LAW, ONTOLOGIES AND THE SEMANTIC WEB
Frontiers in Artificial Intelligence and Applications

FAIA covers all aspects of theoretical and applied artificial intelligence research in the form of monographs, doctoral dissertations, textbooks, handbooks and proceedings volumes. The FAIA series contains several sub-series, including “Information Modelling and Knowledge Bases” and “Knowledge-Based Intelligent Engineering Systems”. It also includes the biennial ECAI, the European Conference on Artificial Intelligence, proceedings volumes, and other ECCAI – the European Coordinating Committee on Artificial Intelligence – sponsored publications. An editorial panel of internationally well-known scholars is appointed to provide a high quality selection.

Series Editors:

Volume 188
Recently published in this series

Vol. 186. G. Lambert-Torres et al. (Eds.), Advances in Technological Applications of Logical and Intelligent Systems – Selected Papers from the Sixth Congress on Logic Applied to Technology
Vol. 185. A. Biere et al. (Eds.), Handbook of Satisfiability
Vol. 184. T. Alsinet, J. Puyol-Gruart and C. Torras (Eds.), Artificial Intelligence Research and Development – Proceedings of the 11th International Conference of the Catalan Association for Artificial Intelligence
Vol. 179. A. Cesta and N. Fakotakis (Eds.), STAIRS 2008 – Proceedings of the Fourth Starting AI Researchers’ Symposium
Vol. 178. M. Ghallab et al. (Eds.), ECAI 2008 – 18th European Conference on Artificial Intelligence
Vol. 177. C. Soares et al. (Eds.), Applications of Data Mining in E-Business and Finance
Vol. 176. P. Zaraté et al. (Eds.), Collaborative Decision Making: Perspectives and Challenges

ISSN 0922-6389
Foreword

This collection of articles about legal ontologies and Semantic Web applications has its roots in workshops and conferences on Artificial Intelligence (AI) and Law, most notably, the workshops on ontologies and Semantic Web technology in the legal domain, held in June 2007 at Stanford University. The domain of law belongs to the early adopters of ontologies and semantic web technology to support its enormous and fast growing demand for effective information management; it is probably only surpassed in this respect by the bio-sciences. Having easily access to relevant legal information among the rising flood of legal documentation is not only the concern of legal practitioners, but also the life and work of citizens becomes more and more entangled with legal issues.

A first monograph on legal ontologies was published as early as 1995, also as a volume of this FAIA series by IOS Press [Valente, 1995]. The first workshop on legal ontologies was held in 1997, as part of the biannual conference on AI & Law (ICAIL-1997). This volume reflects the wisdom, abstracted from experiences accumulated over more than a decade of research and development in this area. It contains a representative overview of the state of the art, covering both theoretical aspects and practical systems. The latter has been an important driver of research in this area, which can be observed from the geographical origins: the work discussed is mainly European, due to R & D initiatives of the European Commission in the various Framework programs. Also national initiatives have played an important role as can be concluded from the overrepresentation of articles from Spain and Italy. As the practical needs for legal information management are certainly not more pressing in these countries than in other countries, we hope that this book will also serve as an argument for (further) investments in this endeavor. The Semantic Web is not only an area of research, but also a world wide project where easy to construct applications can directly find their communities of users. However the semantics that are the engines for these applications are still the bottleneck in the development. Therefore one finds in this volume a large range of technologies and tricks to populate ontologies with machine understandable meaning of terms. This varies from the use of top-ontologies via design patterns to extraction of terms from text and alignment of existing terminologies. In fact, one may see this book not so much as the report of results of research, but rather as a specification of the elements of an emerging methodology for developing legal ontologies.

Barcelona, October 2008, the Editors
We would like to thank the following reviewers for this book:

Alexander Boer, Leibniz Center for Law, University of Amsterdam, the Netherlands
Bert Bredeweg, HCSL University of Amsterdam, Netherlands
Tom van Engers, Leibniz Center for Law, University of Amsterdam, the Netherlands
Jérôme Euzenat, INRIA Grenoble Rhône-Alpes, France
Tom Gordon, FOKUS Fraunhofer Institute, Berlin, Germany
Carole Hafner, Northwestern University, Chicago, USA
Rinke Hoekstra, Leibniz Center for Law, University of Amsterdam, the Netherlands
Laura Holink, Department of Computer Science, VU University Amsterdam, the Netherlands
Laurens Mommers, Faculty of Law, Leiden University, the Netherlands
Paulo Quaresma, Universidade de Évora, Portugal
Marta Sabou, Knowledge Media Institute, the Open University, UK
Giovanni Sartor, Faculty of Law, Università di Bologna, Italy
Marco Schorlemmer, Artificial Intelligence Research Institute, CSIC, Barcelona, Spain
Ronny Siebes, Department of Computer Science, VU University Amsterdam, the Netherlands
Elisabeth Uijttenbroek, Faculty of Law, VU University Amsterdam, the Netherlands
Denny Vrandecic, AIFB, Universität Karlsruhe, Germany
Matthew West, Shell / University of Leeds, UK
Radboud Winkels, Leibniz Center for Law, University of Amsterdam, the Netherlands
Adam Wyner, The University of Liverpool, UK

References

Contents

Foreword v

Introduction
The Flood, the Channels and the Dykes: Managing Legal Information in a Globalized and Digital World 3
  Joost Breuker, Pompeu Casanovas, Michel C.A. Klein and Enrico Francesconi

Legal Knowledge Modelling and Acquisition

Legal Ontology Construction
LKIF Core: Principled Ontology Development for the Legal Domain 21
  Rinke Hoekstra, Joost Breuker, Marcello Di Bello and Alexander Boer
Introducing Pattern-Based Design for Legal Ontologies 53
  Aldo Gangemi

Information Extraction of Legal Texts
Ontology Learning from Italian Legal Texts 75
  Alessandro Lenci, Simonetta Montemagni, Vito Pirrelli and Giulia Venturi
Definitions in Court Decisions – Automatic Extraction and Ontology Acquisition 95
  Stephan Walter and Manfred Pinkal
Automatic Argumentation Detection and Its Role in Law and the Semantic Web 115
  Raquel Mochales Palau and Marie-Francine Moens

Legal Knowledge Applications

Applying Semantic Web Technologies
Matching Law Ontologies Using an Extended Argumentation Framework Based on Confidence Degrees 133
  Cassia Trojahn, Paulo Quaresma and Renata Vieira
Copyright Licenses Reasoning Using an OWL-DL Ontology 145
  Roberto Garcia and Rosa Gil

Integrated Applications
An Ontology-Based Decision Support System for Judges 165
  Pompeu Casanovas, Núria Casellas and Joan-Josep Vallbé
A Two-Level Knowledge Approach to Support Multilingual Legislative Drafting
T. Agnoloni, L. Bacci, E. Francesconi, W. Peters, S. Montemagni and G. Venturi

The e-Sentencias Prototype: A Procedural Ontology for Legal Multimedia Applications in the Spanish Civil Courts
Pompeu Casanovas, Xavier Binefa, Ciro Gracia, Emma Teodoro, Núria Galera, Mercedes Blázquez, Marta Poblet, Jordi Carrabina, Màrius Monton, Carlos Montero, Javier Serrano and José Manuel López-Cobo

Authors’ Biographies

Subject Index

Author Index
Introduction
This page intentionally left blank
The Flood, the Channels and the Dykes: Managing Legal Information in a Globalized and Digital World

Joost BREUKER\textsuperscript{a1}, Pompeu CASANOVAS\textsuperscript{b}, Michel C.A. KLEIN\textsuperscript{c}, Enrico FRANCESCONI\textsuperscript{d}

\textsuperscript{a}Leibniz Center for Law (ICL), University of Amsterdam, The Netherlands
\textsuperscript{b}Institute of Law and Technology (IDT), Universitat Autònoma de Barcelona, Spain
\textsuperscript{c}Department of Computer Science, VU University Amsterdam, the Nederlands
\textsuperscript{d}Institute of Legal Information Theory and Techniques, Italian National Research Council (ITTIG-CNR), Florence, Italy

Abstract: Information search and retrieval are part of daily routines of the legal profession. Lawyers, judges, prosecutors, and legal clerks usually access a number of electronic resources to browse, search, select, or update legal contents. Legal databases have currently become large digital libraries where the tasks related to information-seeking may sometimes be cumbersome. Adding semantics to support information search may provide significant results in terms of efficiency, efficacy, and user satisfaction. Semantic technologies may be able to improve legal information search in the judicial and lawyers’ domains. However, legal professionals sometimes prefer following routines than changing their information search behavior. New trends in legal ontologies and Semantic Web technologies may help to improve both professional and laymen’s skills.

