IDT Series

LOAIT 2009

3rd Workshop on Legal Ontologies and Artificial Intelligence Techniques joint with 2nd Workshop on Semantic Processing of Legal Text

Editors:
Núria Casellas ı Enrico Francesconi ı Rinke Hoekstra ı Simonetta Montemagni

Barcelona, Spain
June 8, 2009

IDT Series

Series Editors:

Pompeu Casanovas
Núria Casellas
Pablo Noriega
Marta Poblet
Antoni Roig
Joan-Josep Vallbé

Volume 2

LOAIT 2009
Workshop on Legal Ontologies and Artificial Intelligence Techniques, joint with 2nd Workshop on Semantic Processing of Legal Texts. In conjunction with ICAIL 2009, 12th International Conference on Artificial Intelligence and Law, 8-12 June, Barcelona.

Volume Editors: Núria Casellas, Enrico Francesconi, Rinke Hoekstra, Simonetta Montemagni

Acknowledgements: ARCS. 2008-ARCS200028; TSI-020501-2008-131; TSI-020100-2008-134. CSO-2008-05536-SOCI; CEN2008-1018.

ISSN: 2013-5017
Legal Deposit: B-26200-2009

Also appeared online with CEUR Workshop Proceedings (CEUR-WS.org, ISSN 1613-0073)

Editor's addresses:

Institute of Law and Technology
School of Law
Universitat Autònoma de Barcelona
08193 Bellaterra, Spain
nuria.casellas@uab.cat

Istituto di Teoria e Tecniche dell'Informazione Giuridica (ITTIG-CNR)
Via dei Barucci 20, 50127 Firenze, Italy
francesconi@ittig.cnr.it

Leibniz Center for Law
Faculteit der Rechtsgeleerdheid/Faculty of Law
Universiteit van Amsterdam
Postbus 1030
1000 BA, Amsterdam, The Netherlands
hoekstra@uva.nl

Istituto di Linguistica Computazionale "Antonio Zampolli", ILC-CNR
Area della Ricerca del CNR
Via Giuseppe Moruzzi Nº 1
56124 Pisa, Italy
simonetta.montemagni@ilc.cnr.it

Publisher:

Huygens Editorial
La Costa 44-46, át.1º
08023 Barcelona. Spain.
www.huygens.es

Sponsoring Institutions

Program Chairs

Núria Casellas, Institute of Law and Technology (IDT-UAB), Universitat Autònoma de Barcelona, Spain
Enrico Francesconi, Institute of Legal Information Theory and Techniques (ITTIG-CNR) Florence, Italy
Rinke Hoekstra, Leibniz Center for Law, University of Amsterdam, The Netherlands
Simonetta Montemagni, Institute of Computational Linguistics (ILC-CNR), PISA, Italy

Program Committee

Trevor J.M. Bench-Capon, University of Liverpool, UK
V. Richard Benjamins, Telefónica R&D, Spain
Guido Boella, University of Turin, Italy
Alexander Boer, Leibniz Center for Law, University of Amsterdam, The Netherlands
Joost Breuker, Leibniz Center for Law, University of Amsterdam, The Netherlands
Thomas Bruce, Cornell Law School, USA
Paul Buitelaar, DERI research institute in Galway, Ireland
Pompeu Casanovas, Institute of Law and Technology, UAB, Spain
Aldo Gangemi, Institute of Cognitive Sciences and Technologies (ISTC-CNR), Italy
Roberto García, Universitat de Lleida, Spain
Mustafa Jarrar, University of Cyprus, Cyprus
Michael Klein, VU University Amsterdam, The Netherlands
Alessandro Lenci, Department of Linguistics, University of Pisa, Italy
Wim Peters, NLP Research Group, University of Sheffield, UK
Giovanni Sartor, European University Institute, Florence, Italy
Marco Schorlemmer, IIIA-CSIC, Spain
Erich Schweighofer, University of Vienna, Austria
Barry Smith, University at Buffalo, USA
York Sure, SAP Research, Germany
Daniela Tiscornia, Institute of Legal Information Theory and Techniques (ITTIG-CNR), Italy
Tom van Engers, Leibniz Center for Law, University of Amsterdam, The Netherlands
Réka Vas, Department of Information Systems, University Corvinus of Budapest, Hungary
Radboud Winkels, Leibniz Center for Law, University of Amsterdam, The Netherlands

Foreword

In this third edition of LOAIT joint with the second edition of the Workshop on Semantic Processing of Legal Texts, we focus our attention on two main research areas: Legal Knowledge Representation as a top-down approach, and Ontology Learning from Legal Texts as a bottom-up approach on legal ontologies.

This year 11 papers have been accepted (7 full and 4 short papers) coming from Italy, the Netherlands, Spain, Austria, United Kingdom, and Hungary.

Besides the submissions, this year the Workshop will include two invited lectures coping with the two main areas of interest of the event. The first one will be given by Joost Breuker, whose work on legal ontologies represent an essential reference for the legal informatics community; he will open the workshop with an overview of the main research results in the legal ontology domain as well as dreams and new trends on this topic. The second one will be given by Hugo Zaragoza from Yahoo! Research, on the current methods aimed to exploit linguistic annotations in search.

June 2009

Núria Casellas
Enrico Francesconi
Rinke Hoekstra
Simonetta Montemagni

Table of Contents

Legal Assessment Using Conjunctive Queries

András Förhécz[1] and György Strausz[1]

Budapest University of Technology and Economics
Department of Measurement and Information Systems
Budapest, Hungary

Abstract. Using the Web Ontology Language (OWL) for knowledge representation in the legal domain is very promising but has some limitations. The language is complex thus hard to comprehend, still decidability results in a limited expressiveness which may introduce serious problems in modelling.
An aspect of limited expressiveness is the tree-model property of OWL, which can be overcome using rule formalisms or introducing variables, however losing decidability is not always acceptable.
We propose using conjunctive queries when modelling conditions of legal norms in the HARNESS architecture. Inference services required for modelling legislation and building legal assessment applications can be feasible using grounded queries resulting in a decidable formalism. Since most of the knowledge base remains within the limits of pure OWL2, we can benefit from consistency checking services of OWL2.

1 Introduction

Using the Web Ontology Language (OWL) for knowledge representation in the legal domain is very promising as different types of inference services are provided on top of a relatively expressive formalism, the description logic underlying OWL2.

The main benefits are strict semantics, consistency checking and feasible reasoning. In contrast to this, OWL is very complex thus hard to comprehend, resulting in a knowledge acquisition bottleneck. Remaining decidable also costs limited expressiveness which may introduce serious problems in modelling.

An aspect of limited expressiveness can be described with the tree-model property of OWL: only tree-like axioms are allowed (except for nominals, transitive properties and role inclusion axioms in OWL2). Complex structures cannot be described precisely due to the lack of cycles in axioms or predicates with arbitrary number of arguments: only unary (classes) and binary (properties) predicates are allowed.

Representing diamond-shaped structures is a frequently reoccuring problem. Suppose a sales contract with two actors – seller and customer – where the subject of the transaction should be joined to both the seller and the customer. Users familiar with rule formalisms tend to use rules or other extensions supporting variables to overcome these limitations, although, this way decidable satisfiability checking w.r.t. the T-Box is lost. In certain cases it is possible to represent cyclic structures using knowledge patterns and OWL2, as described in [1], nevertheless this solution cannot enforce `owl:sameAs` relations, only a custom property defined as a replacement.

An interesting approach for describing complex structures in OWL is the representation used in the HermiT reasoner [2]. Here the representation formalism is extended

with description graphs for finite complex structures, where nodes and edges of graph-like structures are labelled with classes and properties respectively. Reasoning remains decidable but using arbitrary OWL axioms for the entities of complex structures is not allowed.

In this article we propose an alternative solution using conjunctive queries to solve legal assessment problems in the HARNESS[1] system. The next section introduces HARNESS, a legal knowledge-based system aimed at solving legal assessment problems. In the following two sections conjunctive queries are introduced and available inference services are described, including the case, when they are mixed with class expressions. In section 4.1. we show how to use conjunctive queries in HARNESS. The last section provides an overview of possible extensions and future plans.

2 Introducing HARNESS

A central task in legal knowledge-based systems is *legal assessment*: deciding whether some case is allowed or disallowed in a certain legal environment. In everyday situations a legal expert can help individuals to answer this sort of question, but due to the increasing size and complexity of legislation this process becomes more and more difficult, although transparency of jurisdictions would demand the opposite.

During the ESTRELLA[2] project an open platform was developed for legal knowledge technologies, including the Legal Knowledge Interchange Format (LKIF), a reference open source legal CMS called eXistrella, an argumentation engine Carneades and a DL-based inference system called HARNESS. The architecture of HARNESS enables solving different tasks including drafting or legal planning, although it is currently aimed solely at legal assessment.

HARNESS greatly exploits current Semantic Web technology by relying on formal ontologies and highly optimized DL reasoners. Legal assessment requires three distinct types of knowledge: a domain ontology, normative knowledge and case descriptions. The domain ontology defines the concepts and constraints in the field of interest and provides building blocks for defining individual cases. This ontology is a specialization of the LKIF Core ontology of basic legal concepts [3]. Normative knowledge describes regulations which govern the situations in question. Case descriptions underpin individual situations to be evaluated by HARNESS.

Each norm is expressed as a generic situation in which a state of action is qualified as undesirable, permitted or prescribed [4]. The situation itself is a conjunction of conditions, naturally expressed as a class expression in OWL, as specified in [5].

When modelling law in the HARNESS architecture using OWL the tree-model property of OWL hinders expressing complex situations. We will present an extension that will solve some of these issues: using *conjunctive queries* for specifying generic cases.

3 Using conjunctive queries

Conjunctive queries (CQ) are well known in database systems and have been standing in the focus of DL research for years now but not yet widely available in OWL appli-

[1] Hybrid Architecture for Reasoning with Norms Exploiting Semantic web Services

[2] European project for Standardized Transparent Representations in order to Extend Legal Accessibility, IST-2004-027655, see `http://www.estrellaproject.org/`.

cations. Practical results for complex DL languages have only appeared recently [6]. A possible syntax for such queries has just been defined in SPARQL-DL [7].

A conjunctive query is a conjunction of concept expressions of the form $C(t)$ and role expressions of the form $r(t, t')$ where C is a concept, r is a role and t, t' are terms, i.e. variables or individual names [6]. All variables are existentially qualified. These conditions are very similar to the body (condition) of SWRL rules. Introducing variables when specifying generic cases basically solves three different issues:

- We are no longer limited by the tree-model property of OWL, generic cases can express arbitrary relational structures.
- In a query, the values of variables can point at the case or part of the case the norm refers to. We can keep track of individuals when identifying obligations, permissions and violations.
- Using variables enlightens modelling. Most knowledge experts are familiar with variables and it is easier for them to specify the condition with conjunctive queries. When additional expressiveness is not required, the CQ can be automatically transformed into an OWL class expression [8].

The following example demonstrates the usage of CQs. In Section 16. paragraph (2) in the Hungarian Law on Duties[3] an exemption is specified for paying duties on a land received as a gift:

> "In order to verify completion of the construction of the residential house [...] the state tax authority shall contact the competent building authority [...] If the building authority provides a certificate in proof of the occupancy permit issued to the name of the property owner, the state tax authority shall cancel the duty assessed, but suspended in respect of payment."

The condition of the exemption can be formalized using conjunctive queries the following way. If a gift ($?g$) is a *plot of land*, and a building ($?b$) has been built on it, which is a *residential house*, and an *occupancy permit* ($?p$) was issued to the name of the *donee* ($?d$), the generic case is fulfilled:

$$\begin{aligned} GC_{S16\_2} \equiv\ & Donation(?t) \wedge donee(?t, ?d) \\ & \wedge subject(?t, ?g) \wedge PlotOfLand(?g) \\ & \wedge built\_onto(?b, ?g) \wedge ResidentialHouse(?b) \\ & \wedge permit\_issued(?b, ?p) \wedge OccupancyPermit(?p) \\ & \wedge issued\_to(?p, ?d) \end{aligned} \tag{1}$$

A conjunctive query can be represented by a graph where each variable in the query give rise to a node in the graph. Concept names appear as node labels, role names as edges at the appropriate variables in the graph. Figure 1 represents the graph for the CQ shown above.

[3] Hungarian Law on Duties, Act XCIII of 1990, only available in Hungarian, `http://net.jogtar.hu/jr/gen/hjegy_doc.cgi?docid=99000093.TV#pr123`

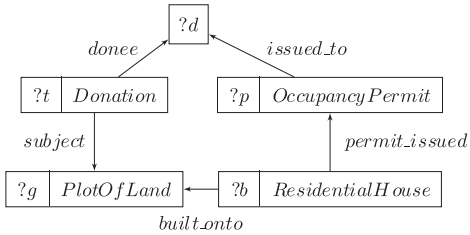


Fig. 1. Graph representation of a conjunctive query

4 Inferences and results

Just like OWL classes, different inference services can be implemented to conjunctive queries:

query entailment is the decision problem to answer whether a query is true in all models of a knowledge base.

query answering is the problem of finding all answer tuples for a query. If the entailment is false, there will be no answers. Otherwise, there may be one or more tuples fulfilling the constraints described in a query.

satisfiability is to decide if a knowledge base has at least one model in which the query is true. When building a model and respective queries, a non-satisfiable query may indicate inconsistency in the model, as the corresponding legal condition will never be fulfilled.

subsumption of CQs can be defined in a similar manner as class subsumption: with respect to a knowledge base $\mathcal{K}$ a query Q_1 subsumes the query Q_2 if in all models where Q_2 is true, Q_1 is also true. A more specific condition may correspond to a norm with higher priority based on *lex specialis*.

Not all of these problems are solved for the description logic underlying OWL2. Satisfiability can be easily answered using a DL reasoner, but it is still an open issue whether the other problems are decidable for the DL $\mathcal{SHOIN}(\mathcal{D})$. Latest results showed that query entailment is decidable for $\mathcal{SHIQ}$ [9] and $\mathcal{SHOQ}$ [10] which are slightly restricted sublanguages of OWL2.

However in the general interpretation variables in a CQ are not required to correspond to a named individual in the ABox. For so-called *non-distinguished variables* only the existence of a suitable element is required in the model, and *answer variables* are required to have a corresponding named individual. This is important as in the restricted closed-world interpretation of CQs we only use answer variables, and then all inference problems are decidable in OWL2.

A subsumption hierarchy of CQs can be derived the same way as for OWL classes. Conjunctive queries are a generalization of OWL class expressions, as all class expression can be trivially transformed to an atomic CQ with one variable:

$$C \rightarrow Q(x) \equiv C(x)$$

As a result subsumption can be defined across CQs and OWL named classes, and hierarchy of CQs and classes can be merged. As an example for the CQ in equation 1: $GC_{S16\_2} \sqsubseteq Donation$.

Satisfiability of conjunctive queries can be derived from subsumption the same way as for OWL classes, by defining the always unsatisfiable CQ:

$$Q_{\perp}(x) \equiv \perp(x)$$

$$Q(\dots) \text{ is unsatisfiable} \Leftrightarrow Q(\dots) \sqsubseteq Q_{\perp}(x)$$

Subsumption relations across CQs and OWL classes are important for designing tool support and knowledge engineering methodology. Conjunctive queries are not yet natively supported by major ontology editors but can be integrated into e.g. class hierarchy in a relatively straightforward manner giving confidence to knowledge engineers.

4.1 Application in HARNESS

In legal inference we have two distinct modelling issues: creating a domain ontology for enforcing valid case descriptions and formalizing normative knowledge in legal assessment. In the assessment part we are using the HARNESS architecture [5] providing definitions for generic cases (GC). A generic case represents the situation describing the condition part of a norm.

The domain ontology must be consistent and all kind of inferences (including full consistency) are required, OWL is an adequate formalism. With GC descriptions, however, the only inferences required are hierarchy of GCs (is one description more general than another?) and case entailment (does a case fulfill all conditions of a GC?). The problem of modelling complex structures generally occur with GCs, so we will propose to use conjunctive queries in specifying generic cases.

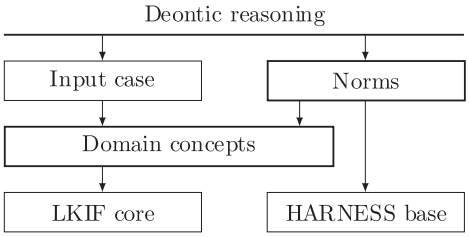


Fig. 2. Interactions of ontology modules in HARNESS

We can use conjunctive queries to describe generic cases while the rest of the model is still specified in OWL2. In Figure 2 parts of a general HARNESS knowledge base are shown. *LKIF-core* and *HARNESS base* are standard ontology modules: providing basic legal concepts, the base class for norms and deontic operators. *Domain concepts* include terminological knowledge for specifying input case descriptions. The normative knowledge is provided in part *Norms*, allowing conjunctive queries for modelling conditions. Consistency of an input case specification can be verified using conventional DL tools and only evaluating the deontic reasoner requires conjunctive query evaluation.

Conjunctive queries are appropriate because inference services are available to cover the tasks required in HARNESS:

– *query answering* matches input case descriptions with relevant norms,

– *subsumption relations* provide exceptions (lex specialis) for norms in the model and
– *satisfiability* ensures that each norm is consistent with the domain model.

An experimental implementation for the closed-world interpretation (using only answer variables) has been provided. The reasoner can use any DL reasoner (black-box reasoning) or Pellet[4] and its highly optimized algorithms. The CQ reasoner can be accessed from Protégé 4[5] as a plug-in and can be selected instead other DL reasoners.

When using HARNESS, legal knowledge bases with conjunctive queries are supported. The generic situations in the normative part of the knowledge base can be expressed both using OWL class expressions or conjunctive queries. Generic situations can be reviewed in a combined hierarchy following *lex specialis* relations and also showing all cases satisfying the condition.

5 Future plans

An important feature of legal expert systems is the ability to provide reliable and comprehensible explanations for inference results. For OWL this can be achieved using a recent feature of the Pellet reasoner: laconic justifications [11, 12]. These are minimal set of axioms supporting a single conclusion, extracting the piece of information required for understanding a single issue. As conjunctive queries are handled by an additional reasoning mechanism, explanation services should be extended to support justifications in these legal knowledge bases.

Unfortunately description logics are hard to comprehend for the casual users, so when non-expert users should be able to interpret explanations, OWL axioms have to be translated to natural language or a graphical representation. The former can be achieved with NLP tools like ROO Rabbit [13] or Ace View [14]. We already took steps on adopting these services to handle queries and translate them to our target language, Hungarian.

References

1. Hoekstra, R., Breuker, J.: Polishing diamonds in OWL 2. In: EKAW. (2008) 64–73
2. Motik, B., Grau, B.C., Sattler, U.: Structured objects in OWL: Representation and reasoning. In: WWW. (2008) 555–564
3. Hoekstra, R., Breuker, J., Bello, M.D., Boer, A.: The LKIF core ontology of basic legal concepts. In: LOAIT. (2007) 43–63
4. Valente, A.: Legal knowledge engineering: A modelling approach. PhD thesis, University of Amsterdam (1995)
5. van de Ven, S., Hoekstra, R., Breuker, J., Wortel, L., El-Ali, A.: Judging amy: Automated legal assessment using OWL 2. In: OWLED. (2008)
6. Glimm, B.: Querying Description Logic Knowledge Bases. PhD thesis, The University of Manchester, Manchester, United Kingdom (2007)
7. Sirin, E., Parsia, B.: SPARQL-DL: SPARQL query for OWL-DL. In: 3rd OWL Experiences and Directions Workshop (OWLED-2007). (2007)

[4] Pellet: an open source OWL 2 DL reasoner developed by Clark & Parsia, LLC `http://clarkparsia.com/pellet/`

[5] Protégé 4 ontology editor, `http://protege.stanford.edu/`

8. Gasse, F., Haarslev, V.: DLRule: A rule editor plug-in for Protege. In: OWL: Experiences and Directions (OWLED). (2008)
9. Glimm, B., Horrocks, I., Lutz, C., Sattler, U.: Conjunctive query answering in the description logic $\mathcal{SHIQ}$. In: Proceedings of the 20th International Joint Conference on Artificial Intelligence (IJCAI 2007). (2007)
10. Glimm, B., Horrocks, I., Sattler, U.: Conjunctive query entailment for shoq. In: Description Logics. (2007)
11. Horridge, M., Parsia, B., Sattler, U.: Laconic and precise justifications in OWL. In: International Semantic Web Conference. (2008) 323–338
12. Kalyanpur, A., Parsia, B., Horridge, M., Sirin, E.: Finding all justifications of OWL DL entailments. In: ISWC/ASWC. (2007) 267–280
13. Dimitrova, V., Denaux, R., Hart, G., Dolbear, C., Holt, I., Cohn, A.G.: Involving domain experts in authoring OWL ontologies. In Sheth, A.P., Staab, S., Dean, M., Paolucci, M., Maynard, D., Finin, T.W., Thirunarayan, K., eds.: International Semantic Web Conference. Volume 5318 of Lecture Notes in Computer Science., Springer (2008) 1–16
14. Kaljurand, K.: ACE View — an ontology and rule editor based on Attempto Controlled English. In: OWLED. (2008)

Legal Taxonomy Syllabus version 2.0

Gianmaria Ajani[1], Guido Boella[2], Leonardo Lesmo[2], Marco Martin[2], Alessandro Mazzei[2], Daniele P. Radicioni[2], and Piercarlo Rossi[3]

[1] Dipartimento di Scienze Giuridiche - Università di Torino
gianmaria.ajani@unito.it
[2] Dipartimento di Informatica - Università di Torino
guido,lesmo,mazzei,radicion@di.unito.it, notmart@gmail.com
[3] Dipartimento di Studi per l'Impresa e il Territorio - Università del Piemonte Orientale
piercarlo.rossi@eco.unipmn.it

Abstract. The need for managing the conceptual representation of European law led to the development of the Legal Taxonomy Syllabus (LTS) and the related methodology. In this paper we consider further legal issues that emerged during the test and use phases, and outline the new features that we added to the new version, the LTS 2.0.

1 Introduction

European Union Directives (EUDs) are sets of norms that have to be implemented by the national legislations and translated into the language of each Member State. The general problem of multilingualism in European legislation has recently been tackled by linguistic and ontological tools [1, 2, 3, 4]. The management of EUD is particularly complex, since the *implementation* of a EUD does not correspond to a straight transposition into a national law.

In previous work we carried out the Legal Taxonomy Syllabus[4] (LTS), a tool to build multilingual conceptual dictionaries aimed at representing and analysing terminologies and concepts from EUDs [5, 6]. LTS is based on the distinction between *terms* and *concepts*. The latter ones are arranged into ontologies that are organised in levels. Only two levels were defined: the European level –containing only one ontology deriving from EUDs annotations–, and the national level –hosting the distinct ontologies deriving from the legislations of EU member states.

While annotating the EUDs, testing and using the system, some more requirements emerged from users expert in law, demanding for a more sophisticated approach along with further developmental efforts: first, it is frequent the case of concepts which are the result of a doctrinal interpretation process rather than of the definition in directives. If, on the one hand, the definitions in directives and their relation with the actual text are required by legal scholars to have a precise model of European law, the layman is more interested in the concepts which results from the doctrinal interpretation. Furthermore, laws are typical objects evolving through time. An open issue to cope with in building legal frameworks both at the European and at the national level is the *normative change* [7, 8]. Concepts in the legal ontologies should not only represent the consolidated legal text, but should also keep trace of the evolution of meaning.

[4] http://www.eulawtaxonomy.org

In this paper we consider not only the terms defined in the directives, but also the interpretation process of legal scholars in the LTSand how to better integrate concepts and the text of EUDs in the LTS. We answer the first question by introducing *abstract* concepts (*abstract* in that they are not related to a single directive), which should be conveniently recognized as a *grouping* of concepts. The users will be thereby allowed to navigate the ontology at different levels of details depending on their goals. Moreover, exploiting natural language processing techniques we greatly simplify the management of legal text associated to concepts. Also, we investigate how to extend the ontology with a temporal dimension to the ends of representing *normative change*, and to allow users to search also for past meanings of terms and the modified norms introducing them. To these ends, we introduce *time* into the ontology, and allow new concepts to replace the old ones while keeping the latter ones in the system as well.

2 Multilingual and Multilevel Ontologies for European Directives

Comparative Law has identified two key points in dealing with EUD, which make more difficult dealing with the polysemy of legal terms. We call them the *terminological* and *conceptual misalignments*. The first problem is determined by the lexical ambiguity of the legal terms (in particular homonymy) in the translation of EUDs. The second problem is determined by the lexical and conceptual ambiguity of the legal terms (in particular polysemy) in the implementation of EUDs. These issues determined the development of the first release of the LTS, and have been illustrated in [6].

We now illustrate further issues in handling EUDs that required to devise further features to enrich the original LTS.

2.1 Concepts Abstraction

The LTS system relies on the concept of *unitary-meaning* or *umeaning*: such atomic concepts can be derived from excerpts of the text of legal norms, such as European directives or national laws, and are arranged into two separate categories of umeanings, as described in [6]. EUDs provide rigorous definitions of some terms, such as the definition of the Italian term *consumatore* (*consumer*), in the Italian version of the *EUD 93/13/EEC*, Art. 2 is:

> [. . .](b) "consumatore": qualsiasi persona fisica che, nei contratti oggetto della presente direttiva, agisce per fini che non rientrano nel quadro della sua attività professionale; [...]
>
> [. . .](b) "consumer": means any natural person who, in contracts covered by this Directive, is acting for purposes which are outside his professional activity; [. . .] (*our literal translation*)

However, two facts must be pointed out. Different EUDs might affect different aspects of the legislation: thus the definition of a term in a EUD only applies to a specific context. Furthermore, EUDs could be written at different points in time, and they can introduce diverging definitions. Let us consider the definition of *consumatore*, as it appears in the Italian version of the *EUD 2002/65/EC, Art. 1*:

> [...](d) "consumatore": qualunque persona fisica che, nei contratti a distanza, agisca per fini che non rientrano nel quadro della propria attività commerciale o professionale; [...]
>
> [...](d) "consumer": means any natural person who, in distance contracts covered by this Directive, is acting for purposes which are outside his business or professional activity; [...] (*our literal translation*)

We remark that in contrast with English, in Italian the second definition of *consumatore* is broader than the first one, since the term *professionale* (*professional*) does not include *commerciale* (*business*). This divergence of term definitions can often occur, since EUDs have usually a sectorial specific target. In this way, EUDs covering different sectors can provide different definitions, and as many views on the same concept. Lawyers and legislators started to put together highly sectorial concepts into more abstract concepts with broader meaning, in order to describe (complex) entities, such as the *consumatore* in all of its aspects.

In recent years, in the Italian legislation EUDs are not being implemented as single laws, but rather as groups of EUDs. The juridical concepts are defined as the union of all the sectorial concepts provided by the individual EUDs, as a result of the doctrinal interpretation process of directives. These problems are common to all European languages. Consider, for instance the definition of *consumer*, in the English version of the EUD 1999/44/EC, Art. 1.2 is:

> [...] (a) consumer: shall mean any natural person who, in the contracts covered by this Directive, is acting for purposes which are not related to his trade, business or profession; [...]

that has a different meaning with respect to the definition of *consumer* given in the Council Directive 90/314/EEC, Art. 2.4:

> [...] "consumer" means the person who takes or agrees to take the package ('the principal contractor'), or any person on whose behalf the principal contractor agrees to purchase the package ('the other beneficiaries') or any person to whom the principal contractor or any of the other beneficiaries transfers the package ('the transferee') [...]

The LTS should be able to represent both the more specific dimension related to the definitions in EUDs and the more abstract one which results from the doctrinal interpretation of European law. The LTS allows inserting the text paragraphs where *umeanings* are defined. However, to gain better understanding of legal concepts, it is often required to consider a broader fragment. For example, in the case of *consumer* the definition is not enough, and it is necessary to collect multiple paragraphs where consumer protection norms are presented and discussed.

2.2 Normative Change

Another big open issue to cope with in building tools for describing legal frameworks both at the European and at the national level is the *normative change* [7]. One major problem, well-known in the literature, is the update of *non-monotonic* ontologies and knowledge bases [8]. In other words, not necessarily ontologies and knowledge bases have a structure constant through time (e.g., see [9]): concepts and relations present in the ontology can become obsolete as new concepts and relations are added. This is

indeed the case of legal frameworks, that are continuously modified as new laws can modify paragraphs of old ones.

We can have two types of normative change: *explicit* change and *implicit* change. In the first case the new norm explicitly states the abrogation of a specific paragraph of an old law (for details on this line of investigation, please refer to [10]). Alternatively, the newer law can state a concept in contradiction to previous laws, but without mentioning them explicitly. In this case the concept stated by the new law becomes the current one; also, the parts of the old laws affected by changes (no longer updated) become obsolete.

3 From LTS 1.0 to LTS 2.0

In this Section we first summarize the functionalities of the existing LTS [6], and then we explain how it has been extended to cope with the new requirements described in the previous Section.

3.1 LTS 1.0

The main assumptions of our methodology come from studies in comparative law [11] and ontologies engineering [12]. Terms –*lexical entries* for legal information–, and concepts must be distinguished; for this purpose we use lightweight ontologies, i.e. simple taxonomic structures of primitive or composite terms together with associated definitions. They are hardly axiomatized as the intended meaning of the terms used by the community is more or less known in advance by all members, and the ontology can be limited to those structural relationships among terms that are considered as relevant.

We distinguish the ontology implicitly defined by EUD, the *EU level*, from the various national ontologies. Each one of these "particular" ontologies belongs to the *national level*: i.e., each national legislation refers to a distinct national legal ontology. We do not assume that the transposition of an EUD automatically introduces in a national ontology the same concepts that are present at the EU level.

Corresponding concepts at the EU level and at the national level can be denoted by different terms in the same national language.

A standard way to properly manage large multilingual lexical databases is to make a clear distinction among terms and their interlingual acceptions (or *axies*) [13].

