates-General Competition, Agriculture and Secretariat-General). It may be noted that users get easily frustrated by the complex structures of publication (e.g. in EUR-Lex, the term "animal welfare" produces 1497 hits but relevant information can only be found if the user knows that a restriction to "Other Documents" leading to 299 documents; only if the user is aware that the Guidelines for State aid in the agriculture sector have been recently published and Commission decisions are summarised under "Summary information communicated by Member States Ĕ" then a more detailed analysis of results can be done).

This text corpus is stored in an information retrieval system (we are using askSam and the Open Source free text standalone enterprise search server Solr). The core of value-added constitutes the knowledge base containing a lexical ontology (similar to that developed in the LOIS project, stored in askSam and XML) and some statistical tools.

Quite valuable support for improving the lexical ontology provided also the GATE tools for linguistic analysis (www.gate.ac.uk). In addition to that, programs developed within the LOIS and KONTERM projects are reused if possible (e.g. term clustering using context, document classification, clustering and labelling of documents etc. (Schweighofer, 1999).

Solr is based on the Lucene Java search library providing also indexing XML documents. The Lucene Query Language is sufficiently powerful and flexible to offer standard legal search options but also query expansion ranking functions.

The overall objective of our prototype is to show that the search result quality of legal information systems can be significantly improved by using artificial intelligence and natural language processing techniques, in a first step in particular by query expansion.

## 6. Experimental test results

First tests have been done in the domain of State aid law using a highly sophisticated lexical ontology. Evaluation results are still tentative and mostly based on the so-called Delphi method (Linstone and Turoff, 1975). The first tests concerned the improvement of retrieval results using query expansion with synonyms in the other Community languages. The results were - not really surprising - quite good. If the knowledge base has sufficient coverage and quality it remains the best way of finding and summarising documents in other languages than the query language. Using sub-terms, umbrella terms and definition terms delivers a much higher number of relevant results, thus more information hints - but results have to be properly presented.

A typical example of our test series: A Czech farmer is displeased by high subsidies given to German farmers doing animal welfare. In particular, he does not understand why 150 euros are paid every year for each cow that has a bigger stable, can get out in free air as it wants and is offered free access to drinking water. He is considering a State aid complaint. The Czech query is extended using the ILI synsets of the other 22 Community languages (e.g. animal welfare, Tierschutz, le bien-être des animaux, el bienestar animal) but also synonyms, umbrella terms, sub-terms and definition terms and weighted accordingly (see example above). This quite complex search is done on the test information retrieval system (in practice it would be accomplished using the indexes of the various databases and websites). Relevant documents are grouped according to main term and Member State and then presented according to document type and chronological order.