Keywords: Information overload, legal ontologies, Semantic Web, information retrieval, lawyers, globalization, e-court, e-discovery, e-governance, e-administration

1. The flood: legal information overload

There is a paradoxical situation in the modern world: although there is an overabundance of available information, it is often difficult to obtain relevant information when it is needed [21]. In addition, researchers in organization and knowledge management have found that the quality and efficiency of decision-making vary with the amount of information people are exposed to. The performance correlates positively with the amount of received information, but up to a certain point. If further information is provided beyond this point, the performance of the individual will rapidly decline [22].

\textsuperscript{1} Corresponding author, breuker@science.uva.nl
These kind of related phenomena have dysfunctional consequences such as stress or anxiety, and have been termed in several ways in the literature: cognitive overload, sensory overload, communication overload, knowledge overload, information fatigue syndrome, data smog or analysis paralysis [21, 22]. However, the most common and generic term is information overload. Information overload “occurs when information received becomes a hindrance rather than a help when the information is potentially useful” [5, 30].

Information overload is not a new phenomenon. As early as 1545 Conrad Gesner complained in his *Bibliotheca Universalis* (Zurich, 1545) of “that confusing and harming abundance of books” and Adrien Baillet wrote in the *Jugemens des scavans sur les principaux ouvrages des auteurs* (Paris, 1685, quoted by [10]):

*We have reason to fear that the multitude of books which grows every day in a prodigious fashion will make the following centuries fall into a state as barbarous as that of the centuries that followed the fall of the Roman Empire.*

This excess of information is also prominently present in the legal domain, which started already centuries ago. In Europe, *la raison écrite* that led to the development of Civil law, and legislation and case law records in the Common law, produced a massive amount of documents since the 16th century. It seems that what Jack Goody [28] called the “domestication of savage mind” —the modern process of thought through technological means— had indeed a strong legal side. At the end of the eighteenth century the English Parliament had quadrupled its legislative output. The Houses of Commons of George III (1760-1820) legislated four times more than those of Whilhelm III (1689-1702) [35, 49]. In USA, the growth of case law was fast (from 18 volumes of legal reports in 1810 to nearly 3800 in 1885). Therefore, at the end of the nineteenth century the number of caseloads had increased by forty since the beginning of the century [21].

Nowadays, the amount of legal information grows even further because of the ongoing legalization of the society. On the one hand, much new technology requires new and specific law, e.g. around online purchases, security and data protection. On the other hand, law is becoming a suitable application domain for technological developments, as technology can be used to automatically enforce law or act upon it. For example, in case of *digital rights management* the technology guarantees the respect of the copyrights, while the technology in modern *traffic toll systems* automatically charges the driver. In spite of the differences between legal cultures, this creates an increasingly *legalized* society. Today, the legal database of the Publication Office of the European Union, EUR-Lex, contains 1.800.000 documents in 22 European languages. The average number of visits per working day at mid 2006 were 170.000 [8], at mid 2008 approximates 175.000.2

The growth of the legal profession is another factor that contributes to flood of the legal information. Especially law firms in the USA and Europe have experienced a permanent growth since the seventies. Table 1 and Figure 1 may show how the

---

2 We thank Pascale Berteloot for this updated information.
Table 1. Legal markets size. Source: Euromonitor.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>USA</td>
<td>140.3 billion $</td>
<td>5.6 percent</td>
<td>174.1 billion $</td>
</tr>
<tr>
<td>UK</td>
<td>28</td>
<td>3.6</td>
<td>31.6</td>
</tr>
<tr>
<td>France</td>
<td>14.7</td>
<td>14</td>
<td>16.7</td>
</tr>
<tr>
<td>Australia</td>
<td>5.9</td>
<td>11.6</td>
<td>9.4</td>
</tr>
<tr>
<td>South Korea</td>
<td>1.35</td>
<td>-3.3</td>
<td>1.9</td>
</tr>
<tr>
<td>China</td>
<td>1.34</td>
<td>7</td>
<td>1.9</td>
</tr>
<tr>
<td>Japan</td>
<td>0.9</td>
<td>8</td>
<td>1.6</td>
</tr>
</tbody>
</table>

situation looks like. The size that legal markets have reached implies that law firms have becoming transnational corporations, competing for revenues and income (Table 1). According to the *The American Lawyer 100 Report*—the top-grossing law firms in the United States—the total revenues reached $64.6 billion in 2007, an increase of 13.6%. Lawyers themselves are starting to wonder whether such numbers are sustainable [44]. Even cultures that are traditionally less-prone to litigation—such as the Netherlands and other North-European countries—have showed an increase in the number of lawyers per 100,000 inhabitants (see Figure 1 for the numbers of 2004). In 2006, there were 833,763 lawyers registered in the national bars of 31 European countries [17].

![Figure 1. Lawyers in Europe per 100,000 inhabitants per country (2004). Source: Council of Europe. 2006.](image)
Eventually, the legal information overload might lead to a reduced access for citizens to the judicial system. If the right information is not available at the right moment for the right person, it hinders the accessibility of justice. The mere fact that it can take too long to find the right information could already be seen as problematic, according to the classical quotation “justice delayed is justice denied” attributed to William Gladstone (1809–1898). In the past, financial constraints hindered the access to justice, in later years time constraints formed a bottle-neck. It should be prevented that the information overload forms a new barrier for the access to justice in the future. Therefore, measures are needed to channel the legal information flood.

2. Channelling: Semantic Web technology and XML

The Semantic Web may be defined as a large scale, heterogeneous collection of formal, machine processable, ontology-based statements (semantic metadata) about web resources and other entities in the world, expressed in a XML-based syntax. As such, Semantic Web technology and ontologies can help to organize the information overload. They play the following roles:

1. giving meaning: ontologies define unambiguously the intended meaning of information;
2. structuring: semantic mark-up together with ontologies give structure to documents.

Together, ontologies and semantic web technology facilitates more precise retrieval of legal documents, partial automatic integration and exchange of information, and to some extent automatic reasoning over problems (e.g. question answering, problem solving).

The benefits of Semantic Web technology for the legal domain is leveraged by the increasing availability of legal data in XML format, as such data can be more easily annotated with semantic information. Currently, the adoption and development of standards for legal information, electronic court filing, court documents, transcripts, criminal justice intelligence systems, etc. has become the core activity of a number of projects in several different countries (LegalXML, LEXML, Norme in Rete, MetaLex, LexDania, CHeXML, eLaw, among other initiatives) [9, 26].

All these institutional and research initiatives derive from the awareness of the peculiarities of legal users' information needs (be them law makers, legal experts or citizens) which are increasingly pushing towards the use of advanced Information Technologies in the legal field. Legal information systems should be aimed not only at providing advanced search and retrieval services to the users, but also at maintaining and up-keeping the legal order, at monitoring the impact of new norms, at handling document timeline and versioning.

Moreover, users are mainly interested in accessing norms rather than simply documents; they are particularly interested in knowing the relations between norms and having support to legal reasoning. In this perspective the annotation of legal documents, in particular legislative ones, with shared document standards is
particularly desirable to describe their well defined structures and to provide them with metadata able to manage production, preservation and workflow, which involve Parliaments and Public Administrations.

The use of document open standards and Semantic Web technologies represents, therefore, a pre-condition for the development of services for legislators, legal information providers, legal experts as well as citizens. Recently EU Member States and Institutions have considered the use of Information and Communication Technologies in the legal domain of paramount importance to achieve better quality in legislation as well as to improve legal information management and accessibility across the EU. Institutional initiatives in legal document standards have been undertaken exactly to cope with these requirements, with the aim of providing high quality legal information integrated services.