In the LTS project to properly manage terminological and conceptual misalignment, we distinguish the notion of *legal term* from the notion of *legal concept* and we build a systematic classification based on this distinction. The basic idea in our system is that the conceptual backbone consists in a taxonomy of concepts (ontology) to which the terms can refer in order to express their meaning. One of the main points to keep in mind is that we do not assume the existence of a single taxonomy covering all languages. In fact, the different national systems may organize the concepts in different ways. For instance, the term *contract* corresponds to different concepts in common law and civil law, where it has the meaning of *bargain* and *agreement*, respectively [14]. In most complex instances, there are no homologous between terms-concepts such as *frutto civile* (legal fruit) and *income*, but respectively civil law and common law systems can achieve functionally similar operational rules thanks to the functioning of the entire taxonomy of national legal concepts [15]. Consequently, the LTS includes different ontologies, one for each involved national language plus one for the language of EU

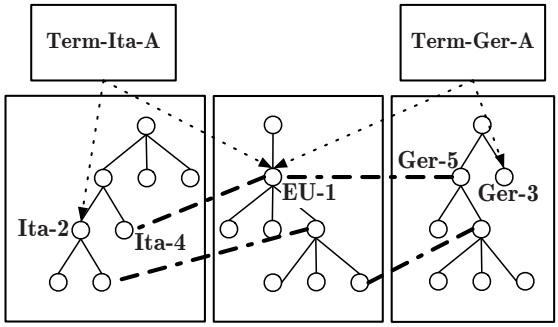


Fig. 1. Relationship between ontologies and terms. The thick arcs represent the inter-ontology "association" link.

documents. Each language-specific ontology is related via a set of *association* links to the EU concepts, as shown in Fig. 1.

Although this picture is conform to intuition, in the basic LTS it has been implemented by taking two issues into account. First, it must be observed that the various national ontologies have a reference language. This is not the case for the EU ontology. For instance, a given term in English could refer either to a concept in the UK ontology or to a concept in the EU ontology. In the first case, the term is used for referring to a concept in the national UK legal system, whilst in the second one, it is used to refer to a concept used in the European directives. This is one of the main advantages of LTS. For example *klar und verständlich* could refer both to concept `Ger-379` (a concept in the German Ontology) and to concept `EU-882` (a concept in the European ontology). This is the LTS solution for facing the possibility of a partial correspondence between the meaning of a term in the national system and the meaning of the same term in the translation of a EU directive. This feature enables the LTS to be more precise about what "translation" means. It makes available a way for asserting that two terms are the translation of each other, but just in case those terms have been used in the translation of an EU directive: within LTS, we can talk about direct EU-to-national translations of terms, and about *implicit* national-to-national translations of terms. In other words, we distinguish between *explicit* and *implicit* associations among concepts belonging to different levels. The former ones are direct links that are explicitly used by legal experts to mark a relation between concepts. The latter ones are indirect links: if we start from a concept at a given national level, by following a direct link we reach another concept at European level. Then, we will be able to see how that concept is mapped onto further concepts at the various national levels.

The situation enforced in LTS is depicted in Fig. 1, where it is represented that the Italian term *Term-Ita-A* and the German term *Term-Ger-A* have been used as corresponding terms in the translation of an EU directive, as shown by the fact that both of them refer to the same EU-concept `EU-1`. In the Italian legal system, *Term-Ita-A* has the meaning `Ita-2`. In the German legal system, *Term-Ger-A* has the meaning `Ger-3`. The EU translations of the directive is correct insofar no terms exist in Italian and German that characterize precisely the concept `EU-1` in the two languages (i.e., the "associated" concepts `Ita-4` and `Ger-5` have no corresponding legal terms). A

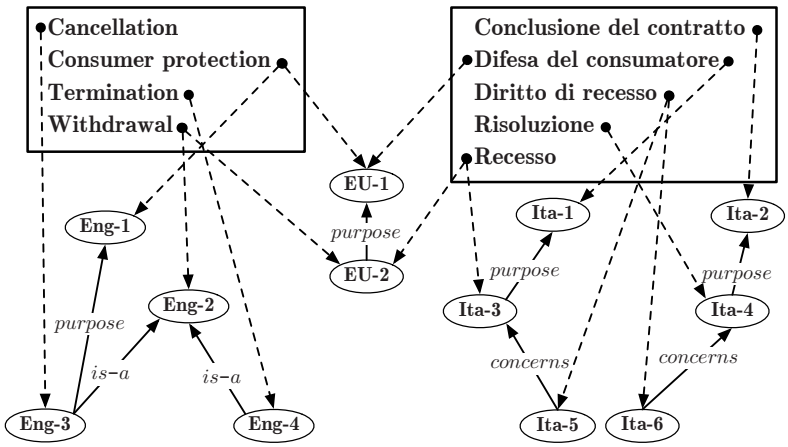


Fig. 2. An example of interconnections among terms.

practical example of such a situation is reported in Fig. 2, where we can see that the ontologies include different types of arcs. Beyond the usual *is-a* (linking a category to its supercategory), there are also the arcs *purpose*, which relate a concept to the legal principle motivating it, and *concerns*, which refer to a general relatedness. The dotted arcs represent the reference from terms to concepts. Some terms have links both to a National ontology and to the EU Ontology (in particular, *withdrawal* vs. *recesso* and *difesa del consumatore* vs. *consumer protection*).

The last item above is especially relevant: note that this configuration of arcs specifies that: 1) *withdrawal* and *recesso* have been used as equivalent terms (concept `EU-2`) in some European Directives (e.g., Directive 90/314/EEC). 2) In that context, the term involved an act having as purpose some kind of protection of the consumer. 3) The terms used for referring to the latter are *consumer protection* in English and *difesa del consumatore* in Italian. 4) In the British legal system, however, not all *withdrawals* have this goal, but only a subtype of them, to which the code refers to as *cancellation* (concept `Eng-3`). 5) In the Italian legal system, the term *diritto di recesso* is ambiguous, since it can be used with reference either to something concerning the *risoluzione* (concept `Ita-4`), or to something concerning the *recesso* proper (concept `Ita-3`).

3.2 Enhancing LTS with interpretation and abstraction

As described in Section 2.1, different pieces of legislations can bear different definitions of terms. Having different detailed definitions is important during the interpretation of very sectorial legal cases, but for the general case it is important to have a view that abstracts from the peculiarities of specific domains.

In order to solve this problem we introduced a new kind of ontologic relation called *INTERPRETED_AS*: it is a non transitive relation where the more general umeaning, that we will call *group leader* represents the abstracted concept that groups the meaning of a number of more specific umeanings, that are the sectorial umeanings defined in the individual EUDs or national laws (see Fig. 3).

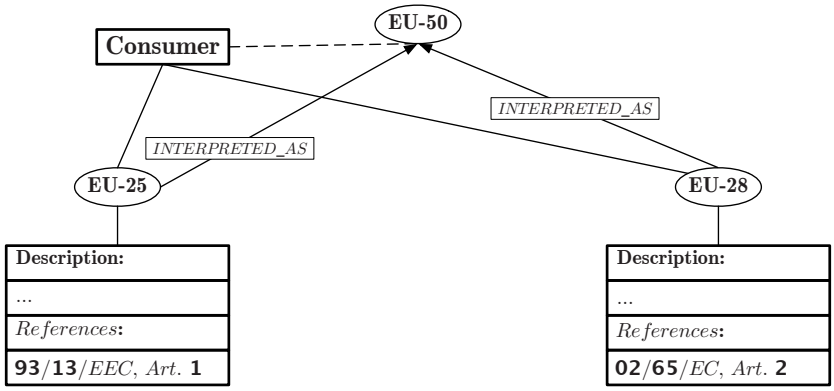


Fig. 3. Umeanings `Eu-25` and `Eu-28` are interpreted by the more abstract umeaning `Eu-50`, the link between `Eu-50` and the term "consumer" is implicit.

We have also introduced a number of constraints and integrity checks to ensure that the semantics of the grouping concept is respected and to improve the usability of the system: *i*) each umeaning can belong to a single group; *ii*) a group leader cannot exist without group members; *iii*) when the user searches into the umeaning database, more specific umeanings are excluded from the results unless the user explicitly asks to show them, i.e. only the group leaders are shown in the results. The need to contextualize concepts to the EUDs defining them leads to the need of more complex instruments to deal with the language of the norms. An umeaning is defined by the legal texts themselves; this makes clear that the creation of umeaning is a quite long task, because it requires from the user searching and reading a very large number of documents.

In order to ease this process, we developed a database that contains the full versions of the desired EUDs and national laws. In this way, the user can carry out his task according to the following workflow. *1*) The user creates a new *umeaning* linked with the term he wants to define; *2*) He selects relevant citation from legal text; consequently, the browser is redirected to a search page and the main term attached to the umeaning is used as the default query; *3*) After choosing one of the search results, the full text of the legal document is displayed, with the search terms highlighted; *4*) Finally, the user selects the text that will go in the citation with the mouse and confirms the insertion in the references database. Lastly, when the user searches for a term in the documents database, the search is not performed upon the exact words, rather with their roots, so for instance when performing the search on the term "contracts" also documents containing only "contract" will be found, this seems to enhance the information retrieval performance as shown in [16].

3.3 LTS with normative change

When a new normative is approved and enacted it can define a number of new umeanings; moreover it can happen that the same law can change a number of old umeanings defined by old laws. In particular, these old umeaning can become obsolete and no longer valid. We are aware of the difficulties concerning the modelling of the time in

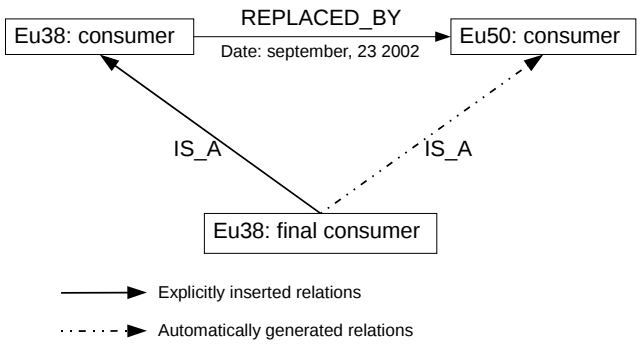


Fig. 4. An example of use of the *REPLACED_BY* relation.

artificial intelligence and in formal ontology too[5]. Anyway, in LTS we adopted a naive solution in order to manage the simpler situation concerning t In the LTS it was necessary to delete all old umeanings, causing the loss of all historic informations from the database, informations that are quite valuable to better understand the evolving of the normative. This problem was resolved by using the same solution adopted for the interpretation and abstraction of the norms (Section 3.2), i.e. empowering LTS with a new ontological relation called *REPLACED_BY*.

When the paragraph of an EUD defining an umeaning has been modified by a new EUD, the new one defines a new umeaning that will replace the old umeaning in the ontology. There will be a relation of type *REPLACED_BY* between the two umeanings, where the child umeaning is replaced by the more general umeaning. Also in this case the new ontological relation has some peculiar characteristics that distinguishes it from the usual ontological relations (Figure 4): *i*) a *REPLACED_BY* relation brings with it a new data field not present in the other relations: the substitution date; *ii*) when the user performs a search in the umeanings database the replaced ones will not be shown, unless the user asks for a certain past date, thus obtaining a snapshot of the legal ontology that was valid in that particular moment; *iii*) when a new umeaning replaces an old one all the ontological relations where the old umeaning appeared are automatically copied in the new umeaning. If some of them are no longer valid with the new umeaning, manual intervention from the user is required.

4 Conclusions

In this paper we discuss some features that have recently been introduced in the LTS, a tool for building multilingual conceptual dictionaries for the EU law. The tool is based on lightweight ontologies to emphasize the distinction between concepts and terms. Different ontologies are built at the EU level and for each national language, to deal with polysemy and terminological and conceptual misalignment.

The present work illustrates how to distinguish between concepts as they are defined in the text of the directives and the concepts representing the doctrinal interpretation

[5] E.g. see [17] for a general survey and [7] for normative systems

of the terms. Moreover, we point out how to deal with normative change by introducing a temporal dimension in ontologies.

Future work will involve exploring how to extend the LTS ontology, with special focus on the issue of populating it at the various levels by semi-automatic approaches [18].

References

1. Després, S., Szulman, S.: Merging of legal micro-ontologies from european directives. Journal of Artificial Intelligence and Law (February 2007)
2. Casanovas, P., Casellas, N., Tempich, C., Vrandecic, D., Benjamins, R.: OPJK modeling methodology. In: Proceedings of the ICAIL Workshop: LOAIT 2005. (2005)
3. Tiscornia, D.: The Lois Project: Lexical Ontologies for Legal Information. In: Proceedings of the V Legislative XML Workshop, European Press Academic Publishing (2007)
4. Vossen, P., Peters, W., Gonzalo, J.: Towards a universal index of meaning. In: Proc. ACL-99 Siglex Workshop. (1999)
5. Ajani, G., Boella, G., Lesmo, L., Martin, M., Mazzei, A., Rossi, P.: A development tool for multilingual ontology-based conceptual dictionaries. In: Proc. of 5th International Conference on Language Resources and Evaluation, LREC06, Genoa (2006) 1–6
6. Ajani, G., Boella, G., Lesmo, L., Mazzei, A., Rossi, P.: Terminological and ontological analysis of european directives: multilinguism in law. In: 11th Internation Conference o Arificial Intelligence and Law (ICAIL). (2007) 43–48
7. Palmirani, M., Brighi, R.: Time model for managing the dynamic of normative system. Electronic Government (2006) 207–218
8. Cadoli, M., Donini, F.M.: A Survey on Knowledge Compilation. AI Communications **10**(3–4) (1997) 137–150
9. The Gene Ontology Consortium: Gene Ontology: tool for the unification of biology. Nature Genetics `http://genetics.nature.com` **25** (2000) 25–29
10. Cherubini, M., Giardiello, G., Marchi, S., Montemagni, S., Spinosa, P., Venturi, G.: NLP-based metadata annotation of textual amendments. In: Proc. of WORKSHOP ON LEGISLATIVE XML 2008, JURIX 2008. (2008)
11. Rossi, P., Vogel, C.: Terms and concepts; towards a syllabus for european private law. European Review of Private Law (ERPL) **12**(2) (2004) 293–300
12. Klein, M.: Combining and relating ontologies: an analysis of problems and solutions. In: Workshop on Ontologies and Information Sharing, IJCAI'01, Seattle, USA (2001)
13. Sérasset, G.: Interlingual lexical organization for multilingual lexical databases in NADIA. In: Proc. COLING94. (1994) 278–282
14. Sacco, R.: Contract. European Review of Private Law **2** (1999) 237–240
15. Graziadei, M.: Tuttifrutti. In Birks, P., Pretto, A., eds.: Themes in Comparative Law. Oxford University Press (2004) –
16. Krovetz, R.J.: Word sense disambiguation for large text databases. PhD thesis, University of Massachusetts (1995)
17. Allen, J.: Towards a general theory of action and time. Artificial Intelligence **23**(2) (1984) 123–154
18. Cimiano, P.: Ontology Learning and Population from Text: Algorithms, Evaluation and Applications. Springer (2006)

Modeling Expert Knowledge in the Mediation Domain: A Mediation Core Ontology

Marta Poblet[1], Núria Casellas[2], Sergi Torralba[2], and Pompeu Casanovas[2]

[1] ICREA Researcher at the Institute of Law and Technology
[2] Institute of Law and Technology, Universitat Autònoma de Barcelona
(Bellaterra) 08193, Barcelona, Spain
marta.poblet, nuria.casellas, sergi.torralba, pompeu.casanovas@uab.cat
http://idt.uab.cat

Abstract. In this paper we introduce the Mediation Core Ontology (MCO), and the steps taken in order to model the expert knowledge on the mediation domain. MCO is created from scratch by eliciting practical knowledge from mediation experts to identify the basic working concepts of the domain. MCO offers initial support towards knowledge acquisition and reasoning and, in later steps, will serve as a general basis for the development of different mediation domain and sub-domain ontologies to be used by the ONTOMEDIA mediation platform, currently also under development.

1 Introduction

Online Dispute Resolution (ODR) is an umbrella domain that covers a full range of processes (i.e. negotiation, early neutral evaluation, conciliation, mediation, and arbitration) to handle disputes online. While it was sometimes viewed as the online equivalent of ADR (Alternative Dispute Resolution) processes, there is a growing consensus in specialized literature that considers ODR more than just the delivery of alternative dispute resolution (ADR) services through the Internet, especially since Katsh first suggested to give technology the role of a "four party" [1]. In this line, the emergence of a panoply of both new terminologies and typologies to systematize current ODR practices proves that the domain is becoming a branch of dispute resolution in its own right [2, 3, 4, 5].

For fifteen years now, ODR processes have evolved with the development of the Internet. As an example, ENSs (e-negotiation systems) deployed in the Web use different Internet technologies to actively assist negotiators, facilitators, and mediators [6]. Yet, some experts have warned that ODR service providers may be lagging behind the curve of recent developments in both Web 2.0 and Semantic Web [7, 8, 9].

The ONTOMEDIA project aims at filling this gap by designing an interactive, web-based mediation platform to assist disputing parties and mediators in identifying different options for the management and resolution of disputes in different domains.[3] One of the objectives of ONTOMEDIA is to model expert knowledge on mediation

[3] ONTOMEDIA: Platform of Web Services for Online Mediation, Spanish Ministry of Industry, Tourism and Commerce (Plan AVANZA I+D, TSI-020501-2008, 2008-2010).

as a domain independent process that, in turn, may be able to encompass different mediation sub-domains (i.e. commerce, family, health, workplace, environment, etc.). The ONTOMEDIA platform will therefore assist users in considering different options of mediation and guiding them throughout the online mediation process.

In this paper we describe the methodological approach taken for modeling expert knowledge on mediation processes, and outline the design of the Mediation Core Ontology (MCO). MCO thus represents the common and reusable structure of mediation processes, which will provide the platform with conceptual machine-processable knowledge regarding mediation events. This is one of the first attempts to design an ontology that models mediation processes within the dispute resolution field.

1.1 Mediation as a domain of knowledge

A meta-analysis of the relational justice domain (the justice produced through cooperative behavior, agreement, negotiation, or dialogue among actors in conflict or post-conflict situations) reveals that there are at least thirty disciplinary areas contributing to the development of the domain [10]. It therefore comes as no surprise if the mediation domain is populated with a full range of concepts, operational definitions, and models [11, 12]. To quote a recent example, Alexander identifies up to six models of mediation practice: settlement mediation, facilitative mediation, transformative mediation, expert advisory mediation, wise counsel mediation, and tradition-based mediation [12]. In addition, as far as it provides a new procedural and communicational framework for interaction, decision-making, and emotion expression [13] online mediation may substantially transform any of those models.

Mediation as a process While bearing in mind the many possible ways in which mediation might be defined and modeled, therefore, we have opted for an approach that emphasizes the representation of the procedural aspects of mediation over the epistemological and theoretical ones. This is not meant to be an entirely agnostic approach, since the focus on procedures already implies epistemological and theoretical choices. Similarly, the emphasis on procedural knowledge does not entail neglecting conceptual knowledge on mediation. Rather, we intend MCO to be a shareable and reusable ontology so that we needed to restrain these ontological commitments to a minimum [14].

Coherently, we propose to define mediation as a voluntary, non-binding process in which a neutral third party, the mediator, assists the parties in reaching a settlement of the dispute. This definition is consistent with the one proposed by the recent Directive 2008/52/EC,[4] and flexible enough to allow any number of disputing parties, roles, and procedural stages of mediation.

[4] The Directive 2008/52/EC of the European Parliament and of the Council of 21 May 2008 on certain aspects of mediation in civil and commercial matters defines mediation in article 3(a) as "a structured process, however named or referred to, whereby two or more parties to a dispute attempt by themselves, on a voluntary basis, to reach an agreement on the settlement of their dispute with the assistance of a mediator. This process may be initiated by the parties or suggested or ordered by a court or prescribed by the law of a Member State".

1.2 Ontologies, mediation and ODR

To date, there is no working ontology dealing with the fundamental concepts of mediation as a process. Certainly, there is precedent work on ontology design within related domains, namely the e-commerce field [15], task collaboration [16], negotiation [17], and negotiation agents [18]. There are also some ontologies that model different conflict events [19, 20] but in these cases the emphasis is put on terrorism and security issues rather than in conflict management.

Finally, there are a number of ongoing research projects that are currently developing ODR-related ontologies. The BEST project (BATNA Establishment using Semantic Web Technology) aims to provide disputing parties with information about their position in the negotiations before they seek professional assistance, and to assist them in the dispute or get information about the legal possibilities to claim compensations [5]. The ALIS Project (Automated Legal Intelligent System) combines game theory, computational logic, and legal reasoning to analyze the compliance of parties' requests in intellectual property disputes [21]. The CEN Workshop on Standardization of Online Dispute Resolution Tools has elaborated a basic ontology of ODR processes [6]. While BEST and ALIS are producing in fact legal domain ontologies (covering damage disputes and intellectual property respectively), the CEN ontology is domain-independent and, thus, the closest precedent to our work [22].

2 Mediation Core Ontology development

The initial stages of the ONTOMEDIA project have run in parallel with the elaboration of the White Book on Mediation in Catalonia, a project coordinated by the UAB Institute of Law and Technology [7]. The main purpose of the White Book is to provide Catalan lawmakers with in-depth research on the state-of-the-art mediation theories and practices as the basis for future legislation and policies. The White Book project has provided a unique opportunity to gather national and international leading experts and practitioners in a number of work sessions and workshops on concepts, methods, techniques and protocols of mediation.

The expert knowledge and support offered by the participants and the outcomes of the White Book project have been integrated in the methodological development cycle of MCO. The methodological steps followed, already established and shared by several ontology development methodologies (such as METHONTOLOGY [23], On-To-Knowledge (OTK) [24], HCOME [25] or UPON [26], etc.), take into account both the analysis of relevant textual materials towards ontology learning and the participation of experts during all the development process. These methodological requirements influence the general steps taken: a preparatory step (establishment of requirements), a development step (knowledge acquisition, conceptualization and formalization), and an evaluation stage [27]. In the following sections, we will describe the preparatory and development steps.

[5] BEST Project, `http://www.best-project.nl/index.shtml`.

[6] CEN Workshop on Standardization of Online Dispute Resolution Tools: `http://www.cen.eu/cenorm/businessdomains/businessdomains/isss/activity/ws_odr.asp`.

[7] White Book on Mediation in Catalonia: `http://idt.uab.es/llibreblanc/index.php?lang=english`.

2.1 Ontology requirements

MCO will serve as a general basis for the development of the mediation domain ontologies and sub-ontologies that will be used by the ONTOMEDIA platform. Therefore, it is directed at knowledge reuse, although it may also offer initial support towards knowledge acquisition and reasoning.

The knowledge acquisition stage is mainly based on the elicitation of expert knowledge. Nevertheless, existing upper ontologies (and legal core ontologies) are taken into account for design purposes. This knowledge acquisition process is guided by a list of questions establishing which knowledge ought to be included in the ontology and what type of answers ought the ontology to be able to give.

Table 1. Mediation Core Ontology (MCO) Requirements Specification Document

Purpose	Explicit expert knowledge in the mediation domain for knowledge reuse and for providing support towards knowledge acquisition and reasoning.	
Methodological approach	An expert-based methodology based on the main steps provided and shared by several current ontology methodologies (METHONTOLOGY, OTK, HCOME or UPON): 1) preparatory step, 2) development step, and 3) evaluation step. The knowledge acquisition process is mainly based on knowledge elicitation from experts, although is supported by knowledge acquisition from texts and guidance from theoretical approaches to the analysis of the mediation domain.	
Sources of knowledge	— —**C. questions** — — —**Other** —**sources**	What types of mediation exist? What characterizes them? Are there separate acts or situations within a mediation process? Which documents or other information sources are produced or used during a mediation process or stage? Which participants can take part in a specific type of mediation process? Which restrictions on the mediation process are caused by the topic of the mediation? What are the limitations on agents regarding the roles they might take in a mediation process? Expert elicitation (White Book project). Relevant regulations and legislation (e.g. Directive 2008/52/EC, EC Recommendation 98/257 & 2001/310).
Tool support	Statistic text analysis tools (JRef, Yoshikoder, AntConc, etc.)	
Ontology editor	Protégé v. 3.4.	
Reuse	No direct reuse of existing upper ontologies (modeling solutions from PROTON [28], LKIF-Core [29], CLO (DOLCE) [30] have been taken into account).	

2.2 Knowledge acquisition

From the knowledge acquisition perspective, the White Book outputs (early drafts, workshop papers, literature reviews, etc.) are a first-hand input for ontology design in ONTOMEDIA. We have analyzed these materials in consensus building sessions to identify a common conceptual framework broader enough to support different models and sub-domains of mediation. As a result, we elicited an initial taxonomy of concepts and relations, guided by the established competency questions (ORSD).

A second source of acquisition of knowledge has been drawn from ethnographic fieldwork, since one member of the team has been participating in a multiparty mediation process involving five mediators (this is work in progress). Participant observation has produced informal interviews with mediators conducted either individually or in group to elicit procedural knowledge used by domain experts in their practice. The translation of ethnographic findings into manageable knowledge leading to the design of ontologies relies on experience from related research projects [31, 32]. In this case, ethnographic research also loosely follows the guidelines of the EthnoModel, which are

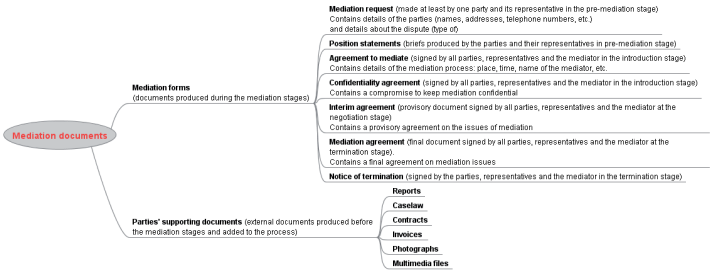


Fig. 1. Expert schema regarding mediation documents

defined as a set of generic heuristics that "may be used both by investigators to conduct ethnographic studies of work and by designers interested in system design" (i.e. plans, procedures, and coordination) [33]. We have complemented these previous inputs with an analysis of mediation procedures as deployed by major mediation services (both online and off line service providers). Again, we have benefitted here from synergies from the White Book project, where we have developed a template to analyze which mediation stages and related mediation forms are most usual among major service providers (up to 23 so far), regardless of the mediation sub-domain involved [34].

Finally, relevant existing regulations within the European Union (e.g. Directive 2008/52/EC of the European Parliament and of the Council of 21 May 2008 on certain aspects of mediation in civil and commercial matters) have been taken into account as regards concept definitions and linguistic use of terms. For example, extracted relevant terms in the mediation domain from European regulations are: *mediation, parties, dispute, agreement, process, mediator, information, resolution, provider*, etc.

3 Mediation Core Ontology (MCO)

The knowledge acquired in the previous phase (list of terms and conceptual schemas regarding knowledge required for the competency questions) from the experts has been further formalized in OWL-DL.[8] The current version of the Mediation Core Ontology has 62 classes. The main objective of the formalization stage was to model formally the main acquired concepts related to the mediation domain and to try to establish the most important relations between them.

Our approach has resulted in an initial taxonomic structure formed by the following concepts:

- `MediationAgent`: Includes all possible agents (actors) in the mediation domain.
- `MediationInformationSource`: All possible information sources, including forms that are created within the mediation process or `MediationForm` (e.g. `Agreement ToMediate`, or `NoticeOfTermination`), and other sources of information that can support the claims of the disputants.

[8] The ontology uses OWL DL constructs such as `owl:unionOf` and `owl:disjointWith`, together with cardinality values different from 0 or 1.

- `MediationTopic`: all topics that configure the different types of MediationProcess, for example, mediation regarding family issues, consumer related complaints, environmental issues, school or labour problems, etc. The mediation process, its agents and other related concepts may require different properties according to the topic or the particular problem underlying the process.
- `MediationProcess`: includes the different processes according to their topic. Thus, it includes as subclasses: `ConsumerMediation`, `SchoolMediation`, etc.
- `MediationProcessStage`: identifiable stages of a mediation process.
- `MediationSession`: identifiable situations taking part during the mediation process involving the different roles.
- `MediationRole`: all the possible roles that participants may assume in a mediation process (`Disputant`, `Mediator`, `ServiceProvider` are some of its subclasses).

Once this main hierarchy of concepts could be established, these concepts were specified and the main relations existing between them, elicited from experts, were also formalized. At the moment, 12 `owl:objectProperty` and 1 `owl:dataTypeProperties` have been included in the ontology.[9] More complex relations and concept definitions have also been specified to allow reasoning on the mediation domain, and facilitate its reuse and specification by the specific ontology for the OntoMedia platform. For example, the ontology includes the specification of the idea that a mediation process requires at least two disputants and one mediator, a termination stage is a mediation process stage that produces a notice of termination (document), an environmental mediation is a mediation process about an environmental topic, etc.

4 Conclusions and future work

In this paper we have introduced the Mediation Core Ontology (MCO), which represents a basic and flexible conceptual structure of mediation processes with minimal ontological commitments. We also offered an overview of the knowledge acquisition process and conceptualization stages leading to its design. Currently, the Mediation Core Ontology includes only the concepts related to the core mediation domain, and may be of use towards knowledge acquisition and reasoning tasks. Future work will include its modular extension to the different mediation subdomains (i.e. labour mediation, family mediation, etc.) and will be adapted for the use of the ONTOMEDIA platform.

Moreover, the Mediation Core Ontology is currently under submission for evaluation to an expert panel from the White Book project, and will be further tested (and refined if necessary) with the instantiation of several currently available mediation services.

Once the ontology has undergone the evaluation and refinement processes it will be made publicly available.

Acknowledgments. We would like to thank the anonymous reviewers for their comments. The research presented in this paper has been developed within the framework of two different projects: (i) ONTOMEDIA: Platform of Web Services for Online Mediation, Spanish Ministry of Industry, Tourism and Commerce (Plan AVANZA

[9] For example, `isMajor` is formalized as a data type property with domain `NaturalPerson` and a boolean range.