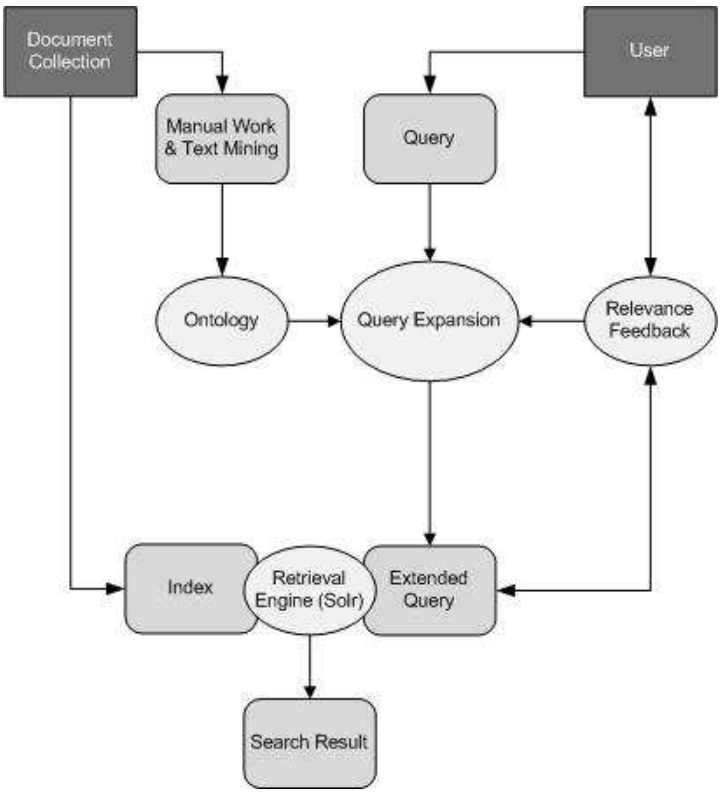



*Figure 1.* Sketch of Prototype IR system with query extension and relevance feedback

For improvement, some models of better presentation and visualisation are currently under examination in order to address the problem of the lower precision of the search results. The clustering of documents allows an easy browsing through Commission decisions and Member States concerning animal welfare leaving beside relevant legislation (Community guidelines for State aid in agriculture and Regulation 1698/2005). Thus, it is quite easy to find the document that is really relevant: the State aid approval of the German notification of State aid for "Gemeinschaftsaufgabe für Agrarstrukturen und Küstenschutz (Common Task for Agrarian Structures and Costal Protection)".

The main improvement consists in the much broader coverage of documents found and thus a broader scope of hints of useful information (e.g. aspects of standards, additional costs and income foregone in all relevant topics, e.g. also in agri-environment). For a practical implementation, the presentation and filtering of results seems to be decisive that has to be a strong focus in future research. Some experimental checks revealed that users may not be able to find a proper search term for the knowledge base as common language may use a different term. Here, integration of terms from relevance feedback research may help but no results of experiments can be reported so far.

## 7. Conclusions and future work

Our research is still quite at the beginning. A sound methodology and a first prototype are now available that are presented in the paper. At the moment, database and knowledge base are focused on the domain of State aid in agriculture. In the future, this application should be enlarged, covering the whole of EU competition law and also the general part of EU law. A text corpus exists also for international law but has to be enlarged substantially. The relevance feedback methodology will be tested in Austrian tax law. At the moment it is still too early to address questions of scaling-up as further test results are still pending. However, it does not seem insurmountable to achieve the required number of entries of a lexical database (also including ILI entries). The success of this approach depends on the quality of the knowledge base and the ability of the knowledge team to build and constantly update the lexical ontology. A (semi)automatic approach seems to be required and, therefore, tests on that will also be part of our future research.

## References

Bench-Capon, T. J. M. and Visser, P. R. S. Ontologies in legal information systems; the need for explicit specifications of domain conceptualisations. In *ICAIL '97: Proceedings of the 6th international conference on Artificial intelligence and law*, ACM Press, 1997.

Bing, J., editor. *Handbook of Legal Information Retrieval.* Elsevier Science Inc., 1984.

Bing, J. Designing text retrieval systems for conceptual searching. In *ICAIL '87: Proceedings of the 1st international conference on Artificial intelligence and law.*, ACM Press, 1987.

Bing, J. Legal Text Retrieval and Information Services. In *25 Years Anniversary Anthology In Computers and Law*, TANO, 1995.

Bing, J. Text Retrieval and Hypertext: The deep Structure (opening and invited talk). In *DEXA '98 Ũ Data Bases and Expert System Applications*, Wien, 1998.

Breuker, J. and Hoekstra, R. Direct: Ontology based discovery of responsibility and causality in legal case descriptions. In *Legal Knowledge and Information Systems. Jurix 2004: The Seventeenth Annual Conference*, IOS Press, 2004.

Dini, L. and Peters, W. and Liebwald, D. and Schweighofer, E. and Mommers, L. and Voermans, W. Cross-lingual legal information retrieval using a WordNet architecture. In *ICAIL '05: Proceedings of the 10th international conference on Artificial intelligence and law*, ACM Press, 2005.