This development is in line with the trends identified by E. Motta and M. Sabou [41]. In a comparison between the first and the next generation of Semantic Web applications, they identify several features of the new orientations: (i) reuse (vs. semantic data generation); (ii) multi-ontology systems (vs. single-ontology systems); (iii) openness with respect top semantic resources, (iv) scale as important as data quality, (v) openness with respect to Web (non-semantic resources), (vi) compliance with the Web 2.0 paradigm, (vi) openness to services.

3. ICT impact on the law field

Before we can see how the law domain can benefit from Semantic Web technology, we first discuss the impact of ICT in general on the field of law. Researchers in Artificial Intelligence and Law used to separate ICT (Information Communication Technologies) and Law into two big domains: (i) ICT law (data protection, copyright, security, domain names...), (ii) ICT for lawyers (e-government, e-court, Online Dispute Resolution, Multi-Agent Systems, etc.) [37, 38]. The first area would cover regulations and protocols. The second one refers to all the languages, tools, software, etc. that bring support to legal activities at the workplace. From a legal point of view this seems quite reasonable.

However, recent developments in semantic technologies, Natural Language Processing (NLP), legal ontologies, information retrieval technologies, and the Web 2.0 contribute to the convergence of the two approaches into a single techno-legal one. A lawyer seriously interested in meta tagging litigation cannot ignore OWL. A computer scientist developing legal ontologies for procedural legal knowledge must have a clear picture of court proceedings. This may also challenge the traditional “technological gap” or “computer divide” that researches in AI and Law have pointed out when describing the traditional diffidence of the legal field towards technology [34, 37, 38]. Barriers to ICT developments and AI applications certainly persist [33].

It is true that very little AI has been applied to the legal field so far. Detailed descriptions of working e-Court technologies in Europe lead to the same conclusion [24, 25, 42]. It may be true as well that, compared to other kind of company organizations, the legal industry “is a surprisingly fragmented, undercapitalized and
inefficient sector” [34]. Very likely, lawyers may do better. However, looking at the available data, there is no doubt that law firms have put effort in improving their skills and efficiency through ICT investments.

Annual technology surveys are part of the legal marketplace, and law firms spend a substantial part of their budget specifically for technology, according to the American Bar Association Tech reports. The average law firm spends 6% to 7% of gross revenue on technology-related expenses. This is correlated with firm size (from 2% to 7%). The 2007 and 2008 ILTA surveys [31, 32] show that small firms (under 200 attorneys) have higher implementation rates for case management, courtroom technology, docketing software, imaging/scanning/OCR, patch and records management software, while large law firms are more interested for remote access technology, voicemail upgrades, wireless connections and workflow automation.

We may distinguish the following ICT domains that have impact on the legal market [15, 46]: (i) Legal Information Research (LIR); (ii) Electronic Data Discovery (e-Discovery); (iii) Web-based communications; (iv) Collaborative tools; (v) Metadata; (vi) XML technologies; (vii) Technologies in Court-rooms and Judicial and Prosecutors’ offices (e-Court); (viii) Technologies in Administration offices (e-Government, e-Administration); (viii) Multimedia and law.

Some fields such LIR are well trodden-paths, with three big legal information providers (Thomson-Westlaw, Reed Elsevier-Lexis-Nexis, and Wolters Kluwer) controlling about 85% of the market. Others, such as e-Discovery —litigation support software that process, collect, preserve, review and produce electronically stored information, including e-mails— are still emerging. However, the market has gone up to 150 million US dollar. As recently reported in The Wall Street Journal, e-discovery conflicts have aroused in recent times between tech-law firms and software companies in the Common Law areas, because this is reducing the need for attorneys doing this kind of work [47].

Perhaps investments in technology could be considered one of the ways firms can reduce costs and improve services. This is related to outsourcing services as well. The December 2007 ILTA White Paper on Law Firm Staffing notes: “There are simply too many IT functions to be performed in today's law firm and too many different skill sets required to go it alone; outsourcing some IT functions, either in whole or in part, is the norm for firms of all sizes” [27].

However, there is more needed than outsourcing IT functions. We think that adopting new ICT techniques is the only way of coping with legal information overload and with the changes in customers’ expectations coming up with the Web 2.0 and 3.0. Knowledge Management in law firms may be defined as “the way in which lawyers optimize the relation between knowledge and knowledge processes with the help of Information Technology” [3]. We will show in the next section that this perspective is not only a professional issue, but an economic, political and cultural one. Focusing on the individual attorney perspective and combining different types of knowledge, as Apistola and Lodder [3] suggest, constitutes a good strategy to face the transformation which is taking place in the legal field.
4. The need for Semantic Web technology in the legal domain

The traditional fields and practices of law are changing fast. Legal drafting, private contracting, judicial sentencing and administrative management have been enlarged with online dispute resolution initiatives and new forms of self-regulation and access to justice. Citizens, customers and consumers require a greater participation and faster and more effective ways of facing their legal activities. The emergence of auto or self-made law within the Internet out of the law firms influence has been already noticed and checked by some market analysts [50].

Therefore, there is a clear need for less expensive lawyering, less adjudication procedures, more dialogue, more participation, and more flexibility and autonomy. These also seem to be the aims of new legal forms of relational administration and relational justice [14]. The urgency of this is clearly illustrated by a quote of a forthcoming paper by Colin Rule[45], affiliated with e-Bay:

“If you have any doubt that consumers are moving to online commerce, take a look at eBay, the online auction company. In the 13 years since it was founded, eBay has grown into the largest marketplace in the world. In the first half of 2008, there were more than one billion product listings added to eBay worldwide. At any given moment, there are more than 100 million listings around the world, and approximately 7.1 million listings are added each day. eBay users trade almost every kind of item imaginable, in more than 50,000 categories. On eBay, a pair of shoes sells every 7 seconds, a cell phone sells every 7 seconds, and a car sells every 56 seconds. The daily volume of trade on eBay is greater than the daily volume of the NASDAQ.

Unsurprisingly, all of these transactions generate a lot of consumer disputes. Even though less than 1 percent of purchases generate a problem, the incredible volume on the site means eBay handles more than 40 million disputes a year, in more than 16 different languages.”

Modern Web 2.0 developments [40] and Multi Agent Systems technology [12, 13] seems to be able to provide (partial) answers to these needs. It has been highlighted that the Web 2.0 implies a democratic model. People can cooperate and jointly build their ideas. Enriching this process in the direction of Web-mediated dispute resolutions seems a quite natural move from the Semantic Web perspective. However, it is not that easy. Semantic Web developers themselves have pointed out some obstacles or limitations of the original purposes of the Semantic Web [7]. For example, search in the World Wide Web is the great unfulfilled promise. Bridging Semantic Web and information retrieval technologies face scientific problems on knowledge representation and natural language understanding that remain still unsolved. In Baeza-Yates, Mika and Zaragoza’s words “IR research is strongly driven by a problem, whereas Semantic Web is driven by a solution” [4].

Although this may be true, hybrid approaches and perspectives seem to guide the ontological work, as long as folksonomies, wikis, data mining, NLP techniques, and upper, middle and domain ontologies develop [2, 11, 48, 55]. Moreover, to optimize

---

3 We thank the author for allowing us to quote his still unpublished paper.
the possibilities of success, SW strategies should be grounded on detailed economic, sociological and cognitive studies on the daily behavior and real needs of professional ICT users and consumers.

Another need for semantic technology follows from the usage of the web by legal professionals. Not only attorneys, but even judges and magistrates are currently browsing the web when they gather information to build their legal strategy to construct and solve the cases they have in front of them. According to the recent American Bar Association 2008 Legal Technology Survey Report legal research online has grown up from 79% in 2003 to 96 % in 2008 [1]. In 2008 the number of lawyers using free online research services has overtaken for the first time the number using for-fee services (89% vs. 83 %). Lawyers receive information mainly through news websites (79%) and e-mail newsletters (59%), and 72% of respondents report that they or their staff file court documents electronically, up from 55% in the 2007 survey. In addition, legal chats and blogs can be found along with public or private legal databases. Moreover, e-mail boxes are currently being used as personal databases.