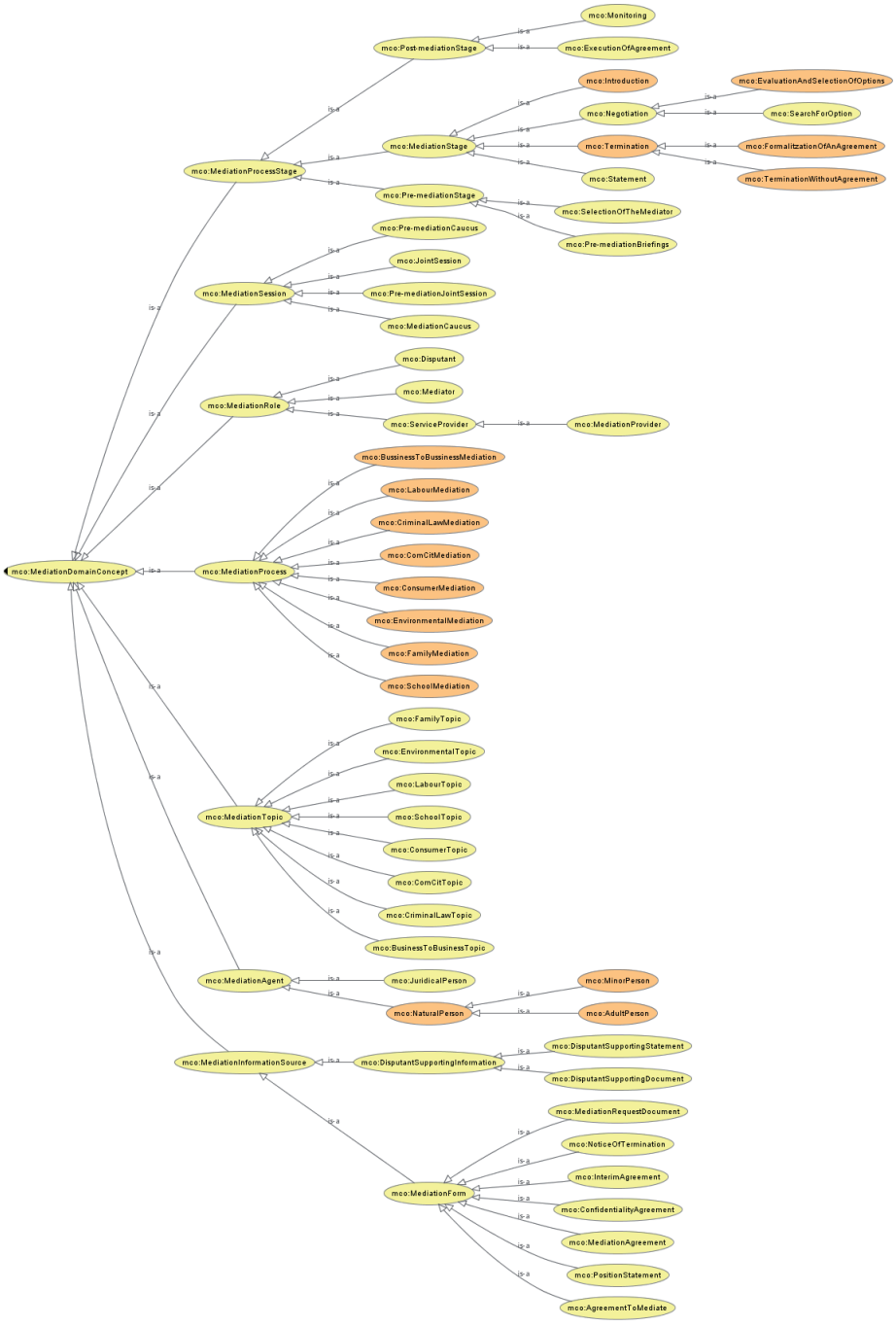


Fig. 2. An outline of the Mediation Core Ontology classes

I+D, TSI-020501-2008, 2008-2010); (ii) ONTOMEDIA: Semantic Web, Ontologies and ODR: Platform of Web Services for Online Mediation (2009-2011), Spanish Ministry of Science and Innovation (CSO-2008-05536-SOCI).

References

1. Katsh, E., Rifkin, J.: Online Dispute Resolution: Resolving Conflicts in Cyberspace. Jossey-Bass Inc., San Francisco (2001)
2. Mann, B.L.: Smoothing some wrinkles in online dispute resolution. International Journal of Law and Information Technology **17**(1) (2008) 83–112
3. Thiessen, E., Zeleznikow, J.: Technical aspects of online dispute resolution - challenges and opportunities. In Tyler, M.C., Katsh, E., Choi, D., eds.: Proceedings of the Third Annual Forum on Online Dispute Resolution, Melbourne, Australia, 5-6 July 2004. (2004)
4. Lodder, A., Thiessen, E.: Artificial intelligence and odr. In Katsh, E., Choi, E., eds.: Proceedings of the United Nations Forum on Online Dispute Resolution (ODR): Technology as the "Fourth Party", Palais des Nations, Geneva, June 30 - July 1, 2003. (2003)
5. Gordon, T., Märker, O.: Mediation systems. In Märker, O., Trénel, M., eds.: Neue Medien in Der Konfliktvermittlung - Mit Beispielen Aus Politik Und Wirtschaft. Edition Sigma, Berlin (2002) 61–84
6. Kersten, G., Lai, H.: Negotiation support and e-negotiation systems: An overview. Group Decision and Negotiation **16** (2007) 553–586
7. Rule, C.: Making peace on ebay: Resolving disputes in the world's largest marketplace. ACResolution **Fall** (2008) 8–11
8. Hattotuwa, S.: The future of online dispute resolution (odr): Technologies to keep an eye on. In: Proceedings Online Dispute Resolution Forum, June 22. (2008)
9. Poblet, M.: Bringing a new vision to online dispute resolution. In Poblet, M., ed.: Expanding the Horizons of ODR: Proceedings of the 5th International Workshop on Online Dispute Resolution (ODR Workshop'08), CEUR-Workshop Proceeding Series. Volume 430. (2008) 1–7
10. Casanovas, P., Poblet, M.: Concepts and fields of relational justice. In Casanovas, P., Sartor, G., Casellas, N., Rubino, R., eds.: Computable Models of the Law: Languages, Dialogue, Games, Ontologies. Volume 4884. Springer-Verlag, Heidelberg (2008) 323–339
11. Herrman, M.S., Hollett, N., Gale, J.: Mediation from beginning to end: A testable model. In Herrman, M., ed.: The Blackwell Handbook of Mediation: Bridging Theory, Research, and Practice. Blackwell Publishing, Malden (MA) (2006) 19–78
12. Alexander, N.: The mediation metamodel: Understanding practice. Conflict Resolution Quarterly **26**(1) (2008) 97–125
13. Poblet, M., Casanovas, P.: Emotions in odr. International Review of Law, Computers & Technology **21**(2) (2007) 145–156
14. Gruber, T.R.: Toward principles for the design of ontologies used for knowledge sharing. International Journal of Human-Computer Studies **43**(5-6) (1995) 907–928 Substantial revision of paper presented at the International Workshop on Formal Ontology, March, 1993, Padova, Italy. Available as Technical Report KSL 93-04, Knowledge Systems Laboratory, Stanford University.
15. Walton, D., Godden, D.M.: Persuasion dialogue in online dispute resolution. AI and Law **13** (2005) 273–295

16. Tamma, V., Phelps, S., Dickinson, I., Wooldridge, M.: Ontologies for supporting negotiation in e-commerce. Engineering Applications of Artificial Intelligence **18** (2005) 223–236
17. Ermolayev, V., Keberle, N., Tolok, V.: Oil ontologies for collaborative task performance in coalitions of self-interested actors. In Arisawa, H., Kambayashi, Y., Kumar, V., Mayr, H.C., Hunt, I., eds.: ER Workshops 2001. Volume 2465 of Lecture Notes and Computer Science. Springer-Verlag, Heidelberg (2002) 390–402
18. Anumba, C.J., Ren, Z., Thorpe, A., Ugwu, O., Newnham, L.: Negotiation within a multiagent system for the collaborative design of light industrial buildings. Advances in Engineering Software **34** (2003) 389–401
19. Tanev, H., Wennerberg, P.: Learning to populate an ontology of politically motivated violent events. In Fogelman-Soulié, F., ed.: Mining Massive Data Sets for Security. IOS Press, Amsterdam (2008) 311–322
20. Smart, P.R., Russell, A., Shadbolt, N.R., Schaefel, M.C., Carr, L.A.: Technical demonstrator system for enhanced situation awareness. The Computer Journal **50**(6) (2007) 703–716
21. Cevenini, C., Fioriglio, G.: Ict-supported dispute resolution. In P. Casanovas, G. Sartor, N.C.R.R., ed.: Computable Models of the Law: Languages, Dialogue, Games, Ontologies. Volume 4884 of Lecture Notes and Artificial intelligence. Springer-Verlag, Heidelberg (2008) 321–322
22. CEN: [draft] workshop agreement on standardisation of online dispute resolution tools (open to comments, final version to be published on june 2009), 2009-02-16,. Technical report, CEN (2009)
23. Gómez-Pérez, A., Fernández-López, M., Corcho, O.: Ontological Engineering. With examples from the areas of Knowledge Management, e-Commerce and the Semantic Web. Advanced Information and Knowlege Processing. Springer-Verlag, London (2003)
24. Sure, Y., Studer, R.: A methodology for ontology-based knowledge management. In Davies, J., Fensel, D., van Harmelen, F., eds.: Towards the Semantic Web. Ontology-driven Knowledge Management. John Wiley & Sons, LTD, Chichester, England (2003) 33–46
25. Kotis, K., Vouros, A.: Human-centered ontology engineering: The hcome methodology. Knowledge and Information Systems **10**(1) (July 2006) 109–131
26. Nicola, A.D., Missikoff, M., Navigli, R.: A proposal for a unified process for ontology building: Upon. In Andersen, K.V., Debenham, J.K., Wagner, R., eds.: Database and Expert Systems Applications (DEXA). Volume 3588 of Lecture Notes in Computer Science., Springer (2005) 655–664
27. Casellas, N.: Modelling Legal Knowledge Through Ontologies. OPJK: the Ontology of Professional Judicial Knowledge. PhD thesis, PhD thesis, Faculty of Law, Universitat Autònoma de Barcelona, Barcelona (2008)., Faculty of Law, Universitat Autònoma de Barcelona, Barcelona (2008)
28. Terziev, I., Kiryakov, A., Manov, D.: D1.8.1 base upper-level ontology (bulo) guidance. SEKT IST-2003-506826 Deliverable 1.8.1, SEKT, EU-IST Project IST-2003-506826, Ontotext Lab, Sirma Group (Bulgaria) (2005)
29. Breuker, J., Hoekstra, R., Boer, A., van den Berg, K., Rubino, R., Sartor, G., Palmirani, M., Wyner, A., Bench-Capon, T.: Owl ontology of basic legal concepts (lkif-core). Deliverable 1.4 D.1.4, ESTRELLA project (IST-2004-027655) (2007)
30. Gangemi, A., Sagri, M.T., Tiscornia, D.: A constructive framework for legal ontologies. In Benjamins, V.R., Casanovas, P., Breuker, J., Gangemi, A., eds.: Law

and the Semantic Web. Legal Ontologies, Methodologies, Legal Information retrieval, and Applications. Volume 3369 of Lecture Notes in Computer Science. Springer-Verlag Berlin Heidelberg (2005) 97–124
31. Casanovas, P., Casellas, N., Vallbé, J.: An ontology-based decision support system for judges. In Casanovas, P., Breuker, J., Klein, M., Francesconi, E., eds.: Channelling the legal information flood. Legal ontologies and the Semantic Web. Volume 188 of Frontiers in Artificial Intelligence and Applications. IOS Press, Amsterdam (2009) 165–176 ISBN 978-1-58603-942-4.
32. Poblet, M., Vallbé, J., Casellas, N., Casanovas, P.: Judges as it users: the iuriservice example. In Cerrillo, A., Fabra, P., eds.: E-justice: Using Information Communication Technologies in the Court System. IGI-Global, USA (2008)
33. Iqbal, R., Gatward, R.A., James, A.E.: A general approach to ethnographic analysis for systems design. In: Proceedings of SIGDOC, Coventry, UK, ACM (2005) 34–40
34. Poblet, M.: Mediació i tecnologia: Estat de l'art. deliverable et.11.1 of the white book on mediation in catalonia. Technical report (2008)

Knowledge Representation and Modelling Legal Norms: The EU Services Directive

Doris Liebwald *

Director of the Vienna Centre for Computers and Law, Austria
d@liebwald.com

Abstract. This paper presents an ontology based model which assists the user in formally specifying her or his information demand and in turn to deliver information across diverse authorities and local and functional jurisdictions, but individualised to the user's needs. For integration of information from and about different sources and relevant authorities, an information layer model is used. Text modules allow for flexibility, the reuse of text, and individualisation of information. Although the focus of this paper is on the transposition of the information duties imposed by the EU Services Directive, most considerations also apply to legal information and transaction portals in general, especially those which need to represent broad as well as in-depth information.

1 Introduction

This paper describes some elements of a draft concept developed by the author within the e-government framework of the Federal Chancellery of Austria with regard to the electronic transposition of the Directive on Services in the Internal Market 2006/123/EC. The Services Directive aims to facilitate the cross-border provision of services within EC-Member States. Legal and administrative barriers, which hinder SMEs from making use of their freedoms to establish and to provide services, are to be removed to boost cross-border service provision. To reach this goal, the Directive enshrines inter alia that all the requirements applicable to providers must be easily accessible at a distance and by electronic means, and that this information must be provided in a clear and unambiguous manner and in plain and intelligible language (Art. 7). Moreover, it must be ensured that all procedures and formalities relating to access to a service activity and to the exercise thereof may be easily completed, at a distance and by electronic means (Art. 8).

Despite the many exceptions, the Services Directive takes a horizontal approach: it establishes common rules for service providers. Services within the meaning of EC Treaty (Art. 49) are all service activities normally provided for remuneration, in particular activities of an industrial character, of commercial character, of craftsman and

* The views expressed in this paper are entirely and solely those of the author. Article published in Casellas, N., Francesconi, E., Hoekstra, R, and Montemagni, S. *LOAIT 2009 3rd Workshop on Legal Ontologies and Artificial Intelligence Techniques, joint with 2nd Workshop on Semantic Processing of Legal Text*, June 8th 2009, Barcelona, Spain.

of the professions. The Austrian taxonomy differs from this EU definition, as a co-extensive concept of "services" does not exist. Many of the relevant Austrian rules are vertical in the sense that they are activity-related (particular rules for directory publishers, chimney sweepers, private tutors, consulting engineers, veterinaries, etc.). Furthermore, Austria is a Federal State, therefore legislation and enforcement of law on national, regional or local levels need to be incorporated, and a variety of different "competent authorities" has to be involved.[1] Bearing the complexity of the Austrian legal framework in mind, representation of cross-linked knowledge about procedures, formalities and other requirements applicable to providers is a very challenging task.

This paper focuses on the information presentation component whilst taking into account that information must be connected to the proper procedures and formalities. Issues relating to administrative back-office processes are not addressed.

2 An Information Portal for Service Providers

The concept presented here establishes an information portal which is capable of interlinking text, text elements and meta-data from different levels and sources, in order to satisfy the user's information demand. According to the specific needs of a user, all relevant text elements have to be identified, selected, and sequenced. To do so, an intelligent, guided navigation system combined with a small question answering system, both based on formal semantic notations, is chosen. Since legal laypeople generally prefer and can better utilize an intelligent navigation system as opposed to searching for foreign or legal concepts, the emphasis is on classification and navigation. For integration of information from and about different sources and relevant authorities, an information layer model is used. This model allows for distributed maintenance of content by the respective authorities. Since the system and its information content have to be developed to a large extent from scratch, and the resulting information portal will have to deliver individualised information units to the user, legal knowledge representation as a top-down approach is employed. A semantic network will not be sufficiently expressive for this task; it must be extended by terminological logic, which allows for negations, non-taxonometric relations and the inference procedures subsumption and instance-classification.[2]

Open textured concepts, the open structure of law and the need for abstract, ex ante interpretation of legal norms and administrative practice are crucial points within such a legal information portal. In many cases constraints will have to be weakly encoded, accompanied by textual explanations and links to further information and supporting bodies. This is not a deficiency of the technical system, but necessary to reflect the special demands of the legal system and to safeguard legal certainty.

The idea of an electronic legal information portal should not be confused with face-to-face legal advice. In a conversation, the adviser will be aware of the individual

[1] To learn more about the Services Directive see [1] and [6], for the Austrian perspective [9]. The situation in Germany is similar, cf. [4], especially Chapter D.

[2] An interesting approach is taken by Salhofer/Stadlhofer [15]. They use a comparatively easy to use and easy to maintain concept tree for goal discovery, on which ontology based forms can be automatically generated. Though it is also their intention to hide complexity from the user, a concern which is supported only to some extent by this paper, the overall approach is of great value for a public information and transaction portal like the one established here.

context of the question. She or he will know why a specific question is asked and what the questioner is going to do with the answer. The adviser may check back, or switch to more adequate language if necessary, and has a good chance to detect misconceptions. On the contrary, an information portal has to work on a more abstract level and without a direct human verification loop. Of course, there already exist some attractive electronic legal advice systems which try to simulate face-to-face advice, but feasible and trustworthy applications are restricted to very narrow areas or specific topics of law, and they usually do not work with cross-border concepts.[3]

2.1 The User's Perspective

A provider who wants to establish in or to deliver services to another EU Member State is in general not familiar with the respective foreign legal system. Perhaps she or he is also not completely familiar with the language and it is likely that she or he lives in a different world of concepts [8]. Therefore a foreign provider will not be able to asses if the activity in question is, for example, covered by the Austrian Crafts and Trade Code, and if so, under which part of it. Maybe a corresponding activity does not exist or exists with a different meaning in the target country, e.g. dental care of horses is in Germany a craft, but in Austria it is just part of the work of the veterinary. Even "traditional" professions may be regulated differently, may allow for more or less activities or may impose deviant or unexpected requirements, which may have no counterpart at all in the provider's home country. In Austria for many activities (within and beyond the Crafts and Trade Code) proficiency has to be proven, and rules of practice may be spread over several laws. Even for activities which do not demand proof of proficiency, a vast number of professional rules may be applicable. Finally, the procedures and formalities a provider must satisfy to access and exercise her or his service activity are not restricted to the professional regulations; one may consider, for example, rules regarding the operating site, the equipment, the commercial register, social insurance duties, etc.

Since the sprit of the Directive is not to make the provider read the law but to make the provider comprehend the law, and in particular to make the provider recognise the requirements imposed on her or him by law and order, it is not sufficient to present the original text of the legal norms. The provider needs intelligible, unambiguous and purposeful information, delivered via an easy-to-use interface, which leads her or him through the labyrinth of the Austrian legal system. On basis of the information provided, the user should be aware of the requirements imposed on her or him, be able to select the effectively necessary formalities and procedures, and to recognize if procedures and formalities depend on each other or are concurrent.

Legal information is better understood by laymen when presented in (real life) context, e.g. based on life- or business events.[4] To bundle user relevant information in life- or business events allows for interconnecting multiple information sources, administrative bodies and other organisations.[5] Additionally, provided that texts are appropriately phrased, structured and annotated, the same information content may be

[3] Cf., e.g. the BEST-project [17], `http://www.best-project.nl`.

[4] For a general description on administrative portals based on life- or business events and further references see [11], pp. 218-230. See also [5].

[5] Proper examples for administrative information portals based on business events *and* trying to involve all administrative levels are the Austrian Amtshelfer `http://help-business.gv.at`, the Dutch Overheid voor ondernemers en organisaties

presented or made accessible under different perspectives. Hence the approach taken here is to assign information and processes to service activities and to organise the service activities under canonical business situations. An intelligent semantic class hierarchy should enable the user to shift to related or overlapping business situations or activities, without having to start from the very beginning.

At a first glance, it seems that an in-depth individualisation assists the user best. On the other hand, individualisation must be restricted for reasons of complexity, maintainability, and liability. Additionally, in-depth individualised information may lead the user to get caught in details while loosing the overall view on the whole issue. Maybe the user did not decide about all the details yet, or perhaps she or he is flexible and is searching for variants and options. Here a middle course is attempted. Because of the complexity of the vertical legal rules some individualisation is inevitable, but the user may not become restricted in her or his course of action by partial information or per computer code. One has to be aware that a reduced representation of the complexity of the legal system as well as of the reality, which is done by describing standardised life- or business situations in plain language, produces incomplete knowledge. If complexity is concealed from the reader, her or his scope of behaviour and action will be restricted. The law may of course be presented in a less complicated manner, but not be shortened to fit on the screen or to virtually satisfy the call for simplification of procedures by the Services Directive (Art. 5). Therefore the user has to be given textual information about the reasons and consequences on the differentiations made. Furthermore, the system must fairly point out its limitations and, when indicated, forward the user with her or his information need to a more proper source or to an individual advisory service.

2.2 Starting with some Questions

The order of structural elements arises from the relevant European and Austrian legal framework. Only the essential considerations are mentioned below. The basic structure is the following:

I Select Country of origin
II Select type of provision of services:
  a. Establishment in Austria
  b. Provision of services in Austria without being established in Austria
III If 2 b. was selected: Posting of workers yes/no
IV Select kind of activity (profession)
V If 2 a. was selected and if relevant/applicable:
VI If relevant/applicable: Select location (place of exercise of service activity)
  a. Select legal form of business (sole trader, private ltd. company,...
  b. Select specific business situation (branch, agency,...

For individualisation of information, the different information needs and different underlying requirements national and foreign EU providers have, must be considered. Moreover, even the differentiation between the citizenship of a natural person and the country of origin of a business may become relevant. A detailed breakdown would

`http://www.overheid.nl/ondernemers`, or the Australian Business Entry Point `http://www.business.gov.au`. See also the BASIS Public Services Broker Study [14] and `http://www.basis.ie`. The focus of the BASIS study was, however, not on information-oriented services but rather, on transaction-oriented services.

significantly increase the complexity of the system. Nevertheless, the system should at least offer the possibility to add text elements tailored to particular needs of providers belonging to specific countries or groups of countries. In case of multilingual content, this option would also allow for language selection.

In the next step a differentiation between providers who want to establish in Austria, and those who want to provide services for a limited period in Austria but are established in another Member State, has to take place. These two situations result in partially differing information needs and differing procedures and formalities. On the one hand, there is the option to describe both situations in one go. This offers the advantage of a more proper way to deal with overlapping contents, and with problems to subsume the real life situation correctly. On the other hand, this would result in longer texts, in which a large number of text elements might be irrelevant for many users. Furthermore, the texts concerning cross-border services must be blanked out for national providers. The final decision was to propose two separate information channels, to develop a common textual connector, to share text elements if applicable, and to allow the user to switch between the two situations.

An essential element is the selection of the kind of activity. Ultimately, the only possibility to guide a foreign provider through the Austrian labyrinth is to connect all relevant information, procedures and formalities to activities or groups of activities and to ask the user for the activity she or he wishes to exercise. Certainly, a valid list of all possible service activities (that would be a few thousand) does not exist, and since law is abstract and the matter a dynamic one, a complete list or description is not achievable at first. However, starting with common, frequently requested activities and working out the feasibility of the rest in the long-term is undoubtedly the best approach.

The question still remains, how can the user find the correct service activity? A foreign provider will be accustomed to different concepts and may assign different meanings to similarly named terms. Therefore the decision was to revert to the relevant parts of the NACE[6] 2.0 classification of economic activities. NACE is a five-level classification primarily used for statistical matters within the EC [7]. Since NACE is based on an EC Regulation,[7] it is available in all official languages of the EU and allows the user to navigate in her or his preferred language. Service providers may also be familiar with it from its use in their home country, e.g. for collecting statistical data. For this purpose the relevant parts of NACE have to be extracted, reduced to the levels necessary, and supplemented by subordinated "Austrian" activities. Additionally, a short "job description" in simple and easy to understand terms has to be assigned to each activity to help the user determine which actions are encompassed by a specific concept. This job description should also include relations to similar or overlapping activities. The classification work may be supported by the Austria-specific subclasses and the alphabeticum as developed by Statistics Austria.[8] Provided the basis is well elaborated, a semantic search could also be implemented at a later stage, e.g. incorporating multilingual thesauri.

Finally, the place of exercise of the activity may be crucial for allocation of information and procedures. In regards to service providers without permanent establishment

[6] Nomenclature générale des activités économiques dans les Communautés européennes.

[7] Regulation 3037/90/EWG recently amended by Regulation 1893/2006/EC.

[8] Statistics Austria, `http://www.statistik.at/web_de/klassifikationen/oenace_2008_implementierung/`.

in Austria, localization is circumstantial and better solved by providing summarized information on regional differences or regional authorities where indeed essential. In respect to cross-border providers who want to establish in Austria, localization on the regional and/or local level may become prerequisite. To serve all relevant constellations the localization tool must be based on postal code level and be connected to an advanced directory reflecting the local, regional and federal jurisdictions. In interaction with the information layer model, localization has to take place at the point where the provision of non-localized information is inadequate.

2.3 Structured Representation of Information

Since law is complex and involved authorities are numerous, and as expert knowledge is usually dispersed over the involved authorities, it will not be possible to develop and maintain all relevant information at one central point. Therefore the information portal presented herein is constructed to be a knowledge base and a directory at the same time. The knowledge base will primarily consist of information on federal level, and basis information on regional level. As regards to electronic procedures, the system must operate as a directory, but may support the development of interoperable processes. To support flexible integration of distributed or shared information sources and processes, text modules and an information layer model are used.

The goal is not to describe any and every activity and business event separately, but to use text elements or text blocks and to assemble them on a case-by-case basis in order to obtain continuous and individualised descriptions, and respectively instructions for the user. The degree of formalisation of course differs, some activities will need to be handled separately, other activities, e.g. those covered by the Crafts and Trade Code, leave more room for formalisation (like common requirements for all or at least groups of professions). The module technique allows also for integration of text elements from external sources into the first layer view, for example job descriptions and professional rules as developed and collected by the Austrian Chamber of Commerce within its own information system.

To work with text modules is, however, sophisticated: not only is the arrangement of the single elements challenging, but it also demands high standards of verbalisation to produce comprehensible and coherent descriptions as a result. Additionally, the editor support must be comprehensive, as changing a text for one instance will change it for all places it is reused as well, and applicability of the modification for all instances of the text block must be checked.

Additionally, an information layer model consisting of two layers is used. The abstract upper level presents basic information on the chosen service activity (or a group of service activities) and the related processes, in context of the chosen business situation. This is done across the diverse authorities (and other stakeholders) and local and functional jurisdictions. Only in a few cases will it be impossible to provide abstract basic information without preceding regionalisation. On this upper level the user shall be given a survey of all relevant requirements, procedures and formalities. The upper level information must be adequate to enable the user to recognize and identify those requirements which apply to her or him, and to further specify any possible supplementary information need and her or his line of action.

The second layer provides detailed information about single elements of the upper level, especially in regard to specific requirements of formalities and procedures. Consequently, specification of authorities and their functional and local jurisdiction must

take place within the second layer. At this point electronic procedures or forms may also be integrated or linked if existing. In a next step a SOA to enable semantic search of Web Services [2], [12], [16] could be modelled.[9]

This approach not only allows for structured integration of information subject to distributed competencies, it also allows for distributed supply of content. At this juncture it does not matter if external actors bring in content or if the second level links to external content. The latter variant will be more attractive for those authorities which do not want to give up their individual appearance or own information portals. Both variants of course assume a coordinated network and some agreement on wording, structuring and quality of texts, especially since the text must be coherent in regard to first layer information.

3 Observations

The considerations within this paper rely to some extent on a small prototype application developed in summer 2008.[10] The prototype was built to visualise the requirements that a system has to meet to fulfill the information duties of the Services Directive. It deals with two rather complex service activities, and since its task was to be just a showcase, it is predominantly hard-coded. The lesson learned from the prototype: aside of structural and layout deficiencies, the application was devised too simplistic and turned out to be unable to transport the complexity of the existing legal framework in a way which is transparent and useful to the user.

The concept provided herein is decidedly more intricate, and thus able to carry the demands of the existing legal framework. The details on its implementation are, however, not yet certain. But even though the Directive's transposition deadline (by 28th December 2009) is pressing, a sustainable system based on an overall plan, which may be finalised in all its intricacies at a later point in time, should be given priority over a hastily constructed portal which is inadequate.

At the end of this paper the reader may question the actual need for such complexity in the law, but this is outside of the scope of this discussion. What is certain is that administrative simplification should not be tackled by means of modern ICT alone, but also deliberate techniques such as legal and regulatory measures and process reengineering [13].

References

1. Breuss, F. et al.: Services Liberalisation in the Internal Market. Springer, Wien (2008).

[9] To develop a SOA to integrate the Austrian e-government landscape is never a trivial task, since there exists a high number of isolated and incompatible applications (some of which offer input or output interfaces for data transfer) on all levels of administration. For preliminary work on a common architecture in regard to the Services Directive see [3].

[10] Available at `http://www.help.gv.at:81/dlr/` (un/pw = dlr/showcase; choose "DLR-Assistent"; only the services "Personenbetreuung" and "Stukkateure und Trockenausbauer" are valid). The prototype is not built on information layers and provides all relevant information at once.

2. Bruijn, J. de et al.: Modeling Semantic Web Services. Springer, Berlin (2008).
3. e-Government Bund-Länder-Gemeinden: E-Government Architektur zur Dienstleistungsrichtlinie v. 1.0.0. (2009), http://www.ref.gv.at/E-Government.1817.0.html.
4. Deutschland-Online: Deutschland-Online-Vorhaben IT-Umsetzung der Europäischen Dienstleistungsrichtlinie. Projektbericht, Stand 24.09.2008, http://213.216.17.150/DOL/Bericht_Anlagen/Projektbericht_Langfassung.pdf.
5. European Commission: Consultation Document of the for a Future Policy Paper on Pan-European Government E-Services (April 2002). ENTR-D-2/PMU D(2002).
6. European Commission: Handbook on Implementation of the Services Directive. EC Publications Office, Luxembourg (2007).
7. Eurostat: Working Paper NACE Rev. 2. EC Publications Office, Luxembourg (2008).
8. Liebwald, D.: Semantic Spaces and Multilingualism in the Law: The Challenge of Legal Knowledge Management. In: Casanovas, P. et al. (eds.) 2nd LOAIT 2007. CEUR Workshop Proceedings, pp. 131–148. CEUR-WS.org (2008).
9. Liebwald, D.: Verwaltungsvereinfachung unter der Dienstleistungsrichtlinie. Zeitschrift für Verwaltung ZfV 6/2008, pp. 751–763. LexisNexis, Wien.
10. Laarschot, R. van et al.: The Legal Concepts and the Layman's Terms. In: 18th JURIX 2005, 115–126. IOS Press, Brussels (2005).
11. Lucke, von, J.: Hochleistungsportale für die öffentliche Verwaltung. Josef EUL Verlag, Köln (2008).
12. Mitrakas, A. et al. (eds.): Secure E-Government Web Services. IGI Global, Hershey (2007).
13. OECD: Cutting Red Tape: National Strategies for Administrative Simplification. OECD Editions, Paris (2006).
14. Price Waterhouse Coopers: BASIS Public Services Broker Study. (Irish) Department of Enterprise, Trade & Employment, Dublin (2001), http://www.epractice.eu/en/library/281326.
15. Salhofer, P., Stadlhofer, B.: Ontology Modeling for Goal Driven E-Government. In: HICSS-42 2009, pp. 1–9, IEEE Press, New York (2009).
16. Studer, R. et al. (eds.): Semantic Web Services. Springer, Berlin (2007).
17. Uijttenbroek, E.M. et al.: Retrieval of Case Law to Provide Layman with Information about Liability: Preliminary Results of the BEST-project. In: Casanovas, P. et al. (eds.) Computable Models of Law. LNCS, vol. 4884, pp. 291–311. Springer, Berlin (2008).