Frakes, W. B. and Baeza-Yates, R., editors *Information Retrieval: Data Structures and Algorithms*. Prentice Hall PTR, 1992.

Gruber, T. R. Ontolingua: A mechanism to support portable ontologies. Stanford University, Knowledge Systems Laboratory, Technical Report KSL-91-66, 1992.

Gruber, T. R. A translation approach to portable ontology specifications. *Knowledge Acquisition*, 5(2):199–220, 1993.

Hafner, C. D. *An Information Retrieval System Based on a Computer Model of Legal Knowledge*. UMI Research Press, 1981.

Hirst, G. Ontology and the Lexicon. In Steffen Staab and Rudi Studer, editors, *Handbook on Ontologies in Information Systems*, Springer, 2003.

Harvold, T. and Bing, J. *Legal decisions and information systems (Publications of the Norwegian Research Center for Computers and Law)*. Universitetsforlaget, 1977.

Linstone, H. and Turoff, M., editors *The Delphi Method: Techniques and Applications)*. Addison Wesley, 1975.

McCarty, L. T. A language for legal Discourse I. basic features. In *ICAIL '89: Proceedings of the 2nd international conference on Artificial intelligence and law*, ACM Press, 1989.

Matthijssen, L. *Interfacing Between Lawyers and Computers:An Architecture for Knowledge-Based Interfaces to Legal Databases (Law and Electronic Commerce)*. Kluwer Law International, 1999.

Moens, M. F. Innovative techniques for legal text retrieval. *Artificial Intelligence and Law*, 9(1):29–57, 2001.

Rose, D. E. *A Symbolic and Connectionist Approach To Legal Information Retrieval*. LEA Inc., 1994.

Salton, G. and McGill, M. J. *Introduction to Modern Information Retrieval*. McGraw-Hill, 1986.

Schweighofer, E. *Legal Knowledge Representation:Automatic Text Analysis in Public International and European Law (Law and Electronic Commerce)*. Kluwer Law International, 1999.

Schweighofer, E. *Wissensrepräsentation in Information Retrieval-Systemen am Beispiel des EU-Rechts (Dissertationen der Universität Wien)*. WUV, 2000.

Schweighofer, E. Computing Law: From Legal Information Systems to Dynamic Legal Electronic Commentaries. In C. M. Sjöberg and P. Wahlgren, editors *Festskrift till Peter Seipel*, Norsteds Juridik AB, 2006.

Stranieri, A. and Zeleznikow, J. *Knowledge Discovery from Legal Databases (Law and Philosophy Library)*. Springer, 2005.

Stamper, R. K. The Role of Semantics in Legal Expert Systems and Legal Reasoning. *Ratio Juris*, 4(2):219–244, 1991.

Turtle, H. Text retrieval in the legal world. *Artificial Intelligence and Law*, 3(1):5–54, 1995.

Valente, A. *Legal Knowledge Engineering, A Modelling Approach*. IOS Press, 1995.

Van Engers, T. and Gerrits, R. and Boekenoogen, M. and Glassee, E. and Kordelaar, P. POWER: using UML/OCL for modeling legislation - an application report. In *ICAIL '01: Proceedings of the 8th international conference on Artificial intelligence and law*, ACM Press, 2001.

Van Kralingen, R. W. *Frame-Based Conceptual Models of Statute Law.* Kluwer Law Intl, 1995.

Visser, P. R. S. *Knowledge Specification for Multiple Legal Tasks: A Case Study of the Interaction Problem in the Legal Domain.* Kluwer Law Intl, 1995.

Winkels, R. and Boer, A. and Hoekstra, R. CLIME: Lessons Learned in Legal Information Serving. In *ECAI 2002: Proceedings of the 15th Eureopean Conference on Artificial Intelligence*, IOS Press, 2002.

Zeleznikow, John and Hunter, Dan *Building Intelligent Legal Information Systems (Computer Law, No 13).* Springer, 1994.

## Author Index