We do not argue that common sense knowledge plays the same role as expert knowledge does when facing a legal case. However, lay people and experts tend to rely on the web for their information needs, and the web offers increasingly accessibility to documents containing legal rules and procedures, past cases and accumulated experiences.

This goes far beyond the practice to look and seek for information through the existing legal databases. We think that the reason for what lawyers do this is because they save time and effort. Therefore, expert knowledge, personal and professional experiences, and common knowledge have to be combined in a new way that reduces the differences among experts and lay people or, at least, approaches the legal perspective to the social perspective of non-legal users. Thus, lawyers (and magistrates) practice law through and within the web. This integration can clearly benefit from semantic technology.

In short, both the amount and character of legal activities, and the increased use of online information by lay and professional users require new ways of handling legal information. More and more people rely on web accumulated information first to find a solution for their administrative or legal problems. This situation makes the legal activity of citizens and the daily routine of experts suitable to be treated with the Semantic Web techniques.

5. Overview of the status on legal ontologies

Given the expected increasing impact of Semantic Web technology on the legal domain, it is worthwhile to have a look at the ontologies for the legal domain that are available. After all, ontologies are often the core of all applications that exploit

---

4 The Survey report is based on 789 completed Baseline and Budgets questionnaires: http://www.abanet.org/abastore/index.cfm?fm=Product.AddToCart&pid=2680079PDF
Semantic Web technology. In this section, we will discuss the different types of legal ontologies and give an overview of a large number of existing ontologies and their role.

The term ‘ontology’ may have different meanings: (i) philosophical discipline, (ii) informal conceptual system, (iii) a formal semantic account, (iv) a specification of a conceptualization, (v) a representation of a conceptual system via logical theory, (vi) the vocabulary used by a logical theory, (vii) a meta-level specification of a logical theory [29].

For Semantic Web technologies, an ontology may be defined as a 4-tuple \(<C,R,I,A>\), where C is a set of concepts, R a set of relations, I a set of instances and A a set of axioms [55]. Ontologies consist of concepts, relations, instances and axioms. They represent knowledge in such a way that can be understood and processed by a machine.

Legal ontologies represent legal knowledge. The problem, then, is to define in a more precise way what “legal knowledge” means.

In the mid-nineties, pioneering work by McCarty, Stamper, van Kralingen, Visser, Breuker, Winkels and Valente tried to bridge the gap between computational and legal concepts. Ontologies were faced as “the missing link between legal theory and AI & Law” [51]. Therefore, concepts traditionally viewed as belonging to the legal theory field such as rights, duties, norms and actions were modelized, in addition to more specific legal terms stemming from a particular domain (criminal, commercial or maritime Dutch law). These early efforts have recently lead to more developed legal-core ontologies, such LKIF-ontology, e.g.

There were four main directions in legal ontology building as identified by Visser and Bench-Capon [53]: (i) legal discourse (McCarty); (ii) legal norms (Stamper); (iii) frame-based ontology of law (Visser and van Kralingen); (iv) functional ontology of law (Breuker and Valente). More general upper and top ontologies lexically-based in Wordnet (Gangemi, Tiscornia, and Sagri) have to be added to these trends. And, to complete the whole picture, fundamental legal concepts and common-sense categories have been recently merged in a general ontology (LKIF) based on the LRI-Core Ontology developed in the nineties.

Ontologies keep growing in the legal field. Table 2 summarizes twenty-three already existing legal ontologies. It should be noticed that other types of legal knowledge are being added to the fundamental ones: legal professional knowledge, multimedia and global contexts (in which digital rights operate e.g), negotiation (ODR), and laymen legal conceptualization set up complementary scenarios.

As stated above, these types of legal knowledge fit into the new trends of the Semantic Web and the development of the Web 2.0. These ontologies are user-centered and even those built for information retrieval purposes are web service-oriented. They intend to operate through the Internet.
Table 2. Extension of André Valente’s table of existing legal ontologies [6, 52].