AGILE: From Source of Law to Business Process Specification

Alexander Boer and Tom van Engers

Leibniz Center for Law
University of Amsterdam
A.W.F.Boer@uva.nl

Abstract. The knowledge management problems involved in managing the consequences of organizational change processes triggered by changes in the law, for instance for business processes, services, databases, fielded applications, forms and documents, and internal education, make a good case for application of some state-of-the-art concepts in legal knowledge representation.
The recently started AGILE project addresses the legal dimension of management of organizational change processes. This paper introduces the AGILE project, and presents an initial overview of relevant relations between sources of law and the business processes and services of the administrative organization, based on concepts familiar in legal theory and legal knowledge representation. It also proposes the application of change-oriented features of the MetaLex XML standard to organizational change.

1 Introduction

Driven by the increasing legal convergence and legal complexity, an increasing pace of organizational change in public administration, and increased use of IT and web services, the interest in legal knowledge representation in public administrations is gradually increasing but also changing in nature.

Initially interest was focused on the utility of fielding computer systems built using a knowledge engineering approach; More recently the focus shifted to the potential utility of knowledge representation for comparative and for maintenance purposes, and for increasing the efficiency of the organizational change process itself.

Compared to the standards set by knowledge engineering research, fielded systems in public administration and elsewhere that use explicit knowledge representation to support decision making processes are technically and theoretically straightforward. The required transparency, and the great challenge real world knowledge representation poses for the people implementing such systems, act as a natural limit to the complexity of these decision support systems. The required functionality rarely by itself justifies state-of-the-art legal knowledge representation.

As we will argue in section 1.1, however, the knowledge management problems involved in managing the consequences of organizational change processes triggered by changes in the law, for business processes, services, databases, fielded applications, forms and documents, internal education, etc, make a considerably better case for some state-of-the-art concepts in legal knowledge representation.

The recently started AGILE project, introduced in section 2, addresses management of institutional change, sometimes driven by changing legislation and sometimes by environmental factors.

Section 3 presents an initial overview of relevant relations between sources of law and the business processes and services of the administrative organization, based on concepts familiar in legal theory and legal knowledge representation. This account is a first, tentative step towards a design method that should help organizations to adapt to new or changing legislation.

1.1 Background

Inside public administrations, and on the interfaces between them, ICT and Internet have a large impact. Some decision making processes are nowadays assisted by computer applications, and others are more or less autonomously performed by the computer.

At the same time, service-oriented architectures are becoming the prominent paradigm for building enterprise information systems, also in administrative agencies. Service-orientation leads to new network arrangements between administrative agencies for sharing data, etc. This development in itself leads to attention for the adaptability and accountability issues that arise (cf. [9]).

The services in question are in an administrative setting often implementations of public legal acts, performed by public legal personalities, based in formal legislation. Legislation gives administrative organizations public personality, defines what the core functions of public organizations are, and what services they provide. It guides how the organization subdivides itself into administrative units, how it organizes business processes inside the organization, and eventually how the functions of the organization are realized by civil servants and computer systems.

Business process design and design of specialized computer systems are both usually based on explicit *models* in various modeling languages of what the business process or computer application should achieve. These models are supposedly used as a specification of the objectives of an organizational change process or application development process. When legislation changes, these models are updated, and the organization's structures and computer programs have to be changed to conform to the models.

In the past these changes were conceived of as temporary interruptions of long periods of everything staying the same. This was certainly the case when the adaptation of existing systems was still considered a frightening prospect: things did change but the changes where carefully orchestrated to not impact existing procedures, network arrangements with other organizations, and computer systems. But as the perceived capacity of organizations to organize change processes increases, and the number of fielded computer applications increases, so does the pace of change in legislation directly affecting existing computer applications.

Tax legislation is for instance changed every year, leading to continuous adaptation of relevant computer applications for next year and the years after that, while the legislation of the present year and previous years is still being applied. In the business process design literature, awareness of this phenomenon has led to a new conception of the organization as an entity that is constantly in the process of changing: the organization is constantly conceptualizing and comparing what it is and what it is becoming.

Attention for *knowledge representation* of sources of law is very often triggered by such administrative change processes driven by new legislation and other sources of law, such as case law and internal written policies.

The change processes triggered by the legal system are increasingly expensive, especially if they involve changes in ICT infrastructure. Knowledge representation is seen as a means to potentially reduce costs and increase efficiency through increased control over the knowledge dimension of the change process. Our past work for the Dutch Tax and Customs Administration (DTCA; cf. for instance [4]) was for instance clearly related to the huge *change process* triggered by the complete overhaul of the Dutch income tax law in 2001. The *Juridisch Loket* (cf. [15]) project on *pro bono* legal assistance, and the *DURP* project on spatial planning (cf. [3]) were for instance also driven by an overhaul of legislation.

These trends have led to the AGILE project, described in the next section of this paper.

2 The AGILE Project

In the AGILE project (acronym for Advanced Governance of Information services through Legal Engineering) we aim at developing a design method, distributed service architecture and supporting tools that enable organizations – administrative and otherwise – to orchestrate their legal information services in a networked environment.

At issue is the adaptivity of ICT infrastructure, of business processes, and of data and knowledge within the organization, given changing legal demands and constraints.

The AGILE project started in the second half of 2008 and will last for four years. The project will use knowledge representation technology developed within the semantic web community, OWL DL, as a starting point, but will extend it where necessary with elements specific for the legal domain or the objectives of the project.

The primary purpose of modeling implementation of legislation in OWL is to account for that implementation, to validate it, and to simulate new service arrangements. Deployment of OWL-based services is not intended: actual technical implementation has to take into account the existing technical infrastructure of an organization, and modernizing infrastructure is not the focus of the project.

Complex Adaptive Systems Based on complex adaptive systems (CAS) theory, the project will develop a service modeling and design method that should help organizations to adapt to new or changing legislation. The essence of CAS theory is the study of systems built of individual agents (being persons, business units, or organizations) that are capable of adapting as they interact with each other and with an environment, in order to understand how the individual affects system-level responses.

The proposed method should take the resilience of existing systems, dependencies on the environment, and the unpredictability of change processes explicitly into account (cf. generally [9]). The objective of this part of the project is to improve the adaptability of ICT infrastructure, of business processes, and of knowledge in the organization.

The underlying premise is that simply determining future needs and requirements is not the right approach, due to the inherent unpredictability of a complex environment and the fact that there are already many working (social and information) systems in place which can not and should not be ignored.

The simulation architecture and tools are out of scope of this paper.

Pilot studies Results from the research tracks discussed will be tested in the context of two actual business cases. One at the Dutch Immigration and Naturalisation Service (IND) and one at the Dutch Tax and Customs Administration (DTCA).

In both organizations, timely and efficient adaptation to changing legislation, case law, and patterns of behaviour accommodating or evading law in the relevant environment is seen as an important organizational objective, whose realization is causing problems.

3 Legal Concepts in Agile

Of specific relevance to a world dominated by written declarations and decisions, databases, web services, and changing sources of law is an account of formal acts, and of the act of providing evidence for a legally relevant proposition. In this account the concepts developed in the MetaLex standardization effort, presented next, play an important role.

Formal legal acts are characterized by 1) the requirement that one intends to bring about a certain institutional change, and 2) that this intent is communicated in writing, i.e. the institutional change is *represented*. Both the act of legislating and the various paper or software-based acts of adminstrative organizations have this nature.

The relation between sources of law and the business processes and services of the administrative organization will be explained in terms of the institutionalization and formalization of normative order (cf. [10, 2]). The notion of services – which usually has no direct counterpart in the relevant sources of law – will be explained in terms of Hohfeld's directed jural relationships (cf. [8]). Both the constitutive rule (cf. generally [11]) and Hohfeld's categories (cf. for instance [12]) are mainstays in legal knowledge representation and legal philosophical logic.

MetaLex To implement traceability from knowledge representation to sources of law, the AGILE project will build on the results of our work on MetaLex XML (cf. for instance [6, 5, 2]).

MetaLex serves as a common document format, processing model, and metadata set for software development. In addition, the MetaLex CEN committee defines a single jurisdiction-neutral and transparent uniform resource identifier (URI) based open, persistent, globally unique, memorizable, meaningful, even to some extent "guessable" naming convention for legislative resources, that can be used productively in OWL modeling.

MetaLex and the MetaLex naming convention strictly distinguish the source of law as a published work from its set of expressions over time, and the expression from its various manifestations, and the various locatable items that exemplify these manifestations, as recommended by the Functional Requirements for Bibliographic Records (FRBR; cf. [13]). MetaLex extends the FRBR with a detailed but jurisdiction-independent model of the lifecycle of sources of law, that models the source of law as a succession of consolidated versions in force, and optionally *ex tunc* consolidations to capture the possibility of correction (errata corrige) or annulment after the fact of modifications by a constitutional court.

In the MetaLex metadata set, represented in an OWL ontology, the `realizes` property between expressions and works represents the connection between the two ontological levels at which documents exist that are of relevance to their real world

use. The source of law on the expression level for instance *cites* other rules on the work level, while the legal rules we represent are necessarily identified by their *representation* in expressions. A citation *(text fragment) w applies to (concept) C* should for instance be read as *each legal rule that is represented by an expression-level text fragment that* `realizes` *work fragment w applies to C*. This representation technique, an implementation of the idea of ontological stratification (cf. [7, 14, 2]), will play an important role in the AGILE project.

In current organizational practice the management of changing sources of law (particularly at the levels below *formal* law) is a notable weak point, and *ex tunc* change is often never heard of.

Institutions and Rules The primary purpose of legal knowledge representation for the administrative organization is to keep track of how it implements law in its organizational structure, business processes, data structures, business rules, and resource allocation practices. On the other hand the administrative organization has a number of good reasons to keep specifications relating to these ontologically distinct from their legal abstractions.

Firstly, there is often a mismatch between the conceptualization of the acts performed as found in the sources of law and the more practical conceptualization within the organization, even if there is a considerable overlap in terms. Law is not the only source of design requirements and constraints, and the implementation of well-defined decision making procedures and software support requires additional commitments.

Secondly, the concurrent use of different regimes within an organization, or of alternative procedures for performing the same legal act (for real or for simulation), make such an identification tricky. While the organization may for instance use the vocabulary of the law to structure it's data structures, it will inevitably be confronted with changes in that vocabulary, and the question which data can be *regrounded* in the new legal vocabulary.

Thirdly, straightforward legal rules are often in practice implemented as formal acts, creating a confusion between proposition and formal representations of such a proposition. For immigration, the proposition that someone is married may for instance be legally relevant; In the implementation of this criterion this however for instance becomes the proposition that someone has supplied a marriage certificate.

Since marriages may end, such a proposition is obviously not an essential quality of a person: the correct way to represent such a proposition is as a participation in a (time-limited) marriage. Moreover, if the organization for instance adds another condition that the marriage certificate must be less than one year old, and a procedure may take more than one year, a service requester may in fact be required to supply a marriage certificate multiple times. A certificate may in fact be still valid at the moment of decision making, while the marriage is not at the point in time in existence.

The administrative organization is conceived of as an implementation of a legal institution. Institutions can be conceptualized as abstract social systems with a well-defined interface with an environment. The structures of the legal institution are defined by the institutional facts that make up the institution, and its mechanisms of change are the constitutive rules – found in the relevant sources of law – that specify what brute act *constitutes*, or counts as, an institutional act. The administrative organization must at least *implement* each of the relevant institutional acts it is supposed to perform in brute reality in a business process and advertise it as a service to the relevant agents.

Conversely, it recognizes a limited number of ways in which agents in the environment can perform the acts that count as a request for the performance of a service.

Relevant patterns in logical propositions describing the functions of legal rules revolve around the notions of *constitutiveness* and *applicability*. The legal rules represented by the source of law appeal to two separate realities – institutional reality and brute reality – and perform a mapping from brute reality – the ontological substratum – into institutional reality – the ontological superstratum. The substratum has an existence independent of the rules, while the superstratum is supervenient on the substratum and exists by virtue of social recognition of the rules of the institution.

Applicability plays a central role as soon as the logical proposition and the legal rule are separated. The law frequently does so: A special class of legal rules, *applicability rules*, constrains the applicability of other rules, or make the application of one legal rule conditional on the application of another legal rule.

The institutional interpretation however tells us little about the functions of law for its users. To explain these functions, we have to appeal to planning and plan recognition. In the AGILE project this aspect is filled in with *agent simulation*.

In some cases such an explanation is straightforward. The analysis of normative rules in terms of normative positions and obligation, i.e. deontic logic, is such a straightforward abstract theory of behaviour, based on the expectation that people generally avoid the circumstances in which they are liable to be punished. To explain the normalizing effect of other rules one must ascribe intentions and preferences to agents: People generally intentionally try to bring about or avoid certain legal facts.

The principal aim of Hohfeld's work (in [8]) was to clarify *jural relationships* between parties. Hohfeld's relationships distinguish between the (legal) competence (or power, ability) and incompetence to play a certain agent role, and therefore to cause a certain change of position, and between the obligation to cause a certain change of position or the absence of such an obligation, and most importantly, between the one who acts and the one who predicts the actions of another. In essence we are dealing with the ability of one agent to infer:

1. that another agent has the *ability* or *inability* to change his (in this case legal) position in relevant ways, and
2. that the other agent has a *preference* for changing or not changing it.

Business process specifications represent an intention to use one's (legal) abilities in a predictable manner. Services publicly advertise this intention, so that it creates an ability (to change their legal position) of prospective clients. These clients use this ability by requesting a service. Of central importance is the adoption of agent roles: the client becomes a client by requesting a service and – thereby – adopting a well-defined role, while the employee of the administrative organization adopts an agent role in an associated business process. Agent simulation as a tool for impact analysis and exploration of design options assumes the development of prototypical agents representing both the organization itself and its relevant environment.

4 Discussion

The discussed elements are all found in legal knowledge representation. A deviation from mainstream legal knowledge representation is found in the rigorous ontological stratification (cf. [7, 14]) of legal entities and organizational entities we consider for

AGILE. Although legal knowledge representation literature discusses the "counts-as" or constitutive rules (cf. for instance [1, 11]), it usually considers them just one type of rule, among other (notably normative) ones, instead of applying the concept throughout.

Acknowledgements

AGILE is a Jacquard project funded by the Netherlands Organisation for Scientific Research (NWO). In the AGILE project, The Leibniz Center for Law of the University of Amsterdam cooperates with the Technical University of Delft, which has experience in the application of CAS theory to organizations. The IND and two companies, O&I and BeInformed, provide matching effort to the project.

References

1. G. Boella and L. W. N. van der Torre. Regulative and constitutive norms in normative multiagent systems. In *Proceedings of the 9th International Conference on the Principles of Knowledge Representation and Reasoning*, Whistler (CA), 2004.
2. A. Boer. *Legal Theory, Sources of Law, & the Semantic Web*. Frontiers in Artificial Intelligence and Applications 195. IOS Press, 2009. To appear.
3. A. Boer, T. van Engers, R. Peters, and R. Winkels. Separating law from geography in gis-based egovernment services. *Artificial Intelligence & Law*, 15(1):49–76, February 2007.
4. A. Boer, T. van Engers, and R. Winkels. Using Ontologies for Comparing and Harmonizing Legislation. In *Proceedings of the International Conference on Artificial Intelligence and Law (ICAIL)*, Edinburgh (UK), 2003. ACM Press.
5. A. Boer, F. Vitali, and E. de Maat. CEN Workshop Agreement on MetaLex XML, an open XML Interchange Format for Legal and Legislative Resources (CWA 15710). Technical report, European Committee for Standardization (CEN), 2006.
6. A. Boer, R. Winkels, and F. Vitali. Metalex XML and the Legal Knowledge Interchange Format. In G. Sartor, P. Casanovas, N. Casellas, and R. Rubino, editors, *Computational Models of the Law*, volume LNCS 4884 of *Lecture Notes in Artificial Intelligence*. Springer, 2008.
7. S. Borgo, N. Guarino, and C. Masolo. Stratified ontologies: The case of physical objects. In *Proceedings of the ECAI-96 Workshop on Ontological Engineering*, 1996.
8. W. Hohfeld. *Fundamental Legal Conceptions as Applied in Legal Reasoning*. Yale University Press, 1919. Edited by W.W. Cook, fourth printing, 1966.
9. M. Janssen. Adaptability and accountability of information architectures in interorganizational networks. In *ICEGOV '07: Proceedings of the 1st international conference on Theory and practice of electronic governance*, pages 57–64, New York, NY, USA, 2007. ACM.
10. N. MacCormick. Norms, institutions, and institutional facts. *Law and Philosophy*, 17(3):301–345, 1998.
11. T. Mazzarese. Towards the semantics of "constitutive" in judicial reasoning. *Ratio Iuris*, 12:252–262, 1999.
12. G. Sartor. Fundamental legal concepts: A formal and teleological characterisation. Technical report, European University Institute, Florence / Cirsfid, University of Bologna, 2006.

13. K. G. Saur. Functional requirements for bibliographic records. *UBCIM Publications - IFLA Section on Cataloguing*, 19, 1998.
14. B. Smith. An essay in formal ontology. *Grazer Philosophische Studien*, 6:39–62, 1978.
15. T. van Engers, R. Winkels, A. Boer, and E. de Maat. Knowledge management and the dutch legal aid service counter. In J. J. Schreinemakers and T. van Engers, editors, *Advances in Knowledge Management*, volume IV, Würzburg, 2006. Ergon Verlag.

Automatic Mark-up of Legislative Documents and its Application to Parallel Text Generation

Lorenzo Bacci[1], Pierluigi Spinosa[1], Carlo Marchetti[2,3], and Roberto Battistoni[3]

[1] Institute of Legal Information Theory and Techniques,
Florence, Italy
bacci@ittig.cnr.it
spinosa@ittig.cnr.it
[2] Dipartimento di Informatica e Sistemistica dell'Università di Roma "La Sapienza"
carlo.marchetti@senato.it
[3] Italian Senate, Rome, Italy
r.battistoni@senato.it

Abstract. In the juridical domain a huge amount of plain legislative acts have been produced since and before the advent of computers and word processors. The conversion of legacy and plain documents in a standard XML format implies great and numerous benefits. In order to accomplish this task, several automatic and semi-automatic tools have been developed in the last ten years. In this paper, xmLegesMarker, developed at Ittig[4], the Italian state-of-art legislative documents parser, is presented. The tool is NIR standard compliant, it's embedded in xmLegesEditor and it was recently adopted for evaluation by the Italian Senate in order to automatically spawn the parallel text (*testo a fronte*), i.e. the document used to highlight the modifications introduced during the debate of a bill.

1 Introduction

xmLegesMarker is a structure parser for legislative documents. Its development started in 2003, within Norme In Rete (NIR) project[1], in order to provide a detailed mark-up of legacy documents and, generally, plain legislative acts. Although it was realized as a stand-alone software, it has been mainly exploited inside xmLegesEditor[2], an XML based legislative editor arisen from the NIR project, to implement the import function, namely the migration of plain texts into the XML environment. The NIR project defined also appropriate DTD and XMLSchema, which represent the basis of both xmLegesEditor and xmLegesMarker and aim at describing in a very detailed way the Italian legislative acts. Thanks to its independent development and the ongoing improvements, xmLegesMarker has also been effectively adopted in the last few years by several public administrations and regional governments, usually to pursue a migration of their plain text databases of local, regional and national laws towards the NIR-XML mark-up.

Several benefits derive from this process: structured documents enhance information retrieval and normative system maintenance, and can represent the ground for a

[4] Institute of Legal Information Theory and Techniques (CNR)

further semantic description of text in terms of provisions [3], but, more important, they can also be exploited for legislative-domain strictly-related operations. As an example, TafWeb, described in section 5, is a smart legislative text comparison software, developed by the Computer Science Department of the University of Bologna under the supervision of the Italian Senate, which exploits the XML mark-up of the versions of a bill under debate in order to generate the parallel text document (*testo a fronte*). The Italian Senate is currently evaluating xmLeges-Marker within the TafWeb suite in order to automatize the production of the first version of a parallel text, starting from the plain Chamber and Senate versions of the bill.

2 Approaching plain documents

Automatically structuring a plain document means creating a software that, given the plain document as input, is able to assign an identity to every piece of text. In the XML world, this task is accomplished by putting the text between tags. In this case, the marker uses NIR defined tags in order to create well formed and valid, with respect to NIR schema, outputs.

Generally speaking, the information that can be obtained from a plain document consists of data and meta-data. In legislative acts, the enacting terms section, in which articles and paragraphs lie, matches the data, while entities like title, number and type of document, subscribers and so on, are considered explicit meta-data. Other meta-data, defined in NIR schema, although not explicitly present in the input, are computed or added by xmLegesMarker, usually exploiting the values of the explicit meta-data (i.e.: automatic generation of URN [4]). The splitting of information in data and meta-data follows the physical structure of the document. While explicit meta-data are typically located in the header or in the footer, the body of the legislative act accommodates the enacting terms, namely the data.

Besides the physical position, there is another important difference between the body and the header (or the footer) in a legislative act: the former is composed by partitions strictly organized and sorted in a hierarchical way, practically a tree of partitions, while the latter appears fuzzy and composed by expected and unexpected elements, often in a random order. This is the reason why xmLegesMarker adopts two different strategies in order to analyze the header, the footer and the body of a document.

2.1 Header and footer

The fuzziness that belongs to both the header and the footer of a legislative act requires a statistical and machine learning approach for meta-data extraction. In order to accomplish this, the theory of Hidden Markov Model (HMM) [5] was chosen and a model, able to understand the most typical information lying in the header and footer of a generic legislative act, was developed. [6]

An important source of information which was exploited to improve the accuracy of the header analysis is represented by the legislative document subtype (act, bill, decree, regional act, etc.). Depending on the subtype in fact, more precise patterns stand out. Therefore, xmLegesMarker applies a specific HMM model if the input subtype is known and supported, the generic model otherwise. Using subtype-crafted models brings two benefits:

– a better formalization of the most important subtype, reaching higher (far higher in some cases, like bills) degrees of accuracy;
– a guess about the subtype when the subtype information isn't given, just applying all the specific model and checking which one fits better.

2.2 Body

The body of a legislative act coincides with the enacting terms section, which is typically well organized in known partitions, hierarchically arranged in a tree structure. On the other hand, the tree representing the enacting terms can be complex, long and nested. An automata approach is required in order to efficiently parse this kind of structure. The Flex scanner generator[5], which has been used to accomplish several tasks for and besides body analysis, allows the creation of very powerful text scanner based on a non deterministic finite state automata (NFA). [6]

The automata that handles the enacting terms strictly follows the constraints imposed by the NIR schema: the parsing process obeys to rules that depend on the automata states (start conditions), which match all the partitions defined in the NIR hierarchy. For example, an alphabetical list can be read only if the automata is in the paragraph state, because, according to NIR and to legislative drafting rules, a list should only stay inside paragraphs.

2.3 Annexes

Let's conclude this survey on plain legislative documents by describing the way the marker handles annexes. In the legislative domain, annexes are very common: they belong to the legislative document itself, just following the, let's say, main act; they can be simple tables, reporting details, prices, fees, or be legislative documents themselves. Their importance sometimes exceeds even the importance of the main act: for example, there are bills containing just one or two paragraph, followed by a legislative decree containing dozens of articles.

The xmLegesMarker strategy comprises the preliminary segmentation of the input into main act and annexes, and the iteration of the header, body and footer parsing functions, described in the previous sections, on the main act as well on every single annex. In this way, possible legislative documents following the main act receive the same treatment and come out completely marked-up.

3 A glance in depth

In order to better understand the working, the capabilities and the potentiality of the marker, in this section the most interesting features and issues are deepened and discussed. After a brief overview about the shape of the input, we focus on how the marker handles partitions and amendments, which factors determine an increase of complexity during the body analysis process and how syntactical errors in the source have been successfully tackled.

[5] http://dinosaur.compilertools.net/flex/

3.1 Input

The input of xmLegesMarker is a plain legislative act in txt, doc, html or pdf format. The marker is able to manage different subtype of plain Italian legislative document, like act, bill, decree, regional act, etc. As discussed, different weights and models are used for the header and footer analysis depending on the subtype.

3.2 Handling partitions

Partitions are used to structure the fragments of the legislative document body. They are arranged in a hierarchical way. The paragraph is the partition that effectively contains the text of the law, while the greater order partitions (article, chapter, section, etc.) can be seen as containers. Even paragraphs can have sub-partitions: lists, which can be alphabetical, numerical or bulleted. Thanks to the use of regular expression, layout reasoning and the application of NIR constraints, as we have seen in 2.2, the marker is able to identify all of these partitions.

Furthermore, xmLegesMarker performs a check concerning partitions numbering, which turns out to be quite useful:

- as a mean to better identify the next partition, often avoiding ambiguity;
- to assign a unique value to the "id" attribute, defined by NIR, in order to permit referencing to every single partition.

3.3 Dealing with textual amendments

Amendments are a widely used mechanism to express modifications from a legislative document to another legislative document. The textual amendments can be briefly categorized in repeal, insertion and modification, and they can act on words as well as on whole partitions (structural amendment). Insertion and modification amendments are typically expressed using quotes. So, for example, it's possible to express through an amendment the substitution of a single noun or the insertion of a brand-new article in a precise position of a regional act.

The marker is able to identify and handle both words and structural amendments, and, in case of structural amendment, it enters the amendment and precisely mark-up all the partitions and data contained.

3.4 Increasing complexity

Even though the body of legislative documents generally follows the same syntactical rules, there are several variables that increase the complexity of the automata:

- paragraphs are sometimes not numbered, this happens especially with legacy documents;
- almost every partition can have a partition title, the *rubrica*, that sometimes is placed just after the declaration of the new partition, sometimes below, sometimes it's included between particular separator characters, sometimes not;
- within a paragraph there are three allowed kind of lists: alphabetical, numerical and bulleted list; however, good drafting rules state that bulleted lists should be avoided, because they can't be directly referenced (the relative position has to be specified), and only alphabetical lists should stay immediately inside paragraphs while numerical lists should only lie within alphabetical lists;

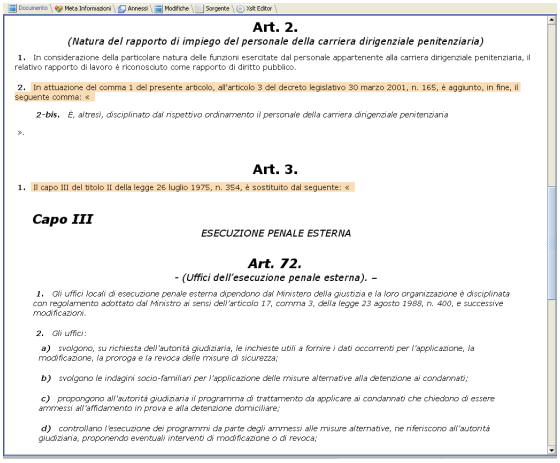


Fig. 1. A pretty nested example visualized in xmLegesEditor: the first paragraph of article 3 contains a chapter substitution amendment.

– parsing the text of amendments sometimes turns out to be a pretty hard task: between quotes there's a no man's land where most of the rules that usually guide the automata don't apply anymore, while every kind of partition is allowed there by NIR schema, thus xmLegesMarker sometimes has to deal with very complex cases (Fig. 1).

3.5 Tackling syntax errors

One of the main problems that have to be faced working with legacy documents and, generally, with documents edited manually, with no drafting support, is represented by the presence of syntactical errors. They can be:

– numbering errors;
– incorrect use of punctuation marks;
– errors in the layout of the document;
– unbalanced quotes;
– other drafting errors.

Some of these errors in the plain document have a limited effect in the XML output of xmLegesMarker, while others may cause totally disruptive behaviors of the automata used for the body parsing. For example, if quotes aren't balanced, the automata jams in the amendment states, forcing all the remaining text into the amendments tags. Another example, with not such a catastrophic effect, is represented by the erroneous or missing punctuation marks that should be used to separate different paragraphs inside an article; in that case, the next paragraph isn't acknowledged because of the erroneous separator, and all the remaining paragraphs in the article aren't acknowledged too, because of the checks on paragraph numbering, until the next article is read, which force a reset of the paragraph counter.

Thus, little errors in the input source often generate huge troubles in the resulting XML, while a little correction of the input saves the user from a painful correction of

the output in an XML editor. For this reason a messaging system which allows the user to operate directly in the source, correct it and process it again was implemented inside xmLegesMarker.

The new version of the marker is able to identify the most typical troublesome situations having reference to errors in the source, embedding a warning message in the output. The message is formatted as a processing instruction in the resulting XML, so it doesn't affect the validity of the document. Moreover, the message contains a warning code that refers to a warning table where typical troubles are classified, described and a guideline to solve each of them is provided.

4 Qualitative evaluation

The main automata, dedicated to the analysis of the enacting terms section (the body of the document), consist of 83 regular expressions, 22 states (or start conditions) and more than 70 rules based on start conditions and stacks of start conditions. It handles ten different kind of partitions (book, part, chapter, title, section, article, paragraph, alphabetical, numerical and bulleted list), and various other entities defined in the NIR schema, like partition title, decoration, amendment, and so on. The lex file that defines the Flex scanner comprises more than one thousand lines of code.

4.1 Shape of legislative acts

Legislative documents may be particularly complex from a structural point of view. Let's have a look at a couple of numerical example.