<table>
<thead>
<tr>
<th>Ontology or Project</th>
<th>Application</th>
<th>Type</th>
<th>Role</th>
<th>Character</th>
<th>Construction</th>
<th>Language</th>
</tr>
</thead>
<tbody>
<tr>
<td>McCarty’s Language of Legal Discourse</td>
<td>General language for expressing legal knowledge</td>
<td>Knowledge representation, highly structured</td>
<td>Understand a domain</td>
<td>General</td>
<td>Manual</td>
<td>English</td>
</tr>
<tr>
<td>Valente &amp; Breuker’s Functional Ontology of Law</td>
<td>General architecture for legal problem solving</td>
<td>Knowledge base in Ontolingua, highly structured</td>
<td>Understand a domain, reasoning and problem solving</td>
<td>General</td>
<td>Manual</td>
<td>English</td>
</tr>
<tr>
<td>Van Kralingen &amp; Visser’s Frame Ontology</td>
<td>General language for expressing legal knowledge, legal KBs</td>
<td>Knowledge representation, moderately structured (also as a knowledge base in Ontolingua)</td>
<td>Understand a domain</td>
<td>General</td>
<td>Manual</td>
<td>English</td>
</tr>
<tr>
<td>Mommer’s Knowledge-based Model of Law</td>
<td>General language for expressing legal knowledge</td>
<td>Knowledge base in English very highly structured</td>
<td>Understand a domain</td>
<td>General</td>
<td>Manual</td>
<td>English</td>
</tr>
<tr>
<td>Breuker &amp; Hoekstra’s LRI-Core Ontology</td>
<td>Support knowledge acquisition for legal domain ontologies</td>
<td>Knowledge base in DAML+OIL/RDF using Protege (converted into OWL)</td>
<td>Understand a domain</td>
<td>General</td>
<td>Manual</td>
<td>English</td>
</tr>
<tr>
<td>Hoekstra &amp; Breuker’s LKIF-Core Ontology</td>
<td>Support knowledge acquisition for legal domain ontologies</td>
<td>Knowledge base in OWL, highly structured</td>
<td>Understand a domain</td>
<td>General</td>
<td>Manual</td>
<td>English</td>
</tr>
<tr>
<td>Gangemi, Sagri &amp; Tiscornia’s JurWordNet</td>
<td>Extension to the legal domain of WordNet</td>
<td>Lexical Knowledge base in DOLCE (DAML), lightly structured</td>
<td>Organize and structure information</td>
<td>General</td>
<td>Manual</td>
<td>Italian</td>
</tr>
<tr>
<td>Benjamins, Casanovas et al. Ontology of Professional Legal Knowledge (OPLK)</td>
<td>Intelligent FAQ system (information retrieval) for judges (Iuriservice)</td>
<td>RDF. Knowledge base in Protégé, highly structured (converted into OWL)</td>
<td>Semantic indexing and search</td>
<td>Domain</td>
<td>Semi-automated</td>
<td>Spanish</td>
</tr>
<tr>
<td>Ontology or Project</td>
<td>Application</td>
<td>Type</td>
<td>Role</td>
<td>Character</td>
<td>Construction</td>
<td>Language</td>
</tr>
<tr>
<td>---------------------</td>
<td>-------------</td>
<td>------</td>
<td>------</td>
<td>-----------</td>
<td>--------------</td>
<td>----------</td>
</tr>
<tr>
<td>Casellas, N. et al.</td>
<td>i-FAQ for</td>
<td>Last version in OWL. Knowledge base in Protégé, highly structured</td>
<td>Semantic indexing and search</td>
<td>Domain</td>
<td>Manual</td>
<td>Spanish</td>
</tr>
<tr>
<td>Ontology of Professional Judicial Knowledge (OPJK)</td>
<td>judges (Juriservice, second version)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lame’s ontologies of French Codes</td>
<td>Legal information retrieval</td>
<td>NLP oriented (lexical), knowledge base, lexical, lightly structured</td>
<td>Semantic indexing and search</td>
<td>Domain</td>
<td>Automated</td>
<td>French</td>
</tr>
<tr>
<td>Legal information retrieval</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Leary, Vanderverghe &amp; Zeleznikow’s Financial Fraud Ontology</td>
<td>Ontology for representing financial fraud cases</td>
<td>Knowledge base (schema) in UML, lightly structured</td>
<td>Semantic indexing and search</td>
<td>Domain</td>
<td>Manual</td>
<td>English</td>
</tr>
<tr>
<td>Asaro et al.’s Italian Crime Ontology</td>
<td>Schema for representing crimes in Italian law</td>
<td>Knowledge base (schema) in UML, lightly structured</td>
<td>Organize and structure information</td>
<td>Domain</td>
<td>Manual</td>
<td>Italian</td>
</tr>
<tr>
<td>Boer, Hoekstra &amp; Winkel’s CLIME Ontology</td>
<td>Legal advice system for maritime law</td>
<td>Knowledge base in Protégé and RDF, moderately structured</td>
<td>Reasoning and problem solving</td>
<td>Domain</td>
<td>Manual</td>
<td>English</td>
</tr>
<tr>
<td>Lehman, Breuker &amp; Brouwer’s Legal Causation Ontology</td>
<td>Representation of causality in the legal domain</td>
<td>Knowledge base lightly structured</td>
<td>Understand a domain</td>
<td>Domain</td>
<td>Manual</td>
<td>English</td>
</tr>
<tr>
<td>Ontology or Project</td>
<td>Application Type</td>
<td>Role</td>
<td>Character</td>
<td>Construction</td>
<td>Language</td>
<td></td>
</tr>
<tr>
<td>---------------------</td>
<td>------------------</td>
<td>------</td>
<td>-----------</td>
<td>--------------</td>
<td>----------</td>
<td></td>
</tr>
<tr>
<td>J. Saias, P. Quaresma, Portuguese Attorney Office</td>
<td>Ontology to semantically enriching legal texts</td>
<td>OWL and logic programming (ISCO and EVOLP)</td>
<td>Organize and structure information</td>
<td>Domain</td>
<td>Automated</td>
<td>Portuguese</td>
</tr>
<tr>
<td>M. Klein, E. Uijttenbroek, A. Lodder, Laymen Ontology</td>
<td>Ontology to represent laymen knowledge on liability cases</td>
<td>OWL and NLP. Knowledge base in laymen natural language</td>
<td>Understand a domain (tort law) and interoperability between NL and legal concepts</td>
<td>Domain</td>
<td>Semi-automated</td>
<td>Dutch</td>
</tr>
<tr>
<td>J. Breuker, A. Elbag’s Crime.NL</td>
<td>Ontology of Dutch Criminal Law</td>
<td>OKBC</td>
<td>Main structure of (Dutch) criminal law; for comparing European CL</td>
<td>Domain/General</td>
<td>Manual</td>
<td>Dutch/English</td>
</tr>
<tr>
<td>Tiscornia, Francesconi, Spinosa et al. - DALOS</td>
<td>Legal drafting</td>
<td>OWL</td>
<td>Support for multilingual legal drafting and document indexing</td>
<td>Domain</td>
<td>Manual/automatic</td>
<td>English, Spanish, Italian, Dutch</td>
</tr>
<tr>
<td>S Despres, S. Szulzman</td>
<td>Micro-ontology</td>
<td>Ontology to represent concepts in European Directives</td>
<td>OWL and NLP (TERMINAE method)</td>
<td>Understand a domain</td>
<td>Domain</td>
<td>Semi-automated</td>
</tr>
<tr>
<td>UCC Ontology</td>
<td>Ontology to represent top-level concepts (e.g. ownership)</td>
<td>NML. Top-level ontology based on NM</td>
<td>Organize and structure information</td>
<td>Domain (top-level)</td>
<td>Manual</td>
<td>English</td>
</tr>
<tr>
<td>E. Schweighofer, D. Liebwald’s CLO (Comprehensive Legal Ontology)</td>
<td>Ontology for information management</td>
<td>Some frame representation</td>
<td>General</td>
<td>Manual with support of legal core ontologies</td>
<td>English?</td>
<td></td>
</tr>
<tr>
<td>E. Melz &amp; A. Valente’s IRC ontology</td>
<td>Ontology of Internal Revenue Code (USA)</td>
<td>OWL</td>
<td>Reasoning about tax cases</td>
<td>Domain</td>
<td>Manual</td>
<td>English</td>
</tr>
</tbody>
</table>
6. About the content of this book

The articles in this book are sorted in two main categories: 1) those which focus on the modeling of legal ontologies and the resulting ontology itself, and 2) those presenting applications, which use ontologies and Semantic Web technology. Within each category one will find a high diversity in kinds, roles and applications of legal ontologies. The message should be clear: in the legal domain – and probably in other domains as well – the nature and use of ontologies is highly diversified. This is no so much due to the fact that there are still differences in believing what an ontology should be about, but rather to the discovery that ontologies play a pivotal role in applications that require some semantic understanding of terms used by these applications. This understanding may range from automated reasoning to semantic support in legal information and knowledge management.

The first article by Hoekstra, Breuker, Di Bello and Boer presents LKIF-Core, a core ontology that contains definitions of concepts that are general and typical for legal domains. The ontology is cast in OWL-DL and its main distinctions are between physical and mental concepts. LKIF-Core is particularly worked out on notions about documents and (legal) roles: roles being the basic terms for social structures. The ontology serves a number of purposes: the usual and main one being a template for modeling legal domain ontologies: an example modeling a EU Directive on driver’s licenses.

The article by Gangemi discusses another way in which ontological definitions can act as templates in knowledge acquisition for legal domains: by using ‘design patterns’: recurring, abstract structures of defined terms. They can be reused as “core components” in ontology construction in (legal) domain. An overview of the typical tasks and services for legal knowledge is presented, the notion of ontology design pattern is introduced, and some excerpts of a reference ontology (CLO) and its related patterns are included, showing their utility in a simple legal modeling case.

If these two articles provide top-down support for developing legal domain ontologies, the article by Lenci, Montemagni, Pirrelli and Venturi explains how legal ontologies can be modelled in a bottom-up way by using a tool that enables learning from text (T2K) that combines a full range of technologies from Artificial Intelligence (AI): natural language processing, statistical text analysis and machine learning. It is applied to two legal text corpora and it is shown that the interleaving of these technologies provide a valid harvest of terms, which form semantic clusters which can be further processed manually as to construct an ontology.

The next article is complementary to the previous one, as it also combines linguistic extraction and statistical techniques to arrive at ontology population. Walter and Pinkal describe how they trace (legal) definitions from 6000 German court verdicts: they observe that definitions are an important element of legal text, and apply linguistic tools for their extraction. These definitions lend themselves easily to ontological modeling. The tools used enable the extraction of these definitions with high precision, although recall is to be improved. Also, here, the actual ontology is a next step, not reported.
The article by Mochales and Moens discusses also natural language processing technologies. However, in this case the aim is not to extract terms for an ontology, but to detect (legal) argument structures in text. It differs from the previous two articles that it is not so much concerned with definitions of terms – the basic stuff of ontologies – but of reasons. Legal reasoning uses (many) definitions of terms, but they are only components in the reasons for justifying decisions. The article provides an overview of technologies for automated argument detection, and shows that argumentative text parts can be separated from non-argumentative ones. They demonstrate that many applications are available for use on the (Semantic) Web, in particular for legal domains.

Argumentation is also a central issue in the article by Trojahn, Quaresma and Vieira. As ontologies become more and more available on the web with similar or overlapping terms, it becomes essential to assess how these overlaps match. In this article a method is presented based upon Value-based Argumentation Framework (VAF) in a multi-agent paradigm. An example is presented, where LKIF-Core (cf. Hoekstra et al., this volume) and LCO (cf. Gangemi, this volume) are compared.

The sharing and reuse of content from the web poses many (new) issues on copyrights. They are discussed in the context of digital right management systems, but it is pointed out by Garcia and Gil that these approaches only make sense if supported also by digital tools. To enable (semantic) interoperability the authors have developed an ontology cast in OWL-DL, which takes as main concepts the notions of works, rights, actions, and copyright licenses, which require the modeling of roles, events and deontic modalities.

Casanovas, Casellas and Vallbé present an ontology based upon the professional experiences of judges in providing legal support in day-to-day police activities. This ontology (Ontology of Professional Judicial Knowledge (OPJK)) drives Iuriservice, a FAQ advisory system, intended for novice judges. OPJK, expressed in OWL, provides the semantics for accessing Iuriservice. The development of OPJK which covers practical rather than formal legal knowledge is described in detail.

The contribution of Agnoloni, Bacci, Francesconi, Peters, Montamegni and Venturi to this book reports the use of ontologies in legal drafting, and in particular the multi-language aspects of the European context (Directives) which require more than simple dictionary mappings, but also a representation of the underlying semantics. The lexical and semantic knowledge form two connected layers.

The last article by Casanovas, Binefa, Gracia et al. describes how a legal ontology (e-Sentencias) enables the management of multi-medial information, obtained in recording hearings of Spanish civil courts. The ultimate goal is to obtain an automatic classification of the episodes of the audio-visual records. The focus in the article is on the knowledge acquisition process.

All together, this book can be read as an overview of ongoing attempts to manage the legal information flood. Ontologies provide the channels that separate the streams on semantic grounds, while Semantic Web technologies take the high ground of dykes to control and prevent overflow.
References


Valente. Types and Roles of Legal Ontologies, in R.V.Benjamins et al. (2005), 65-76.


Legal Knowledge Modelling and Acquisition

Legal Ontology Construction
This page intentionally left blank
LKIF Core: Principled Ontology Development for the Legal Domain

Rinke HOEKSTRA a, Joost BREUKER a, Marcello DI BELLO b, and Alexander BOER a

a Leibniz Center for Law, University of Amsterdam
breuker@uva.nl, hoekstra@uva.nl, aboer@uva.nl
b Department of Philosophy, Stanford University
mdibello@stanford.edu

Abstract. In this paper we describe a legal core ontology that is part of the Legal Knowledge Interchange Format: a knowledge representation formalism that enables the translation of legal knowledge bases written in different representation formats and formalisms. A legal (core) ontology can play an important role in the translation of existing legal knowledge bases to other representation formats, in particular as the basis for articulate knowledge serving. This requires that the ontology has a firm grounding in commonsense and is developed in a principled manner. We describe the theory and methodology underlying the LKIF core ontology, compare it with other ontologies, introduce the concepts it defines, and discuss its use in the formalisation of an EU directive.

Keywords. ontology, legal ontology, ontological engineering, legal concept, LKIF, knowledge representation, framework, methodology, common sense

Introduction

This article describes a legal core ontology that is part of the Legal Knowledge Interchange Format, developed in the Estrella project.1 LKIF specifies a knowledge interchange architecture that serves two main purposes. Firstly, it enables the translation between legal knowledge bases written in different representation formats and formalisms. The second purpose of LKIF is to be a knowledge representation formalism in its own right, i.e. to form the basis for standard reasoning services provided by legal knowledge systems.

Languages used by commercial knowledge-based systems vendors are typically rather idiosyncratic rule formalisms with varying expressiveness [34]. The LKIF language is therefore expressive enough to cover the full range of formalisms used in practice, and has proven to be able to translate back and forth between a number of vendor specific languages [13]. This approach is similar to that of the Knowledge Interchange Format (KIF) of [31] and its successor CommonLogic [46] which have the expressiveness of full first order predicate logic.

1Estrella is a 6th European Framework project (IST-2004-027665). See http://www.estrellaproject.org
However, LKIF differs from KIF in that it is not a uniform formalism, but rather combines two different knowledge representation formalisms: OWL 2 DL, and LKIF Rules [60, 3, 6]. Moreover, the language offers the ability to represent legal argumentation schemes [33, 32]. This versatile, hybrid character of LKIF enables the mapping of many formalisms used in the practice of building legal knowledge systems.

Although indeed many such systems are specified using rule formalisms, it becomes increasingly popular to capture legal knowledge in ontologies, as this book shows. Over the years, ontologies have been cast in a large variety of formalisms and languages ranging from industry standards such as UML and TopicMaps to OWL DL, the W3C’s standard for representing ontologies on the Semantic Web. OWL differs significantly from UML and TopicMaps because it is an explicit knowledge representation formalism that has the attractive properties of being sound, complete and decidable. Moreover, efficient inference engines – tableaux based theorem provers – have been specifically developed and optimised for this language. As we will show below, this is exactly what is needed to use ontologies as knowledge bases. However, these attractive properties come with a price: the language has limited expressiveness. The Estrella project has developed translators that allow translating back and forth between UML and OWL [13].

Ontologies capture that knowledge which is considered to be the undisputed backbone of what is known about a particular domain. It is an axiomatic starting point and necessary foundation for describing a domain’s wide range of less indisputable or more specific knowledge. For instance, although most if not all of our knowledge of the world around us is based on assumptions and beliefs, an ontology captures that (terminological, conceptual) knowledge on which such strong agreement is reached that it is undisputed and universally held to be true (for some purpose). This agreement is often described as an ontological commitment [22]; it defines that which we hold to exist. In fact, the ontology must capture that knowledge shared by humans knowledgeable in the domain: it is a shared commitment. Furthermore, it is shared by the fact that the definitions of terms in an ontology form the common ground for the knowledge captured by more specific domain knowledge bases.

A Core Ontology

Given the central role of ontologies, one would expect currently existing knowledge systems to contain ontologies as a basic resource for reasoning. Unfortunately, this is rarely the case – although a notable exception can be found in the field of qualitative reasoning [26]. Often, ontologies are rather used as domain vocabularies for the management of large collections of documents. Arguably, the legal domain is a typical example where such shared vocabularies can significantly improve accessibility of legal information. Information management problems dealing with law and legal documents are not only of importance to legal professionals, but to citizens as well.

This article describes the LKIF Core ontology, an ontology that can be used as central knowledge component for legal knowledge systems. For instance, the HARNESS ar-

---

2 The vocabulary used in the specification of LKIF Rules and argument schemes is also specified in OWL 2 DL as part of the ontology presented here (see e.g. [47]). A reasoner for unfolding argumentation schemes is implemented as part of the Carneades argument mapping system, http://carneades.berlios.de/.

3 This poses some problems in translating UML or TopicMaps into OWL DL. The main bottleneck is in the fact that in UML and TopicMaps (and for that matter also RDF) allow meta classes and reification (e.g. association classes).
chitecture allows legal reasoning based on this ontology, e.g. the application of norms to cases represented in OWL DL for the purposes of legal assessment [16, 76]. In this respect, the LKIF ontology is similar to the domain ontology developed in the CLIME project [78, 80]. However, it differs in that it is a core ontology, i.e. it contains the terms common to some field of activity or discipline, but which not specific to some domain.

In [12, 44] we identified four main ways for a legal core ontology to support information exchange in the context of LKIF. First of all, the ontology can serve as a resource for special, legal inference. Secondly, the definitions of terms in the ontology can facilitate knowledge acquisition, a terminological framework can facilitate the exchange of knowledge across multiple knowledge bases, and lastly it can be a basis for semantic annotation of legal information sources.

**Resource for Special Legal Inference** Typical and abstract legal concepts are often strongly interrelated and thereby provide the basis for computing equivalencies, or paraphrases, and implications. For instance, by representing an obligation as the opposite of a prohibition, a (legal) knowledge system can make inferences that capture the notion that they are each others’ inverse. A prohibition leaves all options open – except the one that is forbidden – while an obligation is unavoidable when all its requirements, or conditions, are satisfied. Although this implicit knowledge is relevant when reasoning with norms and cases, it does not express the control knowledge of reasoning (as in problem solving methods), but merely elicits the implications of declarative definitions. Specialised legal inference can be based on definitions of concepts in an ontology: an inference engine can generate the implied consequences of explicit concept definitions. LKIF Core defines deontic qualifiers (prohibition, obligation and permission) in such a way that they can be used in normative reasoning (see Figure 7). HARNESS uses these definitions to assess legal cases, i.e. to infer violation of, or compliance with norms. A classical example of specialised inference using the definitions in an ontology and a (general) inference engine is temporal reasoning based on [1]’s ontology of time (Section 4.1). To enable special inference, terms should be highly interrelated and form a coherent cluster with little or no external dependencies ([41], and Section 2). An example of such a cluster in the legal domain is that of the terms that denote deontic qualifications. Clusters of this type are usually found at very high levels of abstraction.