In Italy, the bigger legislative documents are probably the Budget laws. The bill 1183[6] of 2007, for example, representing the 2007 Budget bill, counts in the main act, including the amendments, 1122 paragraphs, 46 articles, 410 alphabetical list partitions and 75 numerical list partitions, altogether 1653 partitions!

The nesting of the body is variable as well, but isn't uncommon to run into bills with almost ten nested partitions. For example, the bill 3328[7] of 2005, has, take a deep breath, a four-points-long numerical list inside the letter "a" in paragraph "3" of article "165-ter" within section "VI-bis" inside an amendment in paragraph "1" within article "6" of chapter "III" in the title "I"!

4.2 On-road test

The Italian Senate provided a big data-set of bills on which several tests have been carried out, aiming at more and more refining the software. Given the fact that is a pretty hard task to conduct a precise statistical analysis about the accuracy on the whole data-set, because of the complexity of the input, in this section we try to give at least a qualitative idea about the capabilities of xmLegesMarker, just reporting the marking-up process outcomes of the two, quite representative, previously discussed bills.

Although quite complex, the mark-up of 3328 body was perfect and the marker didn't miss any partition. On the other hand, the same process on the huge 1183 triggered two warning messages:

[6] http://www.senato.it/leg/15/BGT/Schede/Ddliter/27212.htm
[7] http://www.senato.it/leg/14/BGT/Schede/Ddliter/22640.htm

Table 1. Bill 3328 mark-up details

Partition	Total	Amendment	Missing
Title	6	0	-
Chapter	10	1	-
Section	4	4	-
Article	81	39	-
Paragraph	249	155	-
Alphabetical	178	73	-
Numerical	50	8	-

Table 2. Bill 1183 mark-up details, before and after correction of the original input

	Before correction			After correction		
Partition	Total	Amendment	Missing	Total	Amendment	Missing
Article	35	17	11	46	28	-
Paragraph	1072	867	50	1122	152	-
Alphabetical	398	324	12	410	98	-
Numerical	66	59	9	66	28	9

– erroneous balancing of quotes in art. 18 paragraph 45;
– not numbered comma inside an amendment in art. 18.

As discussed, the first one is a disruptive problem, which effectively causes a pretty bad result. After correcting the two problems in the original plain input, the marking-up process was performed again and the outcome was excellent: the only imperfection is given by two non-standard numerical lists ("1.1", "1.2", etc.), a format not included in the rules for legislative drafting enacted by the Parliament and, consequently, neither included in the NIR drafting standards.

Tables 1 and 2 report, for each type of partition, the number of total occurrences found by the marker, how many of them are found in amendments and how many are missing.

5 An Italian Senate application

This section shows how the promising results of xmLegeMarker are exploited for evaluations within the TafWeb application, supported by the Italian Senate.

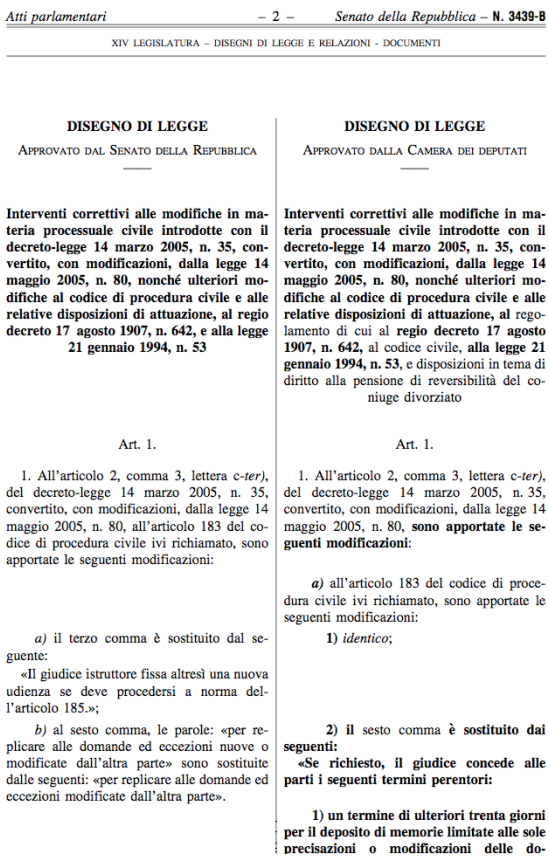
5.1 Scenario

The article by article discussion of a newly proposed bill is scheduled within the so-called "ordinary legislative procedure", in one of the two chambers of the Italian Parliament. During the discussion, amendments are voted and applied to the bill. Once the agreement is reached, the amended bill moves to the other Chamber, which is entitled to apply further modifications and send it back to the previous Chamber, until the bill

is applied in a Chamber without introducing new modifications, which terminates the process.

During the process, the effects of approved amendments, i.e., the differences between the two versions of a bill under debate, are represented using a TAF document, which stands for Testo A Fronte, a parallel text organized in two columns, with the original text on the left and the modified text on the right (Fig. 2). The TAF document is very useful for two reasons:

- it highlights the effects of amendments by using specific textual representations for each kind of modification, making it easier to understand where and how a bill has been modified;
- it can be used to limit the analysis and the discussion of the bill only to the changed parts.



Atti parlamentari – 2 – *Senato della Repubblica* – **N. 3439-B**

XIV LEGISLATURA – DISEGNI DI LEGGE E RELAZIONI - DOCUMENTI

DISEGNO DI LEGGE

APPROVATO DAL SENATO DELLA REPUBBLICA

Interventi correttivi alle modifiche in materia processuale civile introdotte con il decreto-legge 14 marzo 2005, n. 35, convertito, con modificazioni, dalla legge 14 maggio 2005, n. 80, nonché ulteriori modifiche al codice di procedura civile e alle relative disposizioni di attuazione, al regio decreto 17 agosto 1907, n. 642, e alla legge 21 gennaio 1994, n. 53

Art. 1.

1. All'articolo 2, comma 3, lettera c-*ter)*, del decreto-legge 14 marzo 2005, n. 35, convertito, con modificazioni, dalla legge 14 maggio 2005, n. 80, all'articolo 183 del codice di procedura civile ivi richiamato, sono apportate le seguenti modificazioni:

a) il terzo comma è sostituito dal seguente:

«Il giudice istruttore fissa altresì una nuova udienza se deve procedersi a norma dell'articolo 185.»;

b) al sesto comma, le parole: «per replicare alle domande ed eccezioni nuove o modificate dall'altra parte» sono sostituite dalle seguenti: «per replicare alle domande ed eccezioni modificate dall'altra parte».

DISEGNO DI LEGGE

APPROVATO DALLA CAMERA DEI DEPUTATI

Interventi correttivi alle modifiche in materia processuale civile introdotte con il decreto-legge 14 marzo 2005, n. 35, convertito, con modificazioni, dalla legge 14 maggio 2005, n. 80, nonché ulteriori modifiche al codice di procedura civile e alle relative disposizioni di attuazione, al regolamento di cui al **regio decreto 17 agosto 1907, n. 642,** al codice civile, **alla legge 21 gennaio 1994, n. 53**, e disposizioni in tema di diritto alla pensione di reversibilità del coniuge divorziato

Art. 1.

1. All'articolo 2, comma 3, lettera c-*ter)*, del decreto-legge 14 marzo 2005, n. 35, convertito, con modificazioni, dalla legge 14 maggio 2005, n. 80, **sono apportate le seguenti modificazioni**:

a) all'articolo 183 del codice di procedura civile ivi richiamato, sono apportate le seguenti modificazioni:

1) *identico*;

2) il sesto comma **è sostituito dai seguenti:**

«Se richiesto, il giudice concede alle parti i seguenti termini perentori:

1) un termine di ulteriori trenta giorni per il deposito di memorie limitate alle sole precisazioni o modificazioni delle do-

Fig. 2. An excerpt of a TAF document

5.2 TafWeb

TafWeb[8] is an experimental web service developed by the Department of Computer Science of the University of Bologna under the supervision of the Italian Senate, which aims at automatically spawning the first version of a TAF document, in order to reduce the amount of work done for producing these documents from scratch. The overall system is currently under development for evaluation purposes. The core of TafWeb is represented by JNdiff[9], arisen from Ndiff [7][8], a highly configurable algorithm for smartly comparing XML documents.

Thanks to the integration of xmLegesMarker inside the TafWeb environment, it's possible to implement a service which is able, starting from the plain Chamber and Senate version of a bill, to automatically produce a document representing the TAF. The main steps involved are:

- the conversion of the plain Senate and Chamber version of a bill in XML through xmLegesMarker;
- the computation of the "difference document", through JNdiff;
- the application of a style sheet to the original and to the difference document, generating the TAF in the official formats: XHTML, Office Open XML, PDF.

6 Conclusions

In this paper we presented xmLegesMarker, a powerful parser for Italian legislative documents. Its main features and capabilities were deeply analyzed and a qualitative evaluation concerning the marking-up accuracy on plain bills was provided.

Besides the benefits in the information retrieval field, the conversion of a legislative corpus into the XML language permits the development of useful, legislative domain related software. The scientific collaboration between Italian Senate and Ittig, currently focusing on the promising outcomes of the marker, enabled the realization of a process to automatically produce the first version of a parallel text from the plain Chamber and Senate versions of a bill under debate. Within the TafWeb environment, JNDiff, a comparison algorithm for XML documents, exploits the very detailed mark-up generated by xmLegesMarker in order to capture amendments and structural differences.

The automatic production of the parallel text, a process manually performed to date, reduces the burden of communication between the two chambers of the Italian Parliament and speeds up the legislative amending process.

7 Acknowledgements

This work is supported by a grant of the Office for the development of the information systems of the Italian Senate.

[8] http://sourceforge.net/projects/tafweb/
[9] http://sourceforge.net/projects/jndiff/

References

1. Biagioli, C., Francesconi, E., Spinosa, P., Taddei, M.: The NIR project: Standards and tools for legislative drafting and legal document web publication. In Proceedings of ICAIL Workshop on e-Government: Modeling Norms and Concepts as Key Issues, pp. 69-78 (2003).
2. Agnoloni, T., Francesconi, E., Spinosa, P.: xmLegesEditor: an opensource visual XML editor for supporting legal national standards. In Proceedings of the V Legislative XML Workshop, European Press Academic Publishing, pp. 239-251 (2007).
3. Biagioli, C.: Ipotesi di modello descrittivo del testo legislativo per l'accesso in rete a informazioni giuridiche. Informatica e Diritto 2:90 (2000).
4. Spinosa, P.: Identification of legal documents through URNs (uniform resource names). In Proceedings of the EuroWeb 2001, The Web in Public Administration (2001).
5. Rabiner, L.R.: A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. Proceedings of the IEEE 77 (2): pp. 81-106 (1989).
6. Francesconi, E.: The "Norme in Rete"- project: Standards and tools for Italian legislation. International Journal of Legal Information, vol. 34, no. 2, pp. 358-376 (2006);
7. Schirinzi, M., Vitali, F., Di Iorio, A.: Ndiff, un approccio naturale al confronto di documenti XML (2007).
8. Di Iorio, A., Marchetti, C., Schirinzi, M., Vitali, F.: A Natural and Multi-layered Approach to Detect Changes in Tree-Based Textual Documents. (to appear) Proceedings of the 11th International Conference on Enterprise Information Systems (ICEIS), (2009).

Text-based Legal Ontology Enrichment

Wim Peters

Department of Computer Science, University of Sheffield, U.K.
w.peters@dcs.shef.ac.uk

Abstract. The acquisition of knowledge from text is an incomplete and incremental process. When anchored to a particular knowledge model it provides potentially useful information to the legal expert in the form of new concepts and relations, in order to improve the domain coverage. This paper explores the feasibility of various legal text-based ontology enrichment techniques, and discusses the transformation of lexical knowledge to an ontological structure.

1 Introduction

Ontology generation and population is a crucial part of knowledge base construction and maintenance that enables us to relate text to ontologies, providing on the one hand a customised ontology related to the data and domain with which we are concerned, and on the other hand a richer ontology, which can be used for a variety of semantic web-related tasks such as knowledge management, information retrieval and question answering.

Ontologies cover a particular knowledge domain in various levels of adequacy. Lacunae in domain coverage, different tasks or changes in the conceptualization require modifications of the ontology [1]. Ontology enrichment is a necessary ingredient of this ontology life cycle.

One source for enrichment of legal ontologies is the analysis of legal texts. It can generally be stated that law depends on language: regulatory knowledge must be communicated, and the written and oral transmission of social or legal rules passes through verbal expression. Therefore legal conceptual knowledge is closely related to language use within the legal domain. Legal discourse can never escape its own textuality [2], which implies that linguistic information plays an important role in its definition. In our work, we base ourselves on the postulation that there is, as in other terminological domains, a relatively high level of dependence between legal concepts and their linguistic realization in the various forms of legal language [3].

The acquisition of knowledge from resources such as texts is an incomplete and incremental process. Knowledge is quite often left implicit in text, or depends on previous analysis steps. This causes a sparseness problem for automatic acquisition.

In our work we attempt to alleviate this problem firstly by bootstrapping and constraining the acquisition process on the basis of an existing legal ontology, which provides a solid conceptual framework. Secondly, perfect automatic knowledge acquisition does not exist. The acquisition results are considered informal suggestions that need expert evaluation and formalization into an enriched ontological structure as concepts and properties. These suggestions are necessarily partial and incremental. Their fragmented nature shows them as building blocks which, under expert supervision and

according to an existing knowledge structure, enables the building and addition of knowledge in a bottom-up fashion.

This paper investigates the (semi-)automatic enrichment of a legal ontology by means of a selection of NLP techniques based on pattern matching and statistical analysis. It is exploratory in character and therefore its methodologies are only indicative of the potential of the applied techniques.

The main task we set ourselves is the investigation into the feasibility of ontology enrichment techniques. This ontology enrichment can take two forms. On the one hand, new relations between existing ontology elements may emerge from textual data. On the other, new candidate concepts with new relations with existing ontology elements may be suggested by an integrated linguistic and statistical text analysis.

Recently, many relation extraction approaches have been proposed focusing on the particular task of ontology development (learning, extension, population). These approaches aim to learn taxonomic or non-taxonomic relations between concepts, instead of lexical items. Therefore, the list of techniques applied in this paper is not exhaustive. It forms a subset of the full set of methodologies available.

Most techniques described in this paper rely on robust and adaptable tools from the GATE architecture [4]. GATE is a framework for language engineering applications, which supports efficient and robust text processing. GATE uses NLP based techniques to assist the knowledge acquisition process for ontological domain modelling, applying automated linguistic analysis to create ontological knowledge from textual resources, or to assist ontology engineers and domain experts by means of semi-automatic techniques.

Our hypothesis is that the integration of corpus material, knowledge-based techniques and the use of rich linguistic processing strategies, can achieve effective results by accurately acquiring relevant relational knowledge [5]. A variety of techniques is helpful to the expert ontology engineer to extend the domain coverage of an existing ontology.

2 The Dalos Ontology

The DALOS domain ontology[1] [10] aims to describe the domain of the consumer protection, which has been chosen as the pilot case in the recently finished DALOS project[2], which resulted in the provision of support for the legal drafting process. It has been implemented as an extension of the Core Legal Ontology (CLO)[3] developed on top of DOLCE foundational ontology [11] and on the "Descriptions and Situations" (DnS) ontology [12] within the DOLCE+ library[4]. The extension covers the entities of the chosen domain and their legal specificities. In this network of ontologies the role of a core legal ontology is to describe concepts, which belong to the general theory of law, bridging the gap between domain-specific concepts and the abstract categories of formal upper level or foundational ontologies from DOLCE.

The domain ontology is populated by the conceptual entities which characterize the consumer protection domain. Such domain-specific concepts are classified according to more general notions, imported from CLO, as Legal role and Legal situation. Examples of consumer law concepts are CommercialTransaction, Consumer, Supplier, Good and Price. The first version of the DALOS Ontological layer contains 121 named classes.

[1] http://turing.ittig.cnr.it/jwn/ontologies/consumer-law.owl.
[2] http://www.dalosproject.eu/.
[3] http://www.loa-cnr.it/ontologies/CLO/CoreLegal.owl.
[4] http://dolce.semanticweb.org.

3 Ontology Enrichment

The DALOS ontology is the result of a manual effort within the DALOS project. Ontological modelling of legal domains is a constant effort. Domain descriptions need to be refined. Legislation evolves in the sense that new directives are issued, and old ones are deprecated. Therefore its coverage of the domain of consumer protection in terms of ontological vocabulary is never complete, and should be constantly adapted on the basis of expert advice and data-driven suggestions. Its incorporation of top level ontologies such as DOLCE make it descriptively adequate and robust for the higher levels of ontological legal description, but in terms of fine-grained domain-specific vocabulary it continuously remains in need of refinement and extension.

Our aim is to provide data-driven suggestions for ontology extension in the form of lexical material from the English legal texts in the DALOS corpus, which consists of directives and judgements (270,000 words in 55 directives and judgements). The results carry no more authority than suggestions for expert evaluation. For our analyses described below, the evaluator is a computational linguist, not a legal expert.

The main task these analyses perform is the general knowledge based identification of text-derived information that is of possible interest for legal ontology enrichment. Legal relevance will be an additional evaluation phase in which the data, deemed relevant from a general perspective, are assessed by an expert, and, if deemed relevant for the legal knowledge expressed by the DALOS ontology, integrated into an extended knowledge structure.

4 Acquisition from Text

The idea of acquiring semantic information from texts dates back to the early 1960s with Harris' distributional hypothesis [7] and Hirschman and Sager's work in the 1970s [8], which focused on determining sets of sublanguage-specific word classes using syntactic patterns from domain-specific corpora. Many techniques have since been proposed for the task of extracting knowledge from texts. Overall, the majority of approaches can be divided into pattern-based (pattern matching in a corpus) and statistically-based extraction [21]. Quite often, the two techniques are mixed (e.g. [24], [25], [26]). A description of several other approaches for conceptual relation extraction aiming at ontology learning can be found in [9].

4.1 GATE

The GATE platform[5] forms the methodological basis for our work [4]. A number of tools have been developed and used for the task of legal ontology enrichment. They all rely on the initial stage of linguistic pre-processing the corpus under examination, in order to obtain valuable linguistic information that will be used in later processing.

4.2 Pre-processing

First, tokenization and sentence splitting divide up the text into manageable units. Then part of speech tagging and lemmatization allow the inclusion of morpho-syntax into the analysis.

[5] `http://www.gate.ac.uk`.

4.3 Term extraction

The extraction tool TermRaider produces term candidates from a corpus by first filtering out possible terms by means of a multi word unit grammar that defines the sequences of part of speech tags constituting noun phrases. The computation of term frequency/inverted document frequency (TF/IDF) [13] [20], a technique widely used in information retrieval and text mining taking into account term frequency and the number of documents in the collection, yields a score that indicates the salience of term candidates for each document in the corpus. All term candidates with a TF/IDF score higher than an empirically determined threshold are then selected.

4.4 Lexico-syntactic pattern matching

Lexico-syntactic patterns are textual patterns that, with morphosyntactic normalization such as lemmatization, are highly indicative of semantic relations between textual elements. Ontology population based on this pattern approach has proven to be reasonably successful for a variety of tasks [6].

The following pattern matching strategies have been applied:

a) Headword matching This technique looks for a match between a pair of elements, of which one is embedded into the other as the head of a syntactic construction. The ontological interpretation of this relation is the insertion of a hyponymic relation between these elements. Examples from the Dalos ontology are:

Contract SuperClassOff DistanceContract Activity SuperClassOff CommercialActivity

Ten head matching relations were found in the ontology. All ten are covered in the ontology by means of superclass relations, except for one: Agent isSuperClassOff PhysicalAgent. PhysicalAgent is an object, and Agent is a top concept. PhysicalAgent is a hypernym of NaturalPerson, and the definition of Agent is: "A natural or legal person which plays the role of legal subject". We can therefore conclude on the basis of this definition that this additional a subsumption relation holds.

Matching the term candidates identified by TermRaider with existing classes resulted in 378 matching pairs. Manual evaluation of this set showed that 115 (around 30%) of them should be considered by experts for possible inclusion into the DALOS ontology. As an illustration, the following candidate subclasses of Contract were extracted, which show the detail of terminological specification in this domain: timeshare contract; purchase contract; credit contract; package travel contract; consumer contract; building contract.

b) Hearst patterns The second acquisition technique is based on Hearst patterns [14], which are a set of lexico-syntactic patterns that indicate hyponymic relations, and have been widely used by other researchers. Typically, they achieve a very high level of precision, but quite low recall [21]: in other words, they are very accurate but only cover a small subset of the possible patterns for finding hyponyms and hypernyms. The patterns can be described by the following rules, where NP stands for a Noun Phrase and the regular expression symbols have their usual meanings[6]:

[6] () for grouping; |for disjunction; *, +, and ? for iteration

```
{ NP such as (NP,)* (or|and) NP
```

Example: "advertising and marketing practises, such as product placement, brand differentiation or the offering of incentives. . . "

```
{ NP (,NP)* (,)? (or|and) (other|another) NP
```

Example: ". . . whereby a creditor grants or promises to grant to a consumer a credit in the form of a deferred payment, a loan or other similar financial accommodation."

No matching patterns between Dalos ontology elements were found. Table 1 below lists the results for obtained patterns between term candidates selected by TermRaider and Dalos ontology elements. The success rate is lower than expected (27% on average), given the reported high precision of Hearst patterns.

Table 1. Results from Hearst pattern matching

Hearst Pattern	Number found	Valid	Success Rate
Such as	31	11	32%
Including	0	0	-
And other	0	0	-
Or other	2	1	50%
Especially	1	0	0%

c) Mutual Information Whereas both a) and b) produce paradigmatic (isa) relations between terms, pointwise mutual information[7] (MI) is a well-known technique that measures the mutual dependence of the two variables as an expression of a syntagmatic relation. It is commonly used as a significance function for the computation of collocations in corpus linguistics [15]. In our case, it measures the statistically-based strength of relatedness through collocation within the same document.

Overall, forty MI relations were found between existing concepts from the Dalos ontology after matching DALOS ontology labels onto textual elements. Nine (22.5%) of the forty are not connected by any relation or concatenation of relations in the ontology. For example, the following pairs with their MI value:

ConsumerGoods	ConsumerProtection	4.10
ConsumerProtection	Consumer	3.37
FinancialService	Supplier	2.60
Producer	RawMaterial	2.55
Seller	ConsumerGoods	2.45
ConsumerGoods	Producer	2.41
ImmovableProperty	Contract	1.54
ImmovableProperty	FinancialService	1.21
FinancialService	Product	1.20

[7] See `http://www.collocations.de/` and `http://en.wikipedia.org/wiki/Mutual_information`.

Thirty one (77.5%) are related within the ontology, expressed by property concatenations in varying degrees of complexity.

Six MI pairs have a direct connection between its members, as illustrated below:

Advertising	subClassOf	CommercialCommunication	4.60
Consumer	isConsumerRoleOf	NaturalPerson	3.90
NaturalPerson	hasRole	Supplier	2.72
NaturalPerson	hasSellerRole	Traider	2.69
Advertising	isAbout	Product	2.34
CreditAgreement	hasParticipant	Consumer	2.25

A number of concepts (Consumer, Supplier, Trader, Producer, Organizer and Seller) are all subconcepts of LegalRole in the DALOS ontology. As co-hyponyms they are not directly related, but indirectly through their hypernym. The 11 MI pairs in which they are collocations seem to express ontological relations that are applicable to this whole set of co-hyponyms, in varying property configurations, such as Contract and CreditAgreement, of which Contract is the strongest indicator.

Supplier	Seller	5.91
Contract	Organizer	4.53
Consumer	Supplier	3.80
Consumer	Seller	3.48
CreditAgreement	Supplier	3.34
Supplier	Contract	3.21
Seller	Contract	2.70
DistanceContract	Consumer	2.61
Trader	Consumer	2.56
Contract	Consumer	2.46
Contract	FinancialService	2.03
Supplier	Producer	2.00

The remaining fourteen of the MI concept pairs have complex indirect links between them, which consist of a concatenation of object properties. For example:

Producer Product 4.21
Product isObjectOf Advertising Isactedin CommercialTransaction
hasParticipant Agent hasRole Producer

Consumer CommercialCommunication 3.37
CommercialCommunication isActedIn CommercialTransaction
Participant Consumer

Consumer GeographicalAddress 2.45
GeographicalAddress isQualityOf NaturalPerson
hasConsumerRole Consumer

Product Consumer 0.21
Product isObjectOf Advertising Isactedin CommercialTransaction
hasParticipant Agent hasRole Consumer

These results indicate the potential for statistical techniques - in this case the computation of mutual information values for pairs of ontology members- for the identification of fine-grained relations between concepts. 77.5% of the extracted MI relations are already attested in the ontology. The 22.5% of the MI pairs without ontological confirmation make ontological sense to the inexpert eye in that they express fine-grained relations that should be expertly evaluated for inclusion into the ontology, and linked to existing ontology elements by means of existing or new object properties.

The value of the MI score does not seem to matter much in terms of validity of a relation between the ontology elements, nor does it seem indicative of the length of the path between the ontology elements. The actual detection of a relation by means of MI computation seems to be crucial in this case, and it is up to experts to determine the granularity of the property vocabulary in the ontology, and decide whether this relation needs to be made explicit by means of one object property, or a concatenation of object properties.

d) Verbal complementation patterns Verbal patterns typically reflect lexicalized semantic relations between its arguments. Patterns defined in GATE can consist of any type of annotation that has been added in GATE, e.g. part of speech, string value, lemma etc. The corpus indexing and querying tool in GATE, called ANNIC[8] (ANNotations In Context) [16], allows the evaluator to enter search patterns over text annotations, and detect semantic relations between ontology elements at the fine-grained text level.

As proof of concept, the following simple pattern was defined, which identifies pairs of elements from the Dalos ontology that are mentioned in the texts as verb arguments. The surface representation restricts the verb context to a two-token window on either side.

```
{DalosConcept}({Token})*2{Token.category=="VERB"}({Token})*2

{DalosConcept}
```

A graphical user interface allows the user to query a corpus and inspect the results from the query. The screenshot in Figure 1 below illustrates how the results are displayed in the GATE interface. Annotations over spans of text are displayed as rows with coloured blocks indicating part of speech, string and DalosConcept. Contexts to the left and right of the text matching the search pattern are displayed at the bottom.

Using this query, 56 patterns were extracted, of which 37 (66%) were evaluated as deserving expert attention. For example:

NaturalPerson	conclude	Contract	with Seller or Supplier
NaturalPerson	buy	Product	
Seller/Supplier	dissolve	Contract	
Consumer	enter into	CreditAgreement	
Consumer	purchase	Product	

[8] `http://videolectures.net/gate06_aswany_ac/`.

Consumer	rely on	Guarantee
Consumer	acquire	Services
CompetentAuthority	assess	Product

5 Formalization of acquired lexical knowledge

Surface patterns and text spans are potential lexical realizations of underlying ontological relations and concepts. Ontologies themselves are conceptual constructs without linguistics. From a formal ontological point of view, concepts are abstract notions whose labels (often constituted by textual elements) are arbitrary. The lexical senses of the lexicalizations that function as labels for these concepts, are only considered to be evocative or indicative of the ontological meaning of the concepts. There is an implicit mapping assumption between lexical and conceptual knowledge, which underlies "ontology lexicalization", namely that (intensional) senses from a lexical model are mapped to (extensional) interpretations on ontology elements (individuals, classes, restrictions, properties) [17].

The reification of lexical material into ontological elements can happen in various ways. Some authors state that there is a direct relation between lexical form and surface syntactic pattern and ontological content [18]. Others advocate a formalization process that transforms surface patterns into ontology concepts and object properties in a number of stages, maintaining the philosophical distinction between lexical meaning and conceptualization, and allowing predication over these various levels of semantic representation [19].

The first stage is a transformation of linguistic elements (abstracted away from surface forms by means of lemmatization and other linguistic normalization processes such as morphological decomposition) into a semantic metamodel, which expresses the semantics of the domain. The next step, the transformation of this semantic domain knowledge into an ontological representation language construct such as OWL[9], decides on the ontological status of the semantic knowledge, e.g. whether it should be encoded as class, an attribute of an object property.

6 Discussion

When applying a variety of NLP techniques for ontology extension, each technique provides its specific spectrum of potential ontological enrichment based on the nature of the linguistic and statistical algorithms involved. Overall, the four acquisition techniques described in this paper (head matching, Hearst patterns, mutual information and simple verb complementation patterns) form a representative combination of acquisition techniques for both paradigmatic and syntagmatic lexical semantic relations. They perform reasonably well for establishing relations between ontology elements (81.2% average success rate excluding Hearst patterns, for which no hits were found). Since Hearst patterns are very sparse at best, future work on text-based ontological relation acquisition will look at the extension of the Hearst pattern set with more textual patterns reflecting the paradigmatic isa-relation (taken from e.g. [27]).

[9] `http://www.w3.org/2004/OWL`.

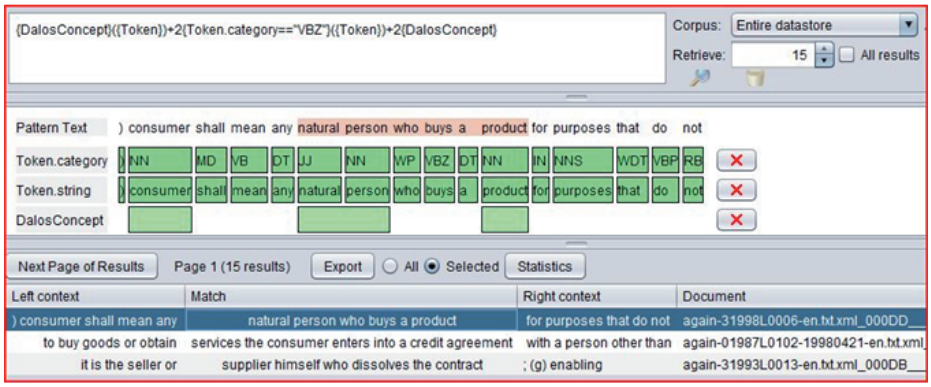


Fig. 1. Snapshot of ANNIC functionality

Head matching and Hearst patterns between term candidates and ontology elements have an average success rate of 28.5%, which is lower than expected. Overall, we can conclude that the techniques work well for identifying relations between ontology elements.

The reification of these surface syntactic and collocational relations may take several forms, depending on the strategy chosen. For some of the extracted relations based on verbal and deverbal lexicalizations, the proposed corresponding ontological relations are not always disjoint. For instance, in a number of cases, it is possible to group certain relations together under synonymy. As an example, the textual fragments "supply of services" and "provision of services" contain deverbal nouns, which, when translated into verbal counterparts, yield the following object properties:

AGENT supply SERVICE AGENT provide SERVICE

Since "supply" and "provide" are synonyms in WordNet [22], the object can be renamed into a common label, which covers both verbal lexicalizations. Further mapping with lexical resources such as VerbNet [23] will further classify the relations into more general classes, and provide semantic role arguments (e.g. agent, instrument etc.). Together with further analysis of the lexicalizations that instantiate these patterns, this will lead to an incremental creation of semantic frames, which then can be transformed into their ontological counterparts with ontologically proper constraints on the domain and range of the reified properties.

References

1. Stojanovic, L., Maedche, A., Motik, B. and Stojanovic, N.: User-driven ontology evolution management. In: European Conference of Knowledge Engineering and Management, EKAW (2002), pp. 285–300, Springer, placeCityHeidelberg (2002).
2. Macdonald, D., Legal bilingualism, McGill Law Journal 1997,42 McGill L.J. 119 (1997).
3. Peters, W,, Tiscornia, D, Sagri, M.T., The Structuring of Legal Knowledge in Lois. In: Artificial Intelligence and Law, Volume 15, Issue 2, Legal knowledge extraction and searching & legal ontology applications, pp. 117-135 (2007).

4. H. Cunningham, H., Maynard, D., Bontcheva, K. and Tablan, V.: Gate: A Framework and Graphical Development Environment for Robust Nlp Tools and Applications. In: Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL'02) (2002).
5. Specia, L., Baldassarre, C., Motta, E., Relation Extraction for Semantic Intranet Annotations, Technical Report kmi-06-17, Open University, Milton keynes (2006).
6. Etzioni, O., Cafarella, M., placeCityDowney, D., Kok, S., Popescu, A., Shaked,T., Soderland, S., Weld, D.S., Yates, A.: Web-scale Information Extraction in KnowItAll. In: Proceedings of WWW-2004 (2004).
7. Z. Harris, Z., Mathematical Structures of Language, Wiley (1968).
8. Hirschman, L., Grishman, R., Sager, N.: Grammatically based automatic word class formation. Information Processing and Retrieval 11 (1975)
9. Gómez-Pérez, A., Manzano-Macho, D., A survey of ontology learning methods and techniques. In: The Knowledge Engineering Review archive, Volume 19, Issue 3,pp. 187-212 (2004).
10. Agnoloni, T., Bacci, L. and Francesconi, E., Ontology Based Legislative Drafting: Design and Implementation of a Multilingual Knowledge Resource..In: Gangemi, A. and Euzenat, J. (Eds.) Knowledge Engineering: Practice and Patterns Knowledge Engineering: Practice and Patterns, 16th International Conference, EKAW 2008, Acitrezza, Italy, September 29 - October 2, 2008, Springer Lecture Notes in Computer Science, Volume 5268/2008 (2008).
11. Gangemi, A., Sagri, M. and Tiscornia, D., A constructive framework for legal ontologies. In: Law and the Semantic Web (Benjamins, R., Casanovas, P, Gangemi, A. and Selic, B. (eds.), Springer Verlag, (2005).
12. C. Masolo, C. Vieu, L. Bottazzi, E. Catenacci, C. Ferrario, R. Gangemi, A. and Guarino, N., Social roles and their descriptions, In: Welty, C. (ed.), Proceedings of the Ninth International Conference on the Principles of Knowledge Representation and Reasoning, Whistler, (2004).
13. Buckley, C. and Salton, G., Term-weighting approaches in automatic text retrieval. In: Information Processing and Management, vol. 24, no. 5, pp. 513-523 (1988).
14. Hearst, M., Automatic acquisition of hyponyms from large text corpora. In: Proceedings of the Fourteenth International Conference on Computational Linguistics, placeCityNantes, country-regionFrance (1992)
15. Smadja F. A & McKeown, K. R. , Automatically extracting and representing collocations for language generation. In: Proceedings of ACL'90, pp. 252–259, placeCityPittsburgh, StatePennsylvania (1990).
16. Aswani, N., Tablan, V., Bontcheva, K. and Cunningham,H., Indexing and Querying Linguistic Metadata and Document Content. In: Proceedings of 5th International Conference on Recent Advances in Natural Language Processing, placeCityBorovets, country-regionBulgaria (2005).
17. Peters, W., Montiel-Ponsoda, E., Aguado de Cea, G. and Gómez-Pérez, A., Localizing Ontologies in Owl. In: Proceedings of the ISWC 2007 workshop Ontolex2007, placeCityBusan, country-regionKorea (2007).
18. Cimiano, P., Haase, P., Herold, M., Mantel, M., and Buitelaar, P., LexOnto: A Model for Ontology lexicons for Ontology-based NLP. In: Proceedings of the ISWC 2007 workshop Ontolex2007, placeCityBusan, country-regionKorea (2007).
19. Picca, D., Gangemi, A. and Gliozzo, A., LMM: an OWL Metamodel to Represent Heterogeneous Lexical Resources. In: Proceedings of the International Conference on Language Resources and Evaluation (LREC), placeCityMarrakech, country-regionMorocco (2008).

20. Maynard, D., Li, Y.and Peters, W., Nlp techniques for term extraction and ontology population. In: Buitelaar, P. and Cimiano, P. (eds.), Ontology Learning and Population: Bridging the Gap between Text and Knowledge, pp. 171-199, IOS Press, placeCityAmsterdam (2008).
21. Cimiano, P., Pivk, A., Schmidt-Thieme, L. and Staab, S. Learning taxonomic relations from heterogeneous sources. In: Proceedings of the ECAI 2004 Ontology Learning and Population Workshop (2004).
22. Fellbaum, Christiane (ed.), WordNet. An Electronic Lexical Database. placeCity-Cambridge, StateMass.: MIT Press. (1998).
23. Kipper, K., Korhonen, A., Ryant, N. and Palmer, M., Extensive Classifications of English verbs. Proceedings of the 12th EURALEX International Congress, placeCity-Turin, country-regionItaly (2006).
24. Pantel, P. and Pennacchiotti, M., Automatically harvesting and ontologizing semantic relations. In Paul Buitelaar and Philipp Cimiano (eds.), Ontology Learning and Population: Bridging the Gap between Text and Knowledge, pp.171-195, IOS Press, placeCityAmsterdam (2008).
25. Cimiano, P., Hartung, M., and Ratsch, E., Finding the Appropriate Generalization Level for Binary Relations Extracted from the Genia Corpus. Proceedings of the International Conference on Language Resources and Evaluation (LREC), placeCityGenoa, country-regionItaly (2006).
26. Schutz, A. and Buitelaar, P., RelExt: A Tool for Relation Extraction from Text in Ontology Extension. International Semantic Web Conference 2005:593-606 (2005).
27. de Cea, G.A., Gómez-Pérez, A., Ponsoda, E.M., Suarez-Figueroa, M.C.: Natural language-based approach for helping in the reuse of ontology design patterns. In: Proceedings of the 16th International Conference on Knowledge Engineering and Knowledge Management Knowledge Patterns (EKAW 2008), placeCityAcitrezza, country-regionItaly (2008).

Towards a FrameNet Resource for the Legal Domain

Giulia Venturi[1], Alessandro Lenci[1],
Simonetta Montemagni[1], Eva Maria Vecchi[1],
Maria Teresa Sagri[2], Daniela Tiscornia[2], and Tommaso Agnoloni[2]

[1]Institute of Computational Linguistics, CNR, Pisa, Italy
[2]Institute of Legal Information Theory and Techniques, CNR, Firenze, Italy
{giulia.venturi,alessandro.lenci,
simonetta.montemagni,evamaria.vecchi}@ilc.cnr.it
{mt.sagri,daniela.tiscornia,tommaso.agnoloni}@ittig.cnr.it

Abstract. In the AI&Law community, the importance of frame–based ontologies has been acknowledged since the early 90's with the Van Kralingen's proposal of a *frame language* for legal knowledge representation. This still appears to be a strongly felt need within the community. In this paper, we propose to face this need by developing a FrameNet resource for the legal domain based on Fillmore's *Frame Semantics*, whose final outcome will include a frame–based lexical ontology and a legal corpus annotated with *frame* information. In particular, the paper focuses on methodological and design issues, ranging from the customization and extension of the general FrameNet for the legal domain to the linking of the developed resource with already existing Legal Ontologies.

Key words: Frame Semantics, Legal Ontologies, Knowledge Representation, Corpus Annotation

1 Introduction

The last few years have seen a growing body of research and practice in the field of Artificial Intelligence and Law (AI&Law) for what concerns the construction of legal ontologies and their application to the law domain. The importance of this research area is testified by the different Workshops and Conferences which have been organized around this topic. However, as [1] points out, existing legal ontologies vary significantly, for what concerns their underlying structure and organization, the way they are constructed (either top–down or bottom–up) and how they are exploited in different applications. In this paper, we will focus on a particular type of ontology, the so–called *lightweight or lexical ontologies* [2], whose main feature consists in bridging the gap between the legal knowledge formalized in domain ontologies on the one hand and the legislative texts on the other hand; this follows from the fact that in this type of ontology legal concepts are paired with their lexical realizations. This feature makes this type of ontology particularly suitable for use in Information Extraction and Semantic Tagging tasks. Note that these ontologies are typically bootstrapped from legal texts (either manually or through ontology learning techniques).

The most notable example of this type of ontology in the legal domain is represented by the JurWordNet ontology–driven semantic lexicon [3], together with its multilingual extension LOIS [4]. Both JurWordNet and LOIS have been developed following the WordNet (hereafter referred to as WN) design, where words expressing legal concepts such as 'liability', 'sanction', 'violation' are organized in *synsets* (i.e. sets of synonyms) in turn linked by hierarchical or taxonomical relations such as hyponymy and hyperonymy. Under this view, the meaning of a word is intended as a distinct, atomic semantic object, fully identified by its position in the general semantic network.

However, the taxonomical organization of legal concepts is not the only possible one. Legal experts claim that, despite their utility, WN–like resources are not completely adequate and satisfactory in order to represent events and situations typically expressed in legal documents: this is a consequence of the WN–model [5] they follow. Interestingly enough, this claim is in line with the Van Kralingen's proposal of a *frame language* as a plausible method for the conceptual representation of legal knowledge [6]; in spite of the fact that this proposal dates back to the early '90s, it still represents a need commonly felt in the AI&Law community.

In this paper we propose to face this need by developing a lexical resource based on Fillmore's *Frame Semantics* [7] and on the organization principles underlying the FrameNet project [8] (hereafter referred to as FN)[1]. In particular, we propose to build a FN–like resource specialized for the legal domain, by extending and refining the general purpose FN resource. By proceeding in this way, it will be possible to overtly represent the inner structure of complex situations in terms of their participants, e.g. "under which *Circumstances*, which *State of affairs* is sanctioned by which *Principle*".

2 Starting points

In order to create a frame–based resource for the legal domain, our idea is to combine two different approaches from two different research communities, i.e AI&Law and Computational Linguistics. In particular, we aim at revisiting Van Kralingen's proposal of a *frame language* [6] for legal knowledge representation in the light of Fillmore's *Frame Semantics* theory [7].

2.1 Frame–based Legal Ontologies

Amongst the bulk of Legal Ontologies built so far (see [1] for a state–of–the–art), the Van Kralingen and Visser studies are the only ones which envisage a frame–based ontology of law. In their collaborative project, Van Kralingen has defined a theoretical model (i.e. a *conceptual* ontology) and Visser has formalized it in an ontology [9]. The proposed *frame language* is based on the concept of a *norm* and of an *act* as legal conceptual primitives of the legal domain which can be conceived as *frames*, i.e. *data–structures for representing a stereotyped situation in which each element is represented.* Thus, the focus is on the inner structure of a *norm* and of a *legal act*, i.e. on what their building elements are. As shown in Table 1, a *norm frame* is defined as a template in which each element of a norm is represented as a slot of the norm frame. Since every legal action has many different aspects, a *legal act* has also been conceived as a frame. As shown in Table 2, each aspect of an action is represented as a slot of the *act frame* as well.

[1] http://framenet.icsi.berkeley.edu

Table 1. A norm frame as defined in the Van Kralingen's frame–based ontology [6]

Element	Description
Norm identifier	The norm identifier (used as a point of reference for the norm).
Norm type	The norm type (norm of conduct or norm of competence).
Promulgation	The promulgation (the source of the norm).
Scope	The scope (the range of application of the norm).
Conditions of application	The conditions of application (the circumstances under which a norm is applicable).
Subject	The norm subject (the person or persons to whom the norm is addressed).
Legal modality	The legal modality (ought, ought not, may, or can).
Act identifier	The act identifier (used as a reference to a separate act description).

2.2 FrameNet

The FN resource we started from is a lexical resource for English, based on *Frame Semantics* and supported by corpus–evidence. The goal of the FN project is to document the range of semantic and syntactic combinatory possibilities of each word in each of its senses. Typically, each sense of a word belongs to different Semantic Frame, conceived in [8] as "a script–like conceptual structure that describes a particular type of situation, object or event along with its participants and properties". For example, the "Apply_heat" frame describes a common situation involving participants such as "Cook" and "Food", etc. , called Frame Elements (FEs), and is evoked by Lexical Units (LUs) such *bake, blanch, boil, broil, brown, simmer*, etc. As shown by the following example, the frame–evoking LU can be a verb (bolded in the example) and its syntactic dependents (those written in subscript) are its FEs: [Matilde $_{Cook}$] **fried** [the catfish $_{Food}$] [in a heavy iron skillet $_{Heathing\_instrument}$].

The type of representation produced by FN is a network of "situation–types" (frames) organized across inheritance relations between Frames, as opposed to a network of meaning nodes, as in the case of WN. In FN, Frame Elements can be also specified with Semantic Types (i.e. ontological categories) employed to indicate the basic typing of fillers that are expected in the Frame Element. Most of these semantic types correspond directly to synset nodes of WN, and can be mapped onto already existing ontologies. FN currently contains more than 800 Frames, covering roughly 10,000 Lexical Units; these are supported by more than 135,000 FN–annotated example sentences.

3 Our approach

This section outlines our approach to the construction of a FN resource for the legal domain. Our eventual goal is to instantiate the Van Kralingen's frame–based approach to the representation of legal knowledge by exploiting the FN model. While the Van Kralingen's methodology is mostly based on domain–theoretical assumptions, we are rather planning to develop a corpus–based lexical–semantic resource which permits accounting for how complex events and situations are expressed within legal documents.

Table 2. An act frame as defined in the Van Kralingen's frame–based ontology [6]

Element	Description
Act identifier	The act identifier (used as a point of reference for the act).
Promulgation	The promulgation (the source of the act description).
Scope	The scope (the range of application of the act description).
Agent	The agent (an individual, a set of individuals, an aggregate or a conglomerate).
Act type	The act type. Both basic acts and acts specified elsewhere can be used.
Means	The modality of means (material objects used in the act or more specific descriptions of the act).
Manner	The modality of manner (the way in which the act has been performed).
Temporal aspects	The temporal aspects (an absolute time specification).
Spatial aspects	The spatial aspects (a specification of the location where the act takes place).
Circumstances	The circumstantial aspects (a description of the circumstances under which the act takes place).
Cause	The cause for the action (a specification of the reason(s) to perform an action).
Aim	The aim of an action (the goal visualized by the agent).
Intentionality	The intentionality of an action (the state of mind of the agent).
Final state	The final state (the results and consequences of an action).

The linguistic–empirical evidence provided by such a corpus–based methodology results in a bottom–up organization of legal knowledge.

As opposed to a WN–like resource, we think that a FN–like approach can be particularly suitable for the legal domain for a number of reasons. While in WN words are organized as hierarchies or taxononies of synsets, according to FN principles word senses are related to each other only by way of their links to common background Frames.

Moreover, as Fellbaum noted in [5], "WordNet reflects the structure of frame semantics to a degree, but suggested that its organization by part of speech would preclude a full frame semantic approach". In FN, on the other hand, the lexical units that evoke a frame are not restricted to a single part part of speech. For example, the Frame "Process_end" is evoked by both verbs such as *to conclude*, nouns such as *end* and adjectives such as *final*. This is a very important FN feature when dealing with corpora of legal language. According to [10], it is very common in legal texts that events are expressed through nominal rather than verbal constructions. It follows that, for example, the Frame "Prohibiting" can be evoked both by the verb 'to prohibit' and by the deverbal noun 'prohibition'.

To our knowledge, the only effort within the AI&Law community devoted to the use of FN is reported in [11]. As part of a layered approach to a legal domain representation, the authors exploit nine Semantic Frames selected from FN. Different Frame Elements from different Frames have been occasionally combined to represent the legal–domain knowledge contained in six judical judgments of the Supreme Court of Justice of Portugal. They overtly argue for "a corpus–based methodology for an ontology construction

that seeks the rigorous linguistic analysis aiming at formalization". Yet, differently from our approach, they do not explicitly aim at creating a domain–specific FrameNet resource.

During the initial design phase, we have considered what has been done in other specialized domains as well. For example, within the bio-medical domain a domain–specific FN extension has been proposed in [12], who successfully developed a BioFrameNet through creating new Semantic Frames relevant to the domain of molecular biology and linking them to domain–specific biomedical ontologies. However, in the construction of such a FN resource for the bio–medical domain the authors faced bio–medical language peculiarities which pose challenges rather different from ours. As laid out in Section 4.1, the specific relationship between the ordinary and legal language (i.e. their closed intertwining) raises more challenging issues.

Following the underlying organization of the FN model, we intend to produce:

1. a legal corpus annotated with frame information,
2. a lexical frame–based resource covering the legal and domain terms occurring in the annotated corpus.

4 Design issues

A number of issues worth discussing has been encountered during the design stage of a FN extension and specialization for the legal domain. They mainly concern the choice of i) whether and to what extent the general FN Frames should be customized for legal text annotation purposes, and ii) how to ontologically type the lexical fillers of Frame Elements for domain–specific purposes.

4.1 FrameNet customization strategies for a *legal FrameNet*

Following the approach laid out in Section 3, we plan to build a legal domain extension of the general FN on the basis of the already existing set of Semantic Frames. An initial stage of corpus annotation has been foreseen as a first 'investigation' phase. In a later stage, in which a suitable amount of annotations will be done, there will be the choice of whether and which kind of customizations are needed according to the corpus evidence and domain requirements.

As pointed out in [12], a key issue encountered while dealing with domain–specific texts is whether or not the creation of a new Frame is warranted. Within the legal domain the situation is made more difficult since the technical language used in the legal domain is closely intertwined with common language. According to linguistic studies (see among others [13]), legal language, still differing from ordinary language, is in fact not dramatically independent from every–day speech. This implies that it is no longer simply an issue of keeping existing Frames or creating new ones from scratch to convey domain–specific semantics. Accordingly, the specialization phase is concerned with the following three customization strategies which differ in their increasing degree of modification to the general FN resource:

1. the exploitation of domain–specific Semantic Types which classify Frame Elements from the general FN repository,
2. the introduction of one or more new Frame Elements within an existing Frame,
3. the splitting with a new Frame.

An example of 1. is provided in the excerpt of annotation reported in Section 5 below, where the Semantic Type "LegalDescription" has been added to the Frame Element "Principle" in order to ontologically type the lexical filler of this participant to the Frame "Prohibiting".

Special attention is paid to the introduction of a new Frame Elements within an existing Frame. It is such the case of the following sentence *Il venditore deve consegnare al consumatore beni conformi al contratto di vendita* 'The seller must deliver goods to the consumer which are in conformity with the contract of sale', which instantiates the Frame "Being_obligated", evoked by the Lexical Unit *deve* 'must'. A new Frame Element "Beneficiary" should need to be added to the list of the semantic roles already existing to the Frame at hand, in order to describe the *addressee of the duty* (i.e. 'to the consumer'). The original Frame only includes a Frame Element "Duty" (in this case 'deliver goods') and "Responsible_party", i.e. 'the person who must perform the Duty' (in this case 'the seller').

In a sentence such as *uno Stato membro può vietare, per motivi di interesse generale, la commercializzazione sul suo territorio, tramite contratti negoziati a distanza, di taluni prodotti e servizi* 'a Member State can prohibit, for reasons of general interest, commercialization on its territory, through contracts negotiated at a distance, of certain products and services', the splitting with a new Frame "Authority_prohibiting" is needed. The syntactic realization of the sentence above shows that it is an *enacting authority* (i.e. 'a Member State') which enacts a normative principle, i.e. a prohibition, rather than a "Principle" which prohibits a "State–of–affairs".

4.2 Towards an ontological typing of Frame Element

In designing a FN extension for the legal domain, we considered the ontological typing of Frame Elements as a fundamental stage. According to [8], the general use of Semantic Types in FN is "to record information that is not representable in our frame and frame elements hierarchies". It is done through the categorization of the sort of lexical fillers that is expected in a Frame Element. We intend to exploit this FN usage in order to domain–specifically categorize Frame Elements involved in a situation expressed by legal texts, on the basis of Legal Ontologies. As pointed out in [14], the real benefit of integrating a lexical and an ontological resource follows from distinguishing lexicalized and not–lexicalized concepts through keeping them as *distinct layers of semantic information* but even linking them.

The domain ontology we intend to use is the Core Legal Ontology (CLO)[2] [2], that specializes the DOLCE foundational ontology library[3] [15]. CLO was chosen since it provides lexicalizations of ontological classes (i.e. juridical concepts), both in Italian and in English. Moreover, it has been exploited as an ontological resource reference in LOIS and in the DALOS project [16].

The possibility of mapping this FN–like resource onto a so–called *lexical ontology*, such as JurWordNet, is still under discussion.

5 An example of legal texts annotation

In this section, we report an example of annotation carried out on the Directive 1999/44/EC of the European Parliament and of the Council of 25 May 1999 on certain

[2] http://www.loa-cnr.it/
[3] http://dolce.semanticweb.org

aspects of the sale of consumer goods and associated guarantees. For the annotation we used the Salsa Tool [17], freely available for research purposes. It offers a graphical representation of a text, already annotated at the syntactic level, and allows the user to annotate Frames and Frame Elements. Figure 1 shows the annotation of the following sentence: *La decisione 90/200 ha vietato l'esportazione dal Regno Unito di taluni tessuti e organi bovini solo il 9 aprile 1990* 'The decision 90/200 prohibited the exportation from the United Kingdom of certain bovine tissues and organs only the 9th April 1990'.

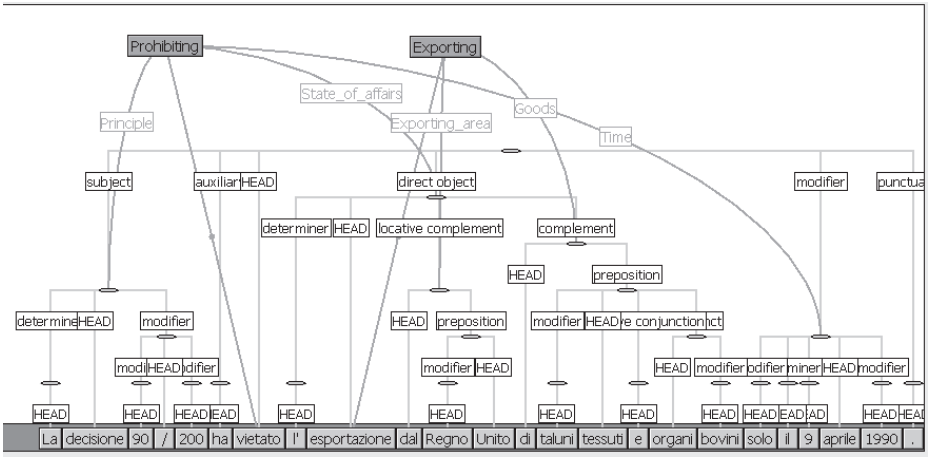


Fig. 1. An annotation example

Two Frames have been annotated: i) a Frame "Prohibiting", evoked by the Lexical Unit *ha vietato* 'prohibited', together with three Frame Elements, i.e "Principle", "State–of–affairs" and "Time", and ii) a Frame "Exporting", evoked by *l'esportazione* 'the exportation', together with "Exporting_area" and "Goods" as participants. It should be noted that the annotated Frames refer respectively to the *legal domain* properly and to the *commerce domain* which is regulated by the Directive at hand. Interestingly, the two Frames are closely intertwined, in the sense that the textual span of the Frame Element "State–of–affairs", part of the Frame "Prohibiting", (i.e. *l'esportazione dal Regno Unito di taluni tessuti e organi bovini* 'the exportation from the United Kingdom of certain bovine tissues and organs') instantiates in turn the Frame "Exporting".

The annotation of the textual span of Frame Elements was carried out on the top of the syntactic dependency relations automatically detected by the DeSR syntactic parser [18] [4]. The use of the Semantic Type "LegalDescription", node of CLO, has been envisaged in order to ontologically type the lexical fillers (i.e. *la decisione 90/200* 'the decision 90/200') of the Frame Element "Principle".

[4] The parser used for this example was trained on a corpus of Italian newespapers; we are currently considering whether to develop a domain–specific version.

6 Conclusion

In this paper we introduced our approach to the construction of a FN resource for the legal domain. Through a customization phase of the general FN, we intend to produce i) an annotated corpus of legal texts and ii) a frame–based lexical–semantic resource. A strategy devoted to ontologically type the lexical fillers of Frame Elements annotated is foreseen as well in order to domain–specifically categorize participants involved in a situation expressed by legal texts. Through this, the developed FN resource will be linked to already existing legal ontologies, thus resulting in a combined resource giving access to both the lexical and ontological aspects of legal texts.

Even though we present a work which is at an early stage of development, we foresee a number of possible applications and future extensions. Firstly, a frame–based annotated corpus of legal texts can be use to train test tools for semantic processing of legal texts, such as Semantic Role Labeling (SRL) tools. Namely, these SRL tools will be developed using language–independent unsupervised or semi–supervised machine learning algorithms, trained on the annotated corpus. The encouraging results achieved so far by SRL systems in the general language domain [19], are seen as an interesting opportunity to advance the state–of–the–art of Textual Case–based Reasoning (CBR) in the legal domain (see [20] for a frame–based approach).

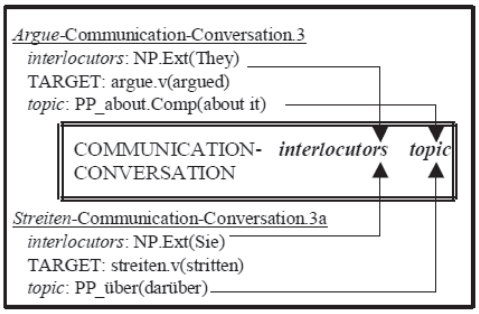


Fig. 2. Semantic frame as an interlingual representation [21]

Secondly, a multilingual FrameNet–like lexical resource can support semantic searching of legal texts in different languages. As reported in [21], where a bilingual German–English dictionary has been built on *Frame Semantics* principles, Semantic Frames are used as structuring devices to link multilingual lexicon fragments. Figure 2, extracted from [21], shows how a given combination of semantic and syntactic combinatorial properties of a given lexical unit in the source language has a correspondence link to its counterpart in the target language.

References

1. Valente A., Types and Roles of Legal Ontologies, in Law and the Semantic Web, LNCS, Volume 3369/2005, Springer Berlin / Heidelberg, 2005, pp. 65-76.

2. Gangemi, A., Sagri, M.T and Tiscornia, D. (2005), A constructive framework for legal ontologies,in Law and the Semantic Web, Benjamins, R., Casanovas, P., Breuker, J., and Gangemi,A., (eds), Springer Verlag.
3. Sagri M–T. (2004), Tiscornia D., Bertagna F., Jur-WordNet, in Proceedings of the Second Global WordNet Conference, pp. 305–310, Brno, Czech Republic, January 20-23.
4. Peters, W. , Sagri, M. T. , Tiscornia, D., The structuring of legal knowledge in lois, Artificial Intelligence and Law, vol. 15, pp. 117-135, 2007.
5. Fellbaum, C. (ed) (1998), WordNet: An electronic lexical database. MIT Press.
6. Kralingen, R., Oskamp, E., and Reurings, E. (1993), Norm frames in the representation of laws, in Proceedings of Jurix.
7. C. J. Fillmore, Frame semantics and the nature of language. Annals of the New York Academy of Sciences, (280):20-32, 1976.
8. Ruppenhofer, J., Ellsworth, M., Petruck, M.R.L., Johnson, C.R., and Scheffczyk, J. (2006), FrameNet II: Extended Theory and Practice, available online at `http://framenet.icsi.berkeley.edu/`
9. Visser, P.R.S., Knowledge Specification for Multiple Legal Tasks; A Case Study of the Interaction Problem in the Legal Domain, Computer/Law Series, No. 17, Kluwer Law International, The Hague, The Netherlands, 1995.
10. Venturi, G., Parsing Legal Texts. A Contrastive Study with a View to Knowledge Management Applications, in Proceedings of Sixth International Conference on Language Resources and Evaluation (LREC 2008), Workshop *Semantic Processing of Legal Texts*, Marrakech, Morocco, May 26-1 June 2008, CD-ROM.
11. Isa Mara da Rosa Alves, Rove Luiza de Oliveira Chishman, Paulo Miguel Torres Duarte Quaresma, The Construction of a Juridical Ontology, in Proceedings of ICAIL '07 June 4-8, Palo Alto, CA USA.
12. Dolbey A, Ellsworth M, Scheffczyk J (2006), BioFrameNet: A Domain-Specific FrameNet Extension with Links to Biomedical Ontologies, O. Bodenreider, ed., Proceedings of KR-MED, 87-94.
13. Mortara Garavelli, B., Le parole e la giustizia. Divagazioni grammaticali e retoriche su testi giuridici italiani, Torino, Einaudi (2001).
14. Prèvot, L., Borgo, S., and Oltramari, A. (2005), Interfacing Ontologies and Lexical Resources, In Proceedings of OntoLex 2005 - Ontologies and Lexical Resources, Jeju Island, Republic of Korea, 15 October.
15. Masolo, C., Gangemi, A., Guarino, N., Oltramari, A., and Schneider L. (2004), Wonderweb deliverable d18: The wonderweb library of foundational ontologies, tech.rep.
16. Agnoloni, T., Bacci, L., Francesconi, E., Peters, W., Montemagni, S., Venturi, G. (2009), A two-level knowledge approach to support multilingual legislative drafting in: J. Breuker, P. Casanovas, E. Francesconi, M. Klein (eds.), Law, Ontologies and the Semantic Web Amsterdam, IOS Press.
17. Erk, K., Kowalski, A., and Pado, S. (2003), The salsa annotation tool-demo description, In Proceedings of the 6th Lorraine-Saarland Workshop, 111-113.
18. Attardi, G., and Dell'Orletta, F., (2009), Reverse Revision and Linear Tree Combination for Dependency Parsing, in In Proceedings of the North American Chapter of the Association for Computational Linguistics -Human Language Technologies short papers (NAACL HLT), Boulder, Colorado. A detailed description of DeSR parser is available at `http://sites.google.com/site/desrparser/`
19. Gildea, D., and Jurafsky, D. (2002), Automatic labeling of semantic roles, In Computational Linguistics, volume 23, 245-288.