**Knowledge Acquisition Support** The classical use of both top and core ontologies in knowledge representation is as a means to support knowledge acquisition. If well designed and explained, they provide an initial structure to which domain terms can be attached as subclasses. Inheritance of properties and other implicit knowledge can then be used to check not only consistency, but also the extra-logical quality of the ontology: whether what is derived (classes, properties) makes sense. The use of a core or top structure that has well tested and evaluated implications, makes it easier to check whether domain refinements are not only consistent, but also arrive at inferences that correspond to what the knowledge engineer or user holds to be valid. The knowledge acquisition support of ontologies is not restricted to just

---

4HARNESS: Hybrid Architecture for Reasoning with Norms Exploiting Semantic web Standards.  
5CLIME, Cooperative Legal Information Management and Explanation, Esprit Project EP25414  
6These inferences are produced via the use of the Pellet DL reasoner (http://pellet.owldl.com/).
ontological or even terminological knowledge. For instance, the CLIME ontology
was successfully used for the incremental specification of normative knowledge
[81].

Preventing Loss in Translation A legal ontology can play an important role in the
translation of existing legal knowledge bases to other representation formats. In
particular when these knowledge bases are converted into LKIF as the basis for
articulate knowledge serving. Similar to a translation between different natural
languages, a formal, ‘syntactic’ translation may clash with the semantics implied
by the original knowledge representation. An ontology, as representation of the
semantics of terms, allows us to keep track of the use of terms across multiple
knowledge bases.

Resource for Semantic Annotation The semantic annotation of legal sources, such as
regulations and jurisprudence is an important contribution to the accessibility and
maintainability of these sources. First of all, an ontology can be a source for infor-
mation retrieval. Secondly, the status of an officially sanctioned legal text is pri-
marily determined by its relation to other legal texts [8]. This status can be made
explicit by expressing it using ontologically defined relations. In fact, these rela-
tions do not just hold between the texts themselves, but between the formal repre-
sentations of their content as well [48].

The following sections describe the theoretical and methodological framework
against which the LKIF core ontology has been developed. Section 1 discusses the per-
spective used in its construction in relation to five other ontologies. The methodology
used to construct the LKIF ontology is discussed in Section 2. Section 3 extends this
methodology by introducing a distinction between ontologies and frameworks, followed
by a discussion of the modules and most important concepts of LKIF Core in Section 4.
Section 5 gives an example of how the ontology can be used in the formalisation of a
regulation.

1. Points of View: Other Ontologies

As we discussed, one of the main roles of ontologies is to ease knowledge acquisition
by providing readily usable concept definitions. The same bootstrap can be applied to
ontology development itself; existing ontologies can be a viable resource for ontology
construction in the form of predefined concepts and relations. These definitions convey
ontological commitments that capture important ontological choices [54]. When these
choices are in concert with our requirements, adopting and extending an existing ontol-
ogy relieves us from the burden of micromanaging these choices ourselves.

The topmost layers of a top or core ontology characteristically distinguish funda-
mental, but highly abstract categories of knowledge, that for centuries have been subject
to philosophical study and debate [71]. Many distinctions prominent in this debate have
found their way to contemporary ontologies. It therefore seems advisable to assess the
currently available top-ontologies as source for re-usable definitions (see also Section 2).
In a similar vein, an ontology may provide abundant inspiration for dealing with recur-
rent problems in knowledge representation in the form of design patterns (viz. [30, 43]).
In this section, we outline two main requirements for the LKIF Core ontology, and show
that even at this highly abstract level, the point of view taken in the development of an ontology has a large impact on how, and which categories are defined.

Besides the general requirements for knowledge representation ontologies, the ontology is to contain a core set of definitions for describing specific legal terms. Valente [75] conceived of law as an instrument used by the legal and political system to identify and control situations and events in social interaction. By far the bulk of social situations, be it in our family life, at work, related to transport, property, crime, etc. is not described in specialised technical terms: their meaning is part of common sense. For instance, the conflicts and problems brought to court – legal cases – are initially described using common sense terms, and are gradually translated into legal technical terminology in the process of coming to a decision. For this legal qualification to be possible, the gap between the vocabulary of a case and legal terminology needs to be bridged [79]. The possibility of legal qualification in general – by legal professionals – is a strong indication that the vocabularies of legal knowledge and common sense are not disjoint. Legal terminology can be reduced to the actual societal events and states governed by law. In other words, the basic categories of the LKIF ontology should reflect Valente’s view that legal world knowledge is an abstraction of common sense [75].

A third requirement for LKIF is given by the ideal of the Semantic Web to achieve understanding both between web services and between human users. In fact, a common sense perspective is also applicable to any serious endeavour towards a Semantic Web. As the the web is about the most diverse information source we know today, a common sense oriented ontology would certainly be an important first step to more uniform web-based knowledge representation. Conversely, the distribution of legal information across various strongly interconnected sources is a demanding use case for semantic web technology [42]. An important requirement is therefore that the LKIF core ontology should be represented using the DL profile of OWL.

Given a commonsense perspective, we expected that (at least parts of) existing ontologies would be reusable, ranging from a source of inspiration to straightforward import of definitions. This would hold, in particular, for top ontologies that include legal terms, as for instance listed in [17]. Unfortunately, it turned out that the amount of reuse and inspiration was rather limited. Not only do existing ontologies diverge on approach, coverage and knowledge representation language used; those ontologies that do claim a common sense or similar perspective differ in their conception as to what such a perspective means.

We consulted several ontologies and evaluated them for their potential contribution to LKIF Core. The main requirement is their suitability to enable the primary roles of the LKIF ontology outlined in the previous section. We pay specific attention to their definition of commonsense and legal terms, and possibilities for safe reuse: reuse should not alter the semantics of the reused ontology [18].

1.1. Suggested Upper Merged Ontology

The SUMO ontology of [63] brings together insights from engineering, philosophy, and information science. It provides definitions for general purpose terms, is intended as a unifying framework for more specific domain level ontologies. As a starting point for a legal core ontology SUMO has several drawbacks. First of all, it does not readily provide definitions of terms relevant to the legal field – e.g. its coverage of mental and social
entities is limited. Because of the way in which SUMO is constructed, it has a bias towards more abstract and theoretical insights coming from engineering and philosophy. Although it is non-revisionist, as in e.g. the distinction between objects and processes, it does not have a real commonsense basis. Furthermore, SUMO is a foundational ontology and uses meta modelling, such as in the definitions of classes, binary relations and sets. As SUMO is represented in the expressive language KIF, and more recently Common-Logic, this practice is not fundamentally problematic. However, it means that safe reuse of SUMO definitions in the construction of an OWL DL ontology is not possible.

1.2. Descriptive Ontology for Linguistic and Cognitive Engineering

DOLCE is part of the WonderWeb library of foundational ontologies, cf. [54, 27]. It was meant as a reference point for the other ontologies in the library, to make explicit the differing rationale and alternatives underlying the ontological choices of each ontology. Rather than a coherent set of ontological commitments, it captures a range of alternatives. This way, the library would form a network of different but systematically related ontology modules. The relation between an ontology, available in the library, and the DOLCE ontology expresses its ontological commitment to particular ontological options. DOLCE was therefore never presented as the foundational ontology it is currently regarded as, but it has been successfully used as such in a large number of projects.

DOLCE is very much an ontology in the philosophical tradition, and differs from the knowledge representation perspective in two significant ways. Firstly, its perspective is philosophical with respect to its content, i.e. it is aimed to directly represent reality. And secondly, it is subject to the epistemological promiscuity of philosophical ontology because it is rather an extension of the knowledge representation formalism at the ontological level [39], than a model expressed using that formalism. The meta-level character of DOLCE means that the ontology is not a representation of knowledge, but of the terms used to describe knowledge – in the same way that the constructs of OWL are. Like SUMO, DOLCE was originally specified in first order logic and the highly expressive KIF language. Its OWL DL representation (DOLCE-Lite) is more restrictive, e.g. it does not consider temporal indexing and relation composition.