20. Mustafaraj, E., Hoof, M., and Freisleben, B. (2006), LARC: Learning to assign knowledge roles to textual cases. In Proceedings of FLAIRS, 370-375.
21. Boas, Hans C. (2002), Bilingual FrameNet Dictionaries for Machine Translation. In M. Gonzalez Rodriguez and C. Paz Suarez Araujo (eds.), Proceedings of the Third International Conference on Language Resources and Evaluation. Las Palmas, Spain. Vol. IV: 1364-1371.

Multilingual Access Modalities to Legal Resources Based on Semantic Disambiguation

G. Peruginelli and E. Francesconi

ITTIG-CNR – Institute of Legal Information Theory and Techniques
Italian National Research Council, Italy

Abstract. An effective access to multilingual legal materials is strictly linked to the peculiarities of legal language as a technical language closely related to the diverse legal systems. This paper proposes an approach for a coherent cross-language information retrieval system based on semantic document indexing able to contextualize queries for terms disambiguation and translation.

1 Multilingualism in the law domain: an overview

Today there is a strong need for worldwide sharing of legal information as internationalization and increasing globalization of market economy and social patterns of life have created a situation where the need for legal information from foreign countries and from different legal systems is greater than ever before. Such requirement is not new, but it is getting increasingly complex to meet under the pressure of the rapid cross-border transactions occurring between people of different legal cultures and languages. It is no doubt that the exchange of information is largely dependent on language, to be intended not only as a system of symbols, but also as a mean of communication [1], a tool for mediating between different cultures. As regards the language of the law, such language properties have a major impact on the exchange of legal information. Cross-Language Information Retrieval (CLIR) refers to a functionality implying the ability of a system to process a query for information in any language, search a multi-language collection and return the most relevant documents. As such, CLIR offers a practical approach towards worldwide sharing of knowledge for its potential to make information accessible across language barriers. The difficult task to effectively access multilingual legal material through information retrieval systems is definitively to match and weight legal terms across languages [2]. This generally implies translating from the language of the query to that of the material to be found or viceversa, and addressing the problem of word disambiguation which is greatly increased when mapping over legal languages. In fact crossing the language barrier between search requests and documents implies facing the problems of the system-bound nature of legal terminology and devising methods to map concepts between different legal systems. It is a matter of fact that in the last decades research and developments on CLIR have progressed rapidly, important cooperative initiatives have been undertaken at international level1 and issues of multilingual querying, presentation and retrieval have been extensively tackled mainly in the area of general domain information. A rather limited number of studies and applications have been produced in domain-specific areas and cross-language retrieval of legal information has received limited attention. In a multilingual access environment information is searched, retrieved and presented effectively

without constraints due to the different languages and scripts used in the material to be searched and in the metadata, that is descriptive and semantic information allowing the retrieval of indexed documents to be found. One main question arising in the context of CLIR of law material is how should the language barrier between the search requests and documents be crossed. This involves decisions about what to translate: search requests into the language of the documents or documents into the language of the request, or even both. Besides this fundamental question, one crucial issue regards the best approach to adopt in carrying out translation and how far translating terms can be successful when dealing with legal information. From a practical point of view the approach for large collections is usually based on the most economical method, consisting in simply translating the query at retrieval time into the document (or metadata) languages, although it would be possible to translate all of the documents into the query language. This presupposes that the query can be translated in a reasonably accurate fashion and that monolingual retrieval systems are available for all of the document languages. Although many experiments have been carried out in general domain information using query translation techniques, in the real world they pose a number of problems related to the need of contextualization and interpretation, which are increased in the law domain [3].

2 Key components of cross-language legal information retrieval

Retrieving information over languages implies facilities such as multiple language recognition, translation, manipulation of information of queries and documents, cross-language search and retrieval, display and merging of results [4]. Basically, these components reflect different sides of the problem of multilingual access, covering technical and linguistic aspects. In such a context the system-bound nature of legal terminology, the complexity of legal languages, legal translations issues and comparative law aspects are major issues having important implications for effective retrieval of law across languages. Legal translation is an essential function for cross-language retrieval systems. One major question concerns the translation strategy to be adopted in order to ensure that users access legal information independently of the language used in a query. The relation between law and language can significantly broaden the scope of legal translation theory. In fact, while it can be assessed that everyday language already implies a formalized way of communication, legal language introduces a supplementary system of formalisation [5]. Although legal translation demands precision and certainty, it is bound to use abstractions, whose meanings derive from particular changing cultural and social contexts. These contexts generate a certain degree of ambiguity, which increases when the legal cultures and systems are vastly different from each other. On a practical level the problems raised by legal translation are strictly connected with those related to the variety and diversity of legal systems and as such to comparative law. Retrieval systems to legal information across different legal systems represent a practical approach to the confrontation and exchange of legal cultures. The whole process of interaction between legal languages can be identified as finding equivalents across legal systems. If no acceptable equivalents can be found in the target-language, subsidiary solutions must be sought, such as no translation and use of source terms, paraphrasing, creating a neologism with explanatory notes. Pure linguistic problems are likely to be

encountered due to legal false friends.[1] Despite the difficulties in establishing the equivalence of legal concepts belonging to different legal systems, a compromise has been adopted in trying to favour the integration of diverse legal cultures, while respecting each national legal system. What is needed is the identification of a common ground, namely common legal concepts and facts which, although not perfectly coinciding with those belonging to other systems, are conceptually close. It is up to legal users, once the material has been examined, to perceive the differences and peculiarities which make these resources unique. It is to be underlined that this does not necessarily lead to noise or unsuccessful searches, but allows for a first-phase search in context, useful to give evidence of the existence or non-existence of a specific concept in other legal systems.

3 Towards solutions for multilingual retrieval of law

Knowledge-based systems can greatly contribute to cross-language retrieval through the structure and function of thesauri and ontologies. In fact these tools have the potential to manage the complexities of terminology in language and provide conceptual relationships, ideally through an embedded classification/ontology [6]. In the domain of law efforts are starting to be made in this direction. These are represented for example by the Lexical Ontologies for Legal Information Sharing (LOIS) project [7], Jurwordnet[2] [8], DALOS project [9] and by a number of linguistic tools like the Legal Taxonomy Syllabus (LTS)[3], Eurovoc Thesaurus[4] and Jurivoc[5], the legal thesaurus of the Swiss Federal Court. But in practice, aligning law vocabularies of two or more languages is a hard process. Ideally a multilingual legal thesaurus should include all concepts needed in searching by any user in any of the source languages, but difficulties arise in making the systems of legal concepts the same for all languages as a different language often suggests a different way of classifying law material and a system needs to be hospitable to all of these. In such a context what cross language retrieval of legal information systems should manage is mapping each query term from the source language to its possible multiple equivalents in the target language. However each of these equivalents may have other meanings in the target language or may not have a precise equivalent, requiring to be mapped to broader or narrower terms, but this can lead to distorting the meaning of the original query. Multiple meanings can be disambiguated through users interaction, but the success of this approach depends on the

[1] For examples the terms "administrative tribunals" cannot be translated in French as "tribunaux administratifs". The English word for the French tribunal is Court and the administrative tribunals are administrative commissions which are comparable, mutatis mutandis, to the French "autorités administratives indépendantes".

[2] The law lexicon is characterized by both taxonomic vertical and associative horizontal relations and it has been developed by the Institute of Legal Information Theory and Techniques (ITTIG-CNR)

[3] LTS consists of both a database and a software development within the European project "Uniform terminology for European Private Law" and is coordinated by the Dipartimento di scienze giuridiche of the University of Turin. Available at: `http://www.eulawtaxonomy.org/index_en.html`

[4] It is the multilingual and polythematic thesaurus of the European Union. `http://europa.eu/eurovoc/`

[5] `http://www.bger.ch/it/index/juridiction/jurisdiction-inherit-template/jurisdiction-jurivoc-home.htm`

quality of the hierarchy of concepts, the provision of well-structured cross-references, and on the interface of the system. The adoption of a common metadata format, where to accommodate semantic classification of legal documents by using categories of law, can ensure a successful legal mapping across languages and systems. Categories of law of a specific legal system, in fact, represent the way how retrieval can be satisfactorily achieved. As often there is no conceptual nor content similarity between the categories of law (i.e. trade law, constitutional law, criminal law) of the different legal systems, mapping between such law categories is necessary to reach proper contextualisation of the query in the diverse legal systems. An example illustrates the need for such mapping. The concepts related to property rights, such as the development of property law, land law, property questions on insolvency, intellectual property, etc. according to UK law belong to the field of property law, whereas in the Italian legal system these legal facts are regulated by private law, agricultural law and industrial law. Below an illustration is given of a possible approach to a coherent multilingual legal information access based on categories of law and on full text and metadata indexing.

4 A possible approach for accessing multilingual legal resources

Let us consider an information system offered to the users where a full text and metadata indexes in a multi-language environment are available. In this context two are the different advanced search modalities that can be envisaged:

1. *metadata-based document querying* (MBDQ);
2. *keyword-based document querying* (KBDQ), combined with category (*category-based document querying* (CBDQ)).

Case 1. Advanced search: the user submits a query filling in the fields related to the adopted metadata schema (for example DC metadata set, taken here as reference).

Case 2. Simple search: the user submits a query, filling an unqualified text box using keywords. Moreover, in order to make the query more focused, the user may choose a legal category of the legal system associated with a language domain.

Dealing with querying and retrieving multi-language documents, basically involves the problem of query translation. As discussed in Section 2, especially in legal domain, a word in the query language can be ambiguous, having therefore different translations in a target language, each corresponding to a legal category in the target legal system (i.e. the Italian word “dolo” has two different translations into English: “fraud” and “malice”, respectively belonging to private law and criminal law). The right sense of an ambiguous word in query language can be obtained only by word contextualization, giving the right sense to the context in terms of a legal category. A legal category in the legal system of the query language can be mapped to the correspondent legal category in the target legal system, therefore the right translation of the ambiguous word can be obtained. If more than one category in the target legal system corresponds to the original legal category, more than one translations of the ambiguous word are selected. Therefore, in both the modalities of querying (MBDQ and KBDQ+CBDQ), the identification of a legal category is essential in order to identify the right translation of an ambiguous word. The procedures used to obtain these results in MBDQ and in KBDQ+CBDQ modalities are described respectively in Section 4.1 and 4.2 (Fig. 1 can be used as reference).

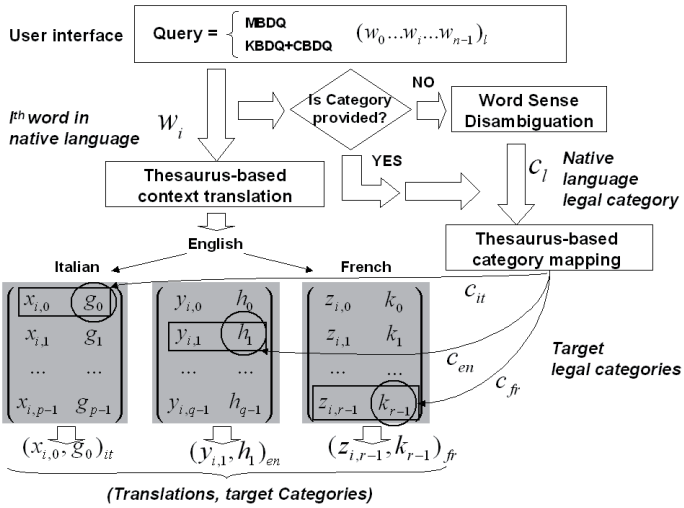


Fig. 1. Query translations of in MBDQ and KBDQ+CBDQ modalities

4.1 Query based on metadata (MBDQ)

MBDQ represents an "advanced search" modality of querying a qualified document index. The user first of all is required to choose a legal system, thus implicitly identifying a language for queries, and a legal category, identifying the right translations of possible ambiguous words. Then each metadata field is filled with a set of words $(w_0, w_1, ..., w_{n-1})_l$, representing a context expressed in the query native language l, that has to be translated by a *thesaurus-based context translation* procedure. Not every field has to be translated. In fact, bibliographic metadata (as for example the Dublin Core metadata set) can be divided into query language-dependent and query language-independent metadata. For example *dc:title* field is query language-independent since, for example, the title of a document has to be queried in its native language, independently from the query-language. Therefore only the contents of query language-dependent metadata have to be translated. While in a multi-language environment the semantic classification (*dc:subject*) is usually query language-independent (or neutral [10]), within a multi-language legal domain this is not true (Section 2). For this reason a semantic category has to be translated, by mapping it from a legal system to different target ones. Also the content of the widely used access point *dc:description* field (the document abstract) is query language-dependent: the information contained is often expressed using a semi-technical language; therefore a *dc:description* field can be considered as important to translate as the *dc:subject* . The contents of *dc:subject* and *dc:description* fields, submitted in a native language are translated in a "pivot" language (English) [11]. Then, from the "pivot" language, the query is translated again to the other languages used by the retrieval system. The use of a "pivot" language in a N-language environment allows the reduction of the number of bilingual thesauri from a factor N^2 to a factor N, and also allows the solution of the problem of the non-availability of some bilingual thesauri. As discussed in Section 2 the main problem with translation is that a single word (w_i) or expression in the native language can have dif-

ferent translations in a target language, depending on the context. For example, let us assume, without loosing generality, that w_i be an ambiguous single word of the context $(w_0, w_1, ..., w_{n-1})_l$ in the *dc:description* field in query native language l. According to Fig. 1, different English translations $\{y_{i,0}, y_{i,1}, ..., y_{i,q-1}\}$ can be associated to w_i, each one corresponding to as many legal categories $\{h_0, h_1, ..., h_{q-1}\}$. For example, being the language l= Italian and w_i= "dolo", possible translations in English are $y_{i,0}$="fraud" related to law category h_0="private law" and $y_{i,1}$="malice" related to law category h_1="criminal law". The right translation can be obtained only by knowing the sense, namely the category h_j, of the context in the query native language, where w_i is contained. Such a context, or legal category, is required and is provided by the user using a *dc:subject* field. When a category c_l (Fig. 1) is selected, the problem arises of different classification schemes in different languages, corresponding to different legal systems (Section 2). The problem can be solved by using a *thesaurus-based category mapping*. In fact, when the category c_l is submitted as a query parameter, the category c_l is mapped in the corresponding, or the closest, categories in the "pivot" language, and from it to the other languages considered by the retrieval system ($c_l \Rightarrow c_{en} \Rightarrow \{c_{it}, c_{fr}\}$), using a classification schema. In accordance with Fig. 1 and without loosing generality, let us assume that only one legal category $c_{en} = h_1$ in the English legal system corresponds to the legal category c_l ($c_l \Rightarrow c_{en} = h_1$). Consequently, the English translation $y_{i,1}$ (Fig. 1) can be selected (in our example, the English word $y_{i,1}$="malice", related to law category h_1="criminal law" is selected as the right translation of the Italian word w_i="dolo"). If more than one category of the target legal system can be associated to c_l, all the corresponding translations of the current w_i are selected. When all the words of the current context are translated in *dc:description*, we obtain the translation of the submitted context $(w_0, w_1, ..., w_{n-1})_l$ from language l to retrieval system target languages. The category c_l is also mapped to the corresponding categories in the target languages. Now queries in different languages are ready to be dispatched to the related domain language indexes (Fig. 2).

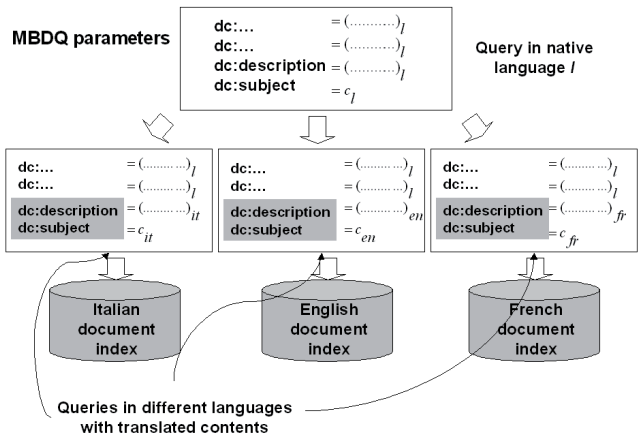


Fig. 2. MBDQ: results of query translation in different languages (in grey metadata whose content is translated)

4.2 Query based on keywords and legal categories (KBDQ+CBDQ)

A query based on keywords and legal categories represents the "simple search" modality of querying our multilingual retrieval system. In this modality the user is provided only with an unqualified text box to be filled with a context $(w_0, w_1, ..., w_{n-1})_l$ of words in a native language l. Words identifying the context will be translated into the target languages of the retrieval system (*thesaurus-based context translation*). Moreover, the user may provide a legal category of the query legal system. If the user selects a legal category c_l, among the values of *dc:subject* in the query legal system, a procedure of *thesaurus-based category mapping* is executed, as described in Section 4.1, obtaining the correspondences of c_l in target legal systems (Fig. 1). If the user fills only the unqualified text box without choosing any value in *dc:subject*, since category is essential for translation, the right sense to the query context can be provided by a procedure of automatic word sense disambiguation, which assigns a legal category to a context as described in Section 5. The legal category thus identified in native query language, is then mapped to the related legal categories in target legal systems (*thesaurus-based category mapping*). At the end of the process, the right translations of ambiguous words can be obtained, as discussed in Section 4.1 (Fig. 1), and as many different queries as target languages considered can be dispatched to the different language indexes (Fig. 3).

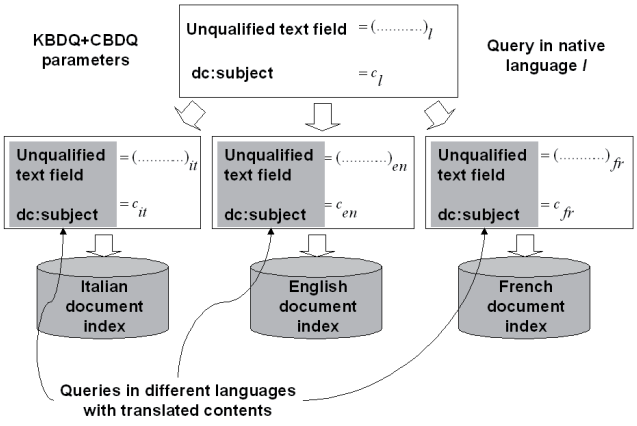


Fig. 3. KBDQ+CBDQ: results of query translation in different languages (in grey metadata whose content is translated).

5 Automatic word sense disambiguation

The problem of assigning the right meaning to a word in context is a problem of assigning the right sense to the context itself out of the various meanings that can be assigned to the ambiguous word. According to the literature lots of methods have been used to solve the problem of automatic disambiguation:

- Thesaurus-based disambiguation [12];
- Disambiguation based on sense definitions [13];
- Disambiguation based on translation in a second-language dictionary [14];
- Bayesian disambiguation [15].

In our retrieval system word disambiguation is a problem of context categorization with respect to the legal categories considered within a legal system. Moreover [15] context categorization is the same problem of document categorization, once we view contexts as documents and word sense as categories. For these reasons in our system we can use different document categorization methods described in literature [16], trained with labelled documents of different legal categories of a particular legal system and language. At the end of the training phase each category profile (for example a vector of weighted terms relevant to it [17]) can also be considered as a context profile to be used for disambiguation function. While experimental results on legal document categorization have been carried on (see [17] [18]) a similar experiment on a set of previously categorized query is to be carried out. It is important to execute automatic word disambiguation prior to translation, because, as discussed, correct word translation depends on contextualization activity of words in their native language.

6 Conclusions

The approach analyzed in this contribution fully reflects the problems illustrated so far, as legal information retrieval is strongly conditioned to the legal orders' specificity, that is to the concepts on which they are based. It is a matter not so much of handling the diversity of languages in which these concepts are expressed, rather considering and managing the peculiarities of the law environment, that is the historical and cultural heritage of a given legal system, whose comparison with other legal orders is often hard, if not impossible. Therefore the real problem is how to establish a correspondence among concepts of diverse legal systems expressed in different languages. A comparative analysis of legal concepts and, parallel to this, the study of translation theory and practice to be intended as search of functional equivalents, are fundamental activities to reach a satisfactory mediation among different legal identities, thus ensuring intercultural communication and at the same time increasing the value of diversity, to be intended as a strength and a challenging factor of integration. Europe is a typical example of this phenomenon: it is praised for its strategies in language policy as a modern relevant experiment of institutional and political innovation which is in the position to open new forms of coexistence and cooperation. In this context multilingual legal information retrieval systems do represent the necessary tools to encourage multilingualism in the law domain and have the chance to make it effective. In particular, in this article an approach is proposed to offer the users a single point of access into multilanguage document collections where categories of law are the key factors to point to relevant material irrespective of the language used in a query. This is done through techniques able to translate legal queries to different target languages, disambiguating ambiguous words if needed. Basically, the approach gives the benefit of accessing multi-language legal documents respecting the identity and the peculiarities of different legal systems.

References

1. L. Wittgenstein, *Philosophical investigations.* Oxford : Blackwell, 1997.

2. R. Sacco, "Droit et langue," in *Rapports italiens au XV Congrès international de droit comparé*, Milano. 1998.
3. E. Francesconi and G. Peruginelli, "Opening the legal literature portal to multilingual access," in *Proceedings of the Dublin Core Conference*, pp. 37-44, 2004.
4. C. Peters and E. Picchi, "Across languages, across cultures: issues in multilinguality and digital libraries," *D-Lib Magazine*, 1997. Retrieved May 11, 2009, from (http://www.dlib.org/dlib/may97/peters/05peters.html).
5. H. Prakken and G. Sartor, "On the relation between legal language and legal argument: assumptions, applicability and dynamic priorities," in *Proceedings of the Fifth International Conference on Artificial Intelligence and Law*, pp. 1-10, New York: ACM, 1995.
6. D. Soergel, "Multilingual thesauri in cross-language text and speech retrieval," in *Working notes of AAAI Symposium on Cross-Language Text and Speech Retrieval*, 24-26 March 1997.
7. W. Peters, M. Sagri, and D. Tiscornia, "The structuring of legal knowledge in lois," *Artificial Intelligence and Law*, vol. 15, pp. 117-135, 2007.
8. A. Gangemi, M. Sagri, and D. Tiscornia, "A constructive framework for legal ontologies," in *Law and the Semantic Web* (Benjamins, Casanovas, Breuker, and Gangemi, eds.), Springer Verlag, 2005.
9. T. Agnoloni, L. Bacci, E. Francesconi, W. Peters, S. Montemagni, and G. Venturi, "A two-level knowledge approach to support multilingual legislative drafting," in *Law, Ontologies and the Semantic Web* (J. Breuker, P. Casanovas, M. Klein, and E. Francesconi, eds.), vol. 188 of *Frontiers in Artificial Intelligence and Applications*, pp. 177-198, IOS Press, 2009.
10. W. Lee, S. Sugimoto, M. Nagamori, T. Sakaguchi, and K. Tabata, "A subject gateway in multiple languages: a prototype development and lessons learned," in *DC*, pp. 59-66, 2003.
11. F. Sebastiani, "Interactive query expansion with automatically generated category-specific thesauri," in Amita G. Chin (ed.), *Text Databases and Document Management: Theory and Practice, Idea Group Publishing*, Hershey, US, pp. 103-117, 2001.
12. D. Yarowsky, "Word sense disambiguation using statistical models of rogets categories trained on large corpora," in *International Conference on Computational Linguistics*, pp. 454-460, 1992.
13. M. Lensk, "Automatic sense disambiguation," in *Proceedings of the SIGDOC Conference*, pp. 24-26, 1986.
14. I. Dagan and A. Itai, "Word sense disambiguation using a second language monolingual corpus," *Computational Linguistic*, no. 20, pp. 563-569, 1994.
15. W. A. Gale, W. K. Church, and D. Yarowsky, "A method for disambiguating word sense in a large corpus," *Computer and Humanities*, vol. 5, no. 26, pp. 415-439, 1993.
16. F. Sebastiani, "Machine learning in automated text categorization," *ACM Computing Surveys*, vol. 34, no. 1, pp. 1-47, 2002.
17. E. Francesconi and G. Peruginelli, "Access to italian legal literature: Integration between structured repositories and web documents," in *Proceedings of the Dublin Core Conference*, pp. 99-107, 2003.
18. E. Francesconi and G. Peruginelli, "Retrieval of italian legal literature: a case of semantic search using legal vocabulary," in *Proceedings of the Dublin Core Conference*, pp. 97-106, 2005.

Learning and Verification of Legal Ontologies by Means of Conceptual Analysis

Erich Schweighofer

Centre for Computers and Law DEICL/AVR, Faculty of Law
University of Vienna Schottenbastei 10-16/2/5, 1010 Wien, Austria
Erich.Schweighofer@univie.ac.at,rechtsinformatik.univie.ac.at

Abstract. A combination of intellectual input, NLP tools and appropriate ontological representation may overcome the existing bottleneck of legal knowledge acquisition of legal ontologies. Such semi-automatic tools rely on easily available input, extensive iterative semiautomatic checking and refining of this knowledge. Preliminary results using the tools of SOM/GHSOM, KONTERM and GATE show the feasibility of this method. However, it remains to be seen if a sufficient number of legal writers will adapt to this new workbench.

1 Introduction

In law, indexing was and is still a very important tool for coping with the vast body of legal materials. Since the advent of information retrieval, legal full text search has been added to the methods of legal research. However, an index of concepts or legal sources is still considered as the best access to the sequential structure of handbooks, textbooks or collections of materials. Such indices may be also used for the production of summaries of cases (head notes) identifying the important parts of court decisions. Huge reference systems on legal materials also exist either based on citations (e.g. the Austrian index [1] or thesauri (e.g. the Swiss thesaurus [2]).

In previous papers, we have argued for the creation of a dynamic electronic legal commentary [3]. Handbooks or commentaries are the most advanced form of traditional explicit knowledge representation. The dynamic electronic legal commentary differs from the traditional legal handbook or commentary mostly in the integration of semiautomatic means of semantic indexing. Legal ontologies constitute the main knowledge base of a dynamic electronic legal commentary. However, creation of such a knowledge base is time consuming and costly. Means of semiautomatic creation and verification are thus highly required.

The example of a legal commentary should point to the required co-operation with legal experts. Only those are able to produce extensive input and to check the vast semiautomatic output. However, legal writers still prefer intellectual analysis without semiautomatic means.

Legal ontologies should be the core of such a knowledge base. However, these ontologies are either too broad and shallow (e.g. LOIS and DALOS) or too small and deep (e.g. LRI Core) in order to meet the standards of semantic indexing. Thus, we propose the development and refinement of such an ontology by means of conceptual analysis. The remainder of this paper is organized as follows: section 2 describes related work, section 3 gives an overview on the method. In section 4, the status of implementation and problems are discussed. Section 5 contains conclusions and future work.

2 Related work

The main components of legal knowledge are the legal retrieval system (or legal information system) as a huge text corpus with a (mostly) textual representation of the legal order and meta knowledge about the text corpus. Computationally speaking, meaningful semantic indexing is linked to a legal text corpus. Such indexing exists in legal brains, legal books but also legal knowledge bases. Legal structuring as such is done by lawyers, in their minds, and is presented and made explicit in their argumentations and writings. As a product of this process, a legal commentary is considered as the highest level of this endeavour.

The semantic web can be considered as an extension to the current web in providing a common framework that allows data to be shared and reused [4]. Semantic search may also improve low and disappointing results of present legal information retrieval [5].

Thesauri (or legal dictionaries) getting more importance now as a traditional tool for representation of knowledge about legal language use. A thesaurus for indexing contains a list of every important term in a given domain of knowledge and a set of related terms for each of these terms [6]. A lexical ontology builds up from this basis with works on glossaries and dictionaries, extends the relations and makes this knowledge computer-usable in order to allow intelligent applications. More advanced representations may formalize complex legal rules and conceptual structures. Ontologies [7] 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.

Legal knowledge representation remains the most important and challenging task of legal ontologies [8]. The frame-based ontology FBO of [9] and [10] as well as the functional ontology FOLaw [11] can still be considered as important work on formalisation. More advanced work consists in the in the development of a core legal ontology called LRI-Core [12] or the impressive standard for the development of a legal ontology called LKIF Core Ontology (Legal Knowledge Interchange Format) [13].

Quite many projects were focused on conceptual information retrieval (see e.g. Iuriservice [14], LOIS (Lexical Ontologies for legal Information Serving) project [15]. The Legal Taxonomy Syllabus [16], DALOS [17] or the Comprehensive Legal Ontology (CLO) [3].

Such powerful ontologies can only be built if resources of robust NLP and machine-learning are exploited. We share the view of [18] that such technologies are "the key to any attempt to successfully face what we termed the acquisition paradox". However, we argue that the quite huge experience of semi-automatic text analysis and conceptual indexing (see e.g. the projects KONTERM/LabelSOM/GHSOM [19, 20], SALOMON [21], FLEXICON [22], SMILE [23], or Support Vector Machines [24]) should be taken into account and reused.

The automated linking of documents constitutes the most advanced work in semantic indexing (e.g. AustLII [25], CiteSeer [26]). It has to be noted that the task is easier due to more formalized language and a controlled vocabulary.

3 Idea and Method

This method adds the idea of an ontological workbench for the lawyer to the already existing tools. A combination of expert knowledge, easy access to intellectual input and

the use of semiautomatic refinement empowers this method for moving on concerning the "scaling up"-problem. It should be noted that legal writing consists to a large degree in structuring, refining and representing legal text corpora in an abridged and more abstract way. So far, legal writers are still not much in favor of semiautomatic analysis as a tool for improving efficiency to a very time-consuming process. The core of our approach consists in re-use of existing legal materials and intellectual input, corpus-based methods of verification and refinement as well as text categorization, conceptual analysis and text extraction. Using our expertise of as a lawyer with extensive practice and an academic in legal informatics, we will develop a workbench for other lawyers for NLP techniques and text analysis.

Due to our corpus-based approach on legal analysis, all tentative results have to checked against a legal text corpus. In our case, the millions of documents of the Austrian legal retrieval system RIS (Rechtsinformationssystem des Bundes) [27] and related private databases RDB and LexisNexis are used for improvement, refinement and verification of the ontological representation.

As a start, we will describe a sketchy picture for a sufficient granularity of an ontological representation of a jurisdiction: about 10 000 thesaurus entries, 5 000 citations, up to 200 document types, a classification structure, 100 text extraction and summarization rules, and, as representation of the dynamic legal electronic commentary, an indefinite number of concepts, rules and procedures. It is an enormous body of knowledge and it should be clear that a stepwise approach has to be taken for higher representations.

Such extensive meta data has to be maintained in a database with different types of knowledge units (or tables):

Thesaurus entries: header, definition (with sources), examples (with sources), relations (synonym, homonym, polysem, hyponym, hyperonym, antonym etc.), classification, other information.

Citations: header, identification (abbreviation or number), synonyms, classification, author, other information.

Document types: header, identification (abbreviation), use, format, other information.

Classification: header, code, definition, relations, other information.

Extraction and summarization rules: header, rule, definition, relations, other information.

Concepts: header, definition (with sources), related thesaurus entries and citations, relations (synonym, homonym, polysem, hyponym, hyperonym, antonym etc.), classification, legal conceptual structure (ontological model), other information.

Rules: header, quasi-logical expression, source, type, classification, legal conceptual structure (ontological model), other information.

Procedures: header, decision tree, source, type, classification, legal conceptual structure (ontological model), other information.

This information can be built and updated only stepwise: for a start, only a list of thesaurus entries and citations is necessary. It is obvious that the very end, the dynamic legal electronic commentary, will take some time to finish.

For the start, we have collected vast information on legal meta knowledge from traditional sources, e.g. concepts, citations, documents and text extraction. Concept lists were taken from the table of contents and indices of text books and commentaries. A quite complete citation list was provided by the Federal High Court of Administration. The list of document types was established using the on-line information on documents in the legal information system RIS. Text extraction rules were intellectually created

by studying the linguistic styles and patterns of Austrian laws, judgments and literature. We took also advantage of the existing experience with the LOIS project. With this method, it was quite easy to achieve a sufficient but still rough representation of conceptual structure of the Austrian legal order. For easier re-use, this information was incorporated in a relational database. It should also be evident that data entries may be quite incomplete at the beginning, e.g. consisting only of the header. An XML representation is also available for later incorporation in higher representations, e.g. the knowledge base of the dynamic electronic legal commentary.

As tools of semi-automatic analysis, we have implemented the modified GHSOM method of classification, the KONTERM conceptual analysis, and the GATE methods of ANNIE and JAPE [28].

The modified GHSOM method is based on the self-organising map, a general unsupervised tool for ordering high-dimensional data in such a way that alike input items are mapped close to each other. In order to use the self-organising map to explore text documents, we represent the various texts as the histogram of its words with a *TFxIDF* vector representation. The methods LabelSOM can properly describe the common similarities of the cluster. An extension to the SOM architecture, the GHSOM [20] can automatically represent the inherent hierarchical structure of the documents. An extension for legal purposes allows the manual refinement of vector weights of the documents with data enrichment tools. The produced output consists in structured maps of clusters with cluster descriptions. These descriptions were used for refinement of the thesaurus, in particular concerning synonyms.

The KONTERM method [3] produces structured lists of term occurrences. The context of a term is used for describing and analyzing the various meanings. These representations were incorporated in the description of homonyms and polysems of thesaurus entries.

The GATE JAPE tool (Regular Expressions Over Annotations) is implemented for a similar purpose. It is much more powerful in bigger text environments but does not allow so sophisticated representations of meanings as the KONTERM method.

The GATE ANNIE (A Nearly New Information Extraction System) tool is very helpful for a more detailed analysis: segmentation of documents (tokenizer), words, gazetteer, sentence splitter and semantic tagger.

One has still to argue that such methods are helpful for understanding and analyzing a legal domain if the vast amount of text and analysis data is considered. It is also evident that these tools are by far not sufficiently adapted for a legal environment. However, in the hands of an expert, such information proves to be very helpful and of similar use that manual research results.

4 Implementation Details and Problems

The documents were taken from the Austrian legal information system RIS. From the Austrian Administrative Court, we got a list of about 5000 citations. The thesaurus entries were scanned from the respective lists of indices of a selection of representative text books. This “rough” ontology was checked and refined with selective document corpora of the Austrian legal information system RIS using GHSOM, KONTERM and GATE tools. For easier checking of results, subfields like telecommunications law or state aid law were selected. The output was then used for extension and enlargement of the knowledge representation.

The work is still ongoing but some preliminary remarks can be made. The output is very helpful for any further analysis of the materials. Analysis is much improved by faster browsing, reading and text extraction. However, the workload of checking the output is enormous and requires substantial resources. It seems that such efforts can only be justified if other knowledge products like handbooks or commentaries are produced. In the very end the success of this method depends mostly on acceptance by legal authors in their analytical work. It has to be noted that such a product is also very helpful for semantic search methods.

5 Conclusions and Future Work

Next steps for a dynamic electronic legal commentary require semantic indexing of legal information systems and extraction of ontological information of these huge data warehouses. A combination of intellectual input, NLP tools and appropriate ontological representation may overcome the existing bottleneck of legal knowledge acquisition of legal ontologies. Preliminary results on the example of Austrian law using the tools of SOM/GHSOM, KONTERM and GATE show the feasibility of this method. However, refinement and adaptation still require important personal resources in practice. It remains to be seen if a sufficient number of legal writers will modify working methods and include this approach in their tool list of legal structural analysis.

References

1. Index 2006, Rechtsprechung und Schrifttum, Jahresübersicht 2006. Band 59, Begründet von Franz Hohenecker. Manz, Wien (2007)
2. Jurivoc. Dreisprachiger Thesaurus des Schweizerischen Bundesgerichts. `http://www.bger.ch/de/index/juridiction/jurisdiction-inherittemplate/jurisdiction-jurivoc-home.htm` (2009).
3. Schweighofer, E.: Computing Law: From Legal Information Systems to Dynamic Legal Electronic Commentaries, In: Magnusson Sjöberg, C., Wahlgren, P. (eds.), Festskrift till Peter Seipel pp. 569-588. Norsteds Juridik AB, Stockholm (2006).
4. Berners-Lee, T. et al.: The Semantic Web. Scientific American Vol. 284, No. 5, 34-53 (2001).
5. Blair, D. C., Maron, M. E.: An Evaluation of Retrieval Effectiveness for a Full-text Document-retrieval System. Comm ACM, Vol. 28, 289-299 (1985).
6. ISO: Documentation. Guidelines for the establishment and development of monolingual thesauri, ISO 2788 (1986).
7. Gruber, T.R.: A Translation Approach to Portable Ontology Specifications. Knowledge Acquisition vol. 5/2, 199-220 (1993).
8. Bench-Capon, T.J.M., Visser, P.R.S.: Ontologies in Legal Information Systems: The Need for Explicit Specifications of Domain Conceptualisations. In: Proceedings of the 6th ICAIL, pp. 132-141. ACM Press, New York, NY (1997).
9. Kralingen, R.W. van: Frame-based Conceptual Models of Staute Law. Ph.D. Thesis, University of Leiden, The Hague (1995).
10. Visser, P.R.S.: Knowledge Specification for Multiple Legal Tasks: A Case Study of the Interaction Problem in the Legal Domain. Computer Law Series Vol. 17, Kluwer Law International, The Hague (1995).
11. Valente, A.: Legal knowledge engineering: A modelling approach. IOS Press, Amsterdam (1995).

12. Breuker, J. and Hoekstra, R.: DIRECT: Ontology-based Discovery of Responsibility and Causality in Legal Case Descriptions. In: Proceedings of the 17th JURIX. IOS Press, Amsterdam et al. (2004).
13. Hoekstra, R., Breuker, J., De Bello, M., Boer, A.: The LKIF Core Ontology of Basic Legal Concepts. In: Casanovas, P., Biasiotti, M. A., Francesconi, E., Sagri, M. T. (eds.) Proceedings of LOAIT 07, II. Workshop on Legal Ontologies and Artificial Intelligence Techniques, pp. 43-64. http://www.ittig.cnr.it/loait/LOAIT07Proceedings.pdf (2007).
14. Casellas, N., Casanovas, P., Vallbé, J.-J., Poblet, M., Blázquez, M., Contreras, J., López-Cobo, J.-M., Benjamins, R.: Semantic Enhancement for Legal Information Retrieval: IURISERVICE performance. In: Eleventh International Conference on Artificial Intelligence and Law, pp. 49-57. ACM Press, New York (2007).
15. Dini, L., Liebwald, D., Mommers, L., Peters, W., Schweighofer, E., Voermans, W.: LOIS Cross-lingual Legal Information Retrieval Using a WordNet Architecture. In: Proc Tenth Int Conf on Artificial Intelligence & Law, pp. 163-167. ACM Press, New York (2005).
16. Ajani, G., Lesmo, L., Boella, G., Mazzei, A., Rossi, P.: Terminological and Ontological Analysis of European Directives: multilinguism in Law. In: Eleventh International Conference on Artificial Intelligence and Law, pp. 43-48. ACM Press, New York (2007).
17. Francesconi, E., Spinosa, P., Tiscorina, D.: A linguistic-ontological support for multilingual legislative drafting: the DALOS Project. In: Casanovas, P., Biasiotti, M. A., Francesconi, E., Sagri, M. T. (eds.) Proceedings of LOAIT 07, II. Workshop on Legal Ontologies and Artificial Intelligence Techniques, pp. 103-112. http://www.ittig.cnr.it/loait/LOAIT07-Proceedings.pdf (2007).
18. Lenci, A., Montemagni, S., Pirrelli, V., Ventur, G.: NLP-based ontology learning from legal texts. A case Study, In: Casanovas, P., Biasiotti, M. A., Francesconi, E., Sagri, M. T. (eds.) Proceedings of LOAIT 07, II. Workshop on Legal Ontologies and Artificial Intelligence Techniques, pp. 103-112. http://www.ittig.cnr.it/loait/LOAIT07Proceedings.pdf (2007).
19. Schweighofer, E.: Legal Knowledge Representation, Automatic Text Analysis in Public International and European Law. Kluwer Law International, The Hague (1999).
20. Schweighofer, E. et al.: Improvement of Vector Representation of Legal Documents with Legal Ontologies. In: Proceedings of the 5th BIS, Poznan University of Economics Press, Poznan (2002).
21. Moens, M.-F. et al.: Abstracting of Legal Cases: The SALOMON Experience. In: Proceedings of the 6th ICAIL pp. 114-122. ACM Press, New York (1997).
22. Smith, J.C. et al.: Artificial Intelligence and Legal Discourse: The Flexlaw Legal Text Management System. Artificial Intelligence and Law Vol. 3/1-2, 55-95 (1995).
23. Brüninghaus, S. and Ashley, K.D.: Improving the Representation of Legal Case Texts with Information extraction Methods. In: Proceedings of the 8th ICAIL pp. 42-51. ACM Press, New York (2001).
24. Gonçalves, T. and Quaresma, P.: Is linguistic information relevant for the classification of legal texts? In: Proceedings of the 10th ICAIL, pp. 168-176. ACM Press, New York (1995).
25. AustLII website. http://www.austlii.edu.au.
26. CiteSeer website. http://citeseer.ist.psu.edu/cs.
27. RIS website. http://www.ris.bka.gv.at.
28. GATE (General Architecture for Text Engineering) Engineering) website. http://gate.ac.uk/.

Enriching Thesauri with Ontological Information: Eurovoc Thesaurus and DALOS Domain Ontology of Consumer Law

Maria Angela Biasiotti[1] and Meritxell Fernández-Barrera[2]

[1] CNR-ITTIG, Via dei Barucci, 20,
50127 Florence (Italy)
biasiotti@ittig.cnr.it,
[2] European University Institute, Via dei Roccettini, 9,
I-50014 San Domenico di Fiesole (FI) Italy
Meritxell.Fernandez@EUI.eu

Abstract. This paper analyses the semantic shortcomings of thesauri in comparison with ontologies in the framework of the trend to building KOS that enable IR by concept search instead of textual search. A particular case study in the domain of the consumer law is presented, in which the differences in terms of semantic depth between the Eurovoc thesaurus and the DALOS ontology are analysed. Moreover the paper analyses the existing technical solutions for semantically enriching thesauri, and explores which would be the possibilities in the case of the Eurovoc thesaurus taking into account that a great number of documents have already been indexed with its descriptors.
Key words:
Thesauri, Ontologies, Legal semantics, Conceptual Information Retrieval

1 Framework

The purpose of the Semantic Web approach is to make web content machine-processable in order to develop new functionalities beyond the mere display of data, such as enabling a better access to relevant information contained in web documents. One of the main problems of the WWW is information overload [8] and the limited software performance in extracting useful information for users. It is thus desirable to develop improved techniques for accessing quickly not only relevant documents, but the bits of information embedded in them that specifically match the user queries.

The paper presents an analysis of the different semantic depth in which a specific legal domain (the consumer law domain) is represented by different KOS: the Eurovoc thesaurus and the DALOS domain ontology of consumer law. Given the shortcomings of Eurovoc and after having reviewed the state of the art in semantic enrichment of thesauri, a transition procedure to support the shift from a traditional KOS, like Eurovoc, towards a full-fledged and semantically rich KOS is suggested.

Before analysing the two different representation models adopted by Eurovoc and by Dalos in the Consumer Protection field we will briefly outline which are the main characters and features of these considered models.

2 Case study. Representation of the consumer law domain

2.1 Knowledge Organization Systems

As to the different metadata structures that can channel the addition of semantics to legal information, we can mention several ways of attaching meaning to the information contained in the web, like catalogs, thesauri and frames.

More specifically and in brief ontologies are controlled vocabularies expressed in an ontology representation language (OWL), whereas taxonomies or semantic nets are a collection of controlled vocabulary terms organized into a hierarchical structure, and thesauri are networked collections of controlled vocabulary terms. Each term in a taxonomy is in one or more parent-child relationships to other terms in the taxonomy, while in a thesaurus associative relationships are used in addition to parent-child relationships. In more detail the latter can be defined as a classification tool to assist libraries, archives or other centres of documentation to manage their records and other information.

This functionality is achieved by establishing paths between terms. The establishment and development of a thesaurus is generally arranged in accordance with the standards of ISO (International Standards Organization), which are especially recognised at international level.

In this context, the main difference between thesauri and ontologies is actually the degree of semantic precision with which they describe contents. On the one hand, a thesaurus embodies a terminological representation of a domain (a particular lexicalisation of a conceptualisation) which is not as semantically complete as the formal conceptual representation provided by an ontology, and its limited structure makes it therefore unsuitable for advanced semantic applications [12]. In particular, the relationships linking the terms (the controlled vocabulary to represent concepts) in a thesaurus (BT, NT, RT) are usually not enough for a deep analysis of the semantics of the indexed documentation. On the other hand, ontologies provide a deeper conceptual representation of the domain (with a richer set of relationships between concepts like *part of*, *instance of*, *role*, among others, depending on the semantic domain) and can therefore enhance better access to the content of specialised documents.

In this framework, the purpose of this paper will be to explore the suitability of different KOS for representing the semantics of a specific legal domain and to analyse how a semantically simpler KOS (a thesaurus) can be enriched. We will show the need to move towards this trend by analysing two different knowledge representation models provided for the consumer protection domain.

2.2 Representation of Consumer Law Domain

The legal domain chosen for our case study is the consumer law domain. It is chosen just as a preliminary study to analyse the semantic depth of Eurovoc as regards legal issues. However, in further work it would be desirable to extend the analysis to other legal domains.

From a textual analysis of the the Consumer law discipline, namely *Consumer Protection*, it arises that relevant concepts to be considered among others are: *Advertising, misleading advertising, commercial communication, Market surveillance, inspection, disclosure, Commercial activity, Contract, selling price, unit price, consumer goods,*

product, raw material, products sold in bulk, agricultural products, finished product, services, all inclusive services, information society services, financial services, producer, buyer, consumer, trader, contract, unfair-terms, credit-agreement and so on.

In particular we will analyse the representation of this particular domain provided by two KOS: (i) the Eurovoc thesaurus and (ii) the Dalos ontology of consumer law.

Case study set-up In this section the specific representation of the consumer protection domain provided by the two considered models will be offered, outlining shortcomings and needs.

The first KOS considered, Eurovoc, has not a proper framework for the consumer protection law within sector Law. It sets it within sector 20 devoted to Trade, in the Micro Thesaurus Consumption. Therefore the consumer protection is just the NT of the top term consumer and has some RT relationship with other relevant concepts such as *advertising, producer's liability, publishing of prices* and so on.

The *consumer* descriptor has then hierarchical relationships with other relevant terms such as *consumer information, European consumer information agency, consumer movement, product quality, product designation, product life, product safety, defective product.*

From the EUROVOC scenario it emerges that the description provided by Eurovoc of *Consumer* has few constraints with respect to the Consumer law protection. In fact, Eurovoc pertains to the category of traditional thesauri, structured on hierarchical and synonymy relations, with a rigid structure and inter-lingual relations but a poor semantics. Its scope is broad (European policy issues), and the components devoted to the normative domain are very weak in precision and granularity. As the focus is on socio-economic issues, depth in law is quite low and the structure is not appropriate to EU law. One of its main limitations is therefore that it is not suitable for the indexation and the search of specialised documents.

It is indeed quite flat and unconsistent with respect to the domain: the consumer is a top term and has only some hierarchical relationship with other descriptors such as *consumer information, product safety*, and a few others.

Generally speaking the Eurovoc model presents:

- *lack of expressivity and granularity*: the concepts considered are very few with respect to those emerging form the legal sources (for instance, a non-descriptor is foreseen for *consumers rights*, but no descriptor exists to refer to a specific instance of this concept, like *withdrawal right*);
- *lack of semantics*: the thesaurus relationships are semantically overloaded, in the sense that the same relationship is used to express different semantic links where three different possible ontological relations have been identified for the relation NT, and two different relations for RT.
- *lack of legal orientation*: relationships are fixed and inexpressive of the domain relationships and meaning. They are limited to BT, NT, RT, UF and are therefore not specific to the domain. For instance, the link between *Consumer and consumers right* is firstly, not direct, since *consumers right* is a non-descriptor for *consumer protection*, which is a NT of Consumer. Secondly, the relation is not specific to the domain, it is just a generic hierarchical relationship, whereas in a deeper semantic model the representation could be that *Consumer has-right-towards* someone else.

As Eurovoc was drafted exclusively for manual indexing and retrieval purposes, the lack of semantic precision generates frequent inconsistencies among several hierarchical

and synonymy relations so that it is mainly suitable for retrieving related terms. In the same way as all existing thesauri it is focused on documentation and lacks sufficient granularity for semantic access to EU law.

The ontology of consumer law developed in the frame of the DALOS project[3] provides a formal representation of the classes of the world entities involved in the domain of consumer law (*agents, actions, legal roles, legal effects*).

Relevant concepts are all present as captured directly from the legal sources by the NLP techniques. They are well identified into five classes Agent, Quality, Region, Event and Object and linked to each other by some significant relationships. The contextualization provided by the Dalos domain ontology is consistent with respect to the domain since:

- there is a higher conceptual expressivity and granularity, as proven by the big number of domain specific concepts: for instance *withdrawal-right; damage-compensation; consumer-complaint.*
- relationships between concepts have a legal orientation (are specific to the domain).
- and relationships are not semantically overloaded, that is, a semantic relation is not used with more than one meaning, unlike RT and NT in Eurovoc.

Comparing the two different approaches in representing the other relevant concepts of the Consumer protection law, such as advertising, we obtain the following scenario:

The two models considered produce two different representations as to granularity, expressiveness, constraint and consistency.

Descriptors in Eurovoc are not consistent with respect to the specific domain as they are identified independently for the concepts embedded in the domain itself. They are identified according to a top-down approach by an expert who considers those terms relevant for describing the domain without any reference to the legal texts. Whereas the Dalos ontology has been built according to a bottom-up approach which enables experts to take into due consideration the richness, the concepts and relations arising directly from the text, that are the sources of law ruling on the Consumer protection field.

Indeed Eurovoc pretends to describe the consumer protection domain with around 23 terms, whereas the DALOS ontology aims at doing the same with over 100 concepts.

Eurovoc shortcomings as regards the representation of the consumer law domain could thus be tackled by enriching the simplified conceptualisation represented by its terminology with deeper conceptual relations specific to the domain. In the following section we will briefly review some technical solutions for designing a more complete semantic model building on a pre-existing semantic resource.

3 Techniques for enriching semantically lexical resources

From an analysis of the literature dealing with the enrichment of metadata resources with a deeper semantics it is possible to identify different techniques. On the one hand, those that rely on the combination of pre-existing resources and on the other hand those that simply restructure a pre-existing resource.

The techniques belonging to the first trend (highlighted in [9]) maintain the autonomy of both resources and therefore avoid an actual *merging*. They are the following:

[3] `http://www.dalosproject.eu/`

- *restructuring* a computational lexicon following ontological principles, focusing thus on the lexical resource as the final output and using the ontology merely as a guiding tool. The resulting lexical resource respects certain ontological restrictions but does not become a full-fledged ontology.
  Some of the suggestions for a better semantics of lexical resources include: making a proper use of the is-a link so that it expresses not only lexical relations but ontological ones (-this would amount to using the is-a relation only to link entities that share similar identity criteria-); avoiding the confusion between concepts and instances, avoiding the subsumption of types by roles as derived from the ONTOCLEAN methodology- and not mixing different levels of generality, among others [5].
- *populating* an ontology with lexical information, which amounts to mapping lexical units to ontological entries;
- *aligning* an ontology with a lexical resource, that is, combining the restructuration of the computational lexicon according to ontological-driven principles and its mapping with the ontological resource.

With regard to the second trend, we can mention those techniques that simply draw on a single pre-existing resource, namely a thesaurus and exploit the possibility of transforming it into a more complex knowledge representation structure, that is, into an ontology. The main difference from the previous techniques is that in this case the input consists just of one initial semantic resource (the thesaurus) and that the nature of the resulting resource changes: the thesaurus not only gains in semantic precision and structure, but it becomes a full-fledged ontology, that is, a rigid semantic representation of the domain with a hierarchy of classes and corresponding slots or properties that increases in constraint density[4] (a higher set of relationships linking its terms). Several works have shown concerns on the connection between thesauri and ontologies ([1]; [6]) and discussed their different semantic scope. A general perspective on the possibility of converting thesauri into ontologies is given by [13] and specific examples of thesaurus reengineering are [10, 4].

4 Applicability of existing techniques to eurovoc: approaches for improvement

As to the techniques that could be applied to achieve our goals, several possibilities arise.

4.1 Mapping of EUROVOC to pre-existing legal ontologies

The first one, corresponds to the mapping of lexical entries and ontological classes (populating the classes of the ontology with the thesaurus terms). This may in some cases be feasible, where in Eurovoc there exists a term *advertising* and in Dalos ontology a concept with the same name. In this case an equivalence relationship could be added to the Dalos ontology to link the Dalos *advertising* concept and corresponding term in Eurovoc, creating a direct link between the two resources. However, in some other cases

[4] The notion of constraint density is introduced by [9] as the *density of the "network of constraints" that holds between the concepts.*

populating the ontology with Eurovoc classes might be more difficult, where Eurovoc has no corresponding term for the Dalos concept *product*.

Indeed, since Eurovoc is a general thesaurus it therefore does not match the semantic requirements of a specialised domain such as the consumer law, and there will be cases in which it will not be possible to find an Eurovoc term corresponding to a class of the DALOS ontology. This is why it will be necessary, in the process of mapping the ontology to the thesaurus, to enrich the semantics of Eurovoc adding more terms. This would amount to making some kind of restructuring of the thesaurus, but in this respect, we have to take into account an important restriction, namely, that Eurovoc is currently used to index documents and in order not to loose this indexation it is necessary to maintain current Eurovoc terms. This imposes limitations on a possible restructuration of the thesaurus, for it will not be possible to delete any descriptors even if semantic consistency might require to do so in some cases.

A further difficulty of following the approach of using the thesaurus to populate a pre-existing ontology is that it would be necessary to identify ontologies for several specialised domains of the law, if the whole structure of Eurovoc referring to the legal field wants to be expanded semantically. Dalos would be the option for the consumer protection domain but it might turn up more difficult to find ontologies of other specific legal areas, such as international law or environmental law.

4.2 Reengineering Eurovoc into a formal ontology

A different option would be to use merely the Eurovoc thesaurus and transform it into a full-fledged ontology. This would require providing the thesaurus with a well structured semantics and to represent it in a highly expressive language like OWL. In order to do that, it would be necessary to:

- increase specificity and granularity by adding more classes corresponding to the legal domain;
- refining the relationships existing between the different terms of the thesaurus (BT, NT, RT, UF, USE) according to the semantics of the domain (adding therefore other types of relations);
- checking ontological consistency (ontological constraints).

5 Results

The proposed approach tries to meet the new trend arising from the Semantic Web towards the development of legal KOS able to allow the user to search by concept instead of searching by words. This implies the use of tools able to expand the query from a semantic point of view, meaning that the concept identified is surrounded by other concepts semantically linked to it. In this framework, the paper has analysed the semantic scope of a traditional KOS (Eurovoc) used for the indexation and retrieval of legal information in the EU institutions and national parliaments.

The paper has provided some concrete evidence on the semantic shortcomings of the Eurovoc thesaurus for the representation of the legal domain by analysing how it represents the particular domain of the consumer protection law in comparison to DALOS domain ontology. The main findings are, firstly, that in Eurovoc there is a lack of semantic granularity, since many relevant concepts of the domain are not represented by its descriptors; relations linking terms are semantically overloaded and therefore only

shallowly expressive; and relations are not specific of the legal domain, but generic (RT, BT, NT, UF). Taking into account that currently one of the main functionalities of the Eurovoc thesaurus is to be used as an indexing and searching tool of legal documentation (it is for instance used by Eur-lex, the gateway for accessing European Union law), it arises the need of equipping the thesaurus with more powerful conceptual structures specific to the legal domain in order to improve search and legal information retrieval.

Secondly, the paper assesses which are the technical possibilities, according to the state of the art, for enriching semantically the Eurovoc thesaurus. Two methods are highlighted as feasible solutions: on the one hand, using Eurovoc to populate pre-existing ontologies on specific legal domains; on the other hand, transforming Eurovoc into a full-fledged ontology by enriching its conceptual structure and expressing it using formal representation languages.

6 Further work

The paper presents some preliminary conclusions as to possible directions to solve the problem of the semantic limitations of a current indexing and retrieval tool used at EU and national level for legal documentation. However, further research is foreseen in order to implement the project:

- A whole assessment of the semantic representation of the legal domain by Eurovoc: building on the case study presented in this paper that analyses the representation of the domain of the consumer law, further analysis of Eurovoc semantic representation of other legal domains is required.
- Analysis of the benefits that the proposed approach would bring about: run experiments to measure the degree of improvement of information retrieval tasks by the use of an ontological structure instead of Eurovoc structure and analyse the benefits for the various users of Eurovoc (EU institutions, national parliaments, private users with licence).
- Analysis of the costs of implementing the approach: in terms firstly, of the KOS reengineering costs: and secondly, the adaptation of current information systems to the new KOS.

References

1. Arano, S.: Thesauruses and ontologies [on line]. Hipertext.net, 3, (2005) <http://www.hipertext.net> [Consulted: 07/01/09]. ISSN 1695-5498
2. Benjamins, Casanovas, P., Gangemi, A., Selic, B. (Eds.): Law and the Semantic Web. Legal Ontologies, Methodologies, Legal Information Retrieval, and Applications. Springer Verlag. Berlin Heidelberg 5, (2005)
3. Coulthard, Malcolm and Johnson, Alison: An Introduction to Forensic Linguistics. Language in Evidence. London and New York Routledge, (2007)
4. Cross, P., Brickley, D., Koch, T.: Conceptual relationships for encoding thesauri, classification systems and organised metadata collections and a proposal for encoding a core set of thesaurus relationships using an RDF Schema. Available in: http://www.desire.org/results/discovery/rdfthesschema.html, (2000)

5. Gangemi, A., Guarino N., Oltramari, A.: Conceptual Analysis of Lexical Taxonomies: The Case of WordNet Top Level In Formal Ontology in Information Systems. Proceedings of FOIS2001, eds. C. Welty and S. Barry, 285-296. New York: Association of Computing Machinery, (2001)
6. García Jiménez, A.: Instrumentos de representación del conocimiento: tesauros versus ontologías. Anales de Documentación, 7, 79–95, (2004)
7. Hirst, D.: Ontology and the Lexicon Handbook on Ontologies. Information Systems. Springer, (2003)
8. Lazonder et al.: Differences between Novice and Experienced Users in Searching Information on the World Wide Web. Journal of the American Society for Information Science, 51(6), 576–581, (2000)
9. Oltramari, Prévot, Borgo: Theoretical and practical aspects of interfacing ontologies and lexical resources. Proc. of the 2nd Italian Semantic Web workshop SWAP 2005 (Semantic Web Applications and Perspectives), Trento, (2005)
10. Qin, J., Paling, S.: Converting a controlled vocabulary into an ontology: the case of GEM. Information Research, 6, 2. Available in: `http://informationr.net/ir/6-2/paper94.html`, (2000-01)
11. Sartor, G.: Legislative information and the web. Biasiotti et al.: Legal Information Management of Legislative Documents, (2008)
12. Soergel, D., Lauser, B., Liang, A., Fisseha, F., Keizer, J, Katz, S.: Reengineering thesauri for new applications: the AGROVOC example. Journal of Digital Information, 4(4), (2004)
13. Wilson, M.: Migrating from Thesauri to Ontologies. Available in: `http://www.w3c.rl.ac.uk/ukofficepasttalksindex.html`, (2002)

Author Index