The DOLCE ontology is descriptive, and is based on the stance that “the surface structure of natural language and human cognition” is ontologically relevant. It is argued that this perspective results in an ontology that captures cognitive artefacts more or less depending on human perception, cultural imprints and social conventions, and not deep philosophical insights of Ontology. DOLCE thus claims an approach that fits more with a commonsense perspective than the science perspective of SUMO. However, the suggestion that this surface structure has any bearing on common sense is not based on evidence. Rather, the methodological commitment to the surface structure of language and cognition almost inevitably resulted in an intricate framework of theoretical notions needed to encompass the idiosyncrasies of human language use. Alternatively, rather than constructing an ontology by studying reality through the kinds of categories implicit in natural language, a pragmatic approach based more directly on the conceptualisation of reality we use and live by in our daily routine is more appropriate [10, 45, and Section 1.5].

---

1Emphasis by the authors. [54]
1.3. Core Legal Ontology

Over the years, DOLCE has been extended in several ways. DOLCE+ is the extension of DOLCE with a theory on descriptions and situations (also called D&S, [28]). CLO, the Core Legal Ontology [29] extends DOLCE+ even further and defines legal concepts and relations based on its formal properties. CLO was designed to support both the definition of domain ontologies, a juridical Wordnet, and the design of legal decision support systems. To a large extent these goals correspond with the requirements of the LKIF ontology.

CLO conceives the legal world as a description of social reality, an ideal view of the behaviour of a social group. It builds on the D&S distinction between descriptions, and situations. Examples of legal descriptions, or conceptualisations, are the contents of laws, norms, and crime types. These descriptions constrain legal situations, i.e. legal facts of cases, such as legal, relevant non-legal and juridical states of affairs. Every legal description classifies a state of affairs. More precisely, a legal description is the reification of a theory that formalises the content of a norm, or a bundle of norms. A legal case is the reification of a state of affairs that is a logical model of that theory. A description is satisfied by a situation when at least some entity in the situation is classified by at least some concept in the description. Classification in CLO is thus not DL classification, and it is unclear as to what extent the two interpretations are compatible.

Viewing the legal system as description, or rather prescription, of reality is not new, cf. [75, others]. However, the CLO distinction between descriptions and situations is rather one dimensional in that it does not commit to an ontological view of the kinds of descriptions involved. In line with the DOLCE ontology, it confounds the distinction between representation and the represented with representation and reality. Although it introduces new levels of abstraction by reification, it does not provide ontological categories that can be used to describe the knowledge at these levels. In a language that itself can be conceived as providing the means to construct descriptions of reality (situations), such as OWL DL, it is unclear what the epistemological status of the classes ‘description’ and ‘situation’ themselves is. For instance, what is the difference between an individual description being classified-in-the-OWL-sense as some description class, and a situation class being classified-in-the-CLO-sense by that same description?

As CLO relies on a subset of DOLCE for the definition of elements of situations, it is subject to the same criticism with respect to its commonsense perspective. Moreover, the lack of ontological commitment at the level of descriptions undermines its suitability for knowledge acquisition support in a legal setting as well. Although for sure a norm can be described as some description of a situation, it is not the norm-as-description that uniquely characterises what a norm is. This holds especially for less obvious ‘descriptions’ (in CLO terms), as e.g. damage or right of way.

1.4. CYC

CYC is a huge unified ontology of commonsense concepts [53, 52]. Although the project has started as early as 1984, its general set-up corresponds to that of later large scale ontology efforts. The main motivation for the Cyc Project was that all software programs would benefit from the availability of commonsense knowledge. This idea is not new, and was acknowledged in the early years of AI: ‘A program has common sense if it
automatically deduces for itself a sufficiently wide class of immediate consequences of anything it is told and what it already knows” [56, p.2].

The idea is that, when enough commonsense knowledge is represented, and a certain threshold is reached, a quantum-leap (“The Singularity”)8 would enable CYC to expand its knowledge through guided learning (as a human child would). This theory is in line with Minsky’s ideas about how computers can become intelligent beings: add enormous amounts of commonsense knowledge [58, 59]. This basic knowledge about the workings of the world would finally allow you to send the kid to school. With currently over 300K concepts, the knowledge base seems well under way in reaching this threshold, however we still have to see the first results.9

The upper part of the CYC ontology is claimed to express a commonsense view and indeed it is more concrete than either SUMO or DOLCE. On the other hand, from a methodological point of view, the CYC approach is not very satisfactory either. Technically, CYC qualifies rather as a terminological knowledge base than as an ontology proper.10 Instead of a meticulous study of the actual workings of the world, as in SUMO, or the surface structure of language and cognition, as in DOLCE, it seems the approach followed is to have a bunch of knowledge engineers simply put everything they know into the CYCL formalism.11 This procedure results in a large portion of the knowledge base being decidedly non-ontological, but rather context dependent, frameworks (see Section 3).

Furthermore, CYC suffers from two more technical impediments for reuse. Firstly, like SUMO and DOLCE, it is specified in the very expressive CYCL representation language, which is based on first-order predicate calculus. Recently a port of the publicly available OpenCYC effort has been made available in OWL Full, but again, this does not cover the full semantics of the ontology.12 CycL admits meta modelling, which indeed is used liberally throughout the ontology – even more so than in DOLCE and SUMO. Secondly, the sheer size of the knowledge base – as with most unified ontologies – introduces significant reasoning overhead for even the simplest tasks. As such, CYC seems more suitable for direct inclusion in a knowledge based system than as a conceptual coat rack for ontology development on the Semantic Web.

1.5. Ontology Reuse

Thus far, the ontologies we reviewed do not appear to meet the requirements for the top structure of a legal core ontology. Although in the past few years the ontologies underwent changes and extensions, these results are in line with the outcome of an earlier review [10]. Firstly, the ontologies are specified at multiple (meta) levels of abstraction, using very expressive languages, which limits possibilities for safe reuse.

Furthermore, in all three ontologies those concepts that are of relevance to law are either scarce and under specified, or overly theoretical. In particular, our requirement

---

8Indeed, CYC is a much-hyped project, and has received a lot of criticism because of it.
9The online game FACTory for teaching CYC is online at http://game.cyc.com.
that a legal core ontology should be built on a commonsense conception of reality is not met. Where a commonsense perspective is claimed, it is not motivated, explained or substantiated. The common sense of CYC is in fact *common knowledge*, or rather “human consensus reality knowledge” [53, p. 33], i.e. that knowledge of things most humans will concede to exist in reality. In contrast, the DOLCE and SUMO ontologies do not commit to the ontological relevance of human consensus, but rather aim to provide an ontological grounding for *all* knowledge. They do this in two distinct ways. SUMO is based on scientific knowledge of objective reality, and for DOLCE the way in which we apply and consciously report our knowledge is ontologically relevant.

Both approaches are generic enough to be the basis for an ontology that describes our common sense, in the sense of common knowledge. However, in our view, common sense refers not to some common body of knowledge of reality, but rather to the commonality of the scope and detail of that knowledge as it manifests itself in individual persons: it is a *level of description*, much akin to the *basic level* of [49] (see Section 2). Characteristic of this level is not just the *kind* of things we commonly know, but more importantly, the way in which this knowledge is *structured*. But for the last requirement, a philosophical approach would be perfectly adequate. However, a commonsense ontology should not be specified in highly specific, theoretical jargon, but should rather have a commonsense structure of its own.

1.5.1. LRI Core

The review in [10] motivated a decision to develop a legal core ontology to support the development of ontologies for criminal law in various European countries, as part of the e-Court project. This ontology, *LRI Core*, was developed with a set of design goals similar to that of LKIF Core, cf. [11, 10]. What sets LRI Core apart from other ontology efforts is that its definitions were aimed to be verifiable through empirical research on how humans relate concepts in actual understanding of the world; and not about revisionist views of how we should view the world as it actually is (as e.g. in correct theories of the physical world) or as it makes up a parsimonious view on reality (as e.g. in philosophical views on the main categories of description). This kind of empirical evidence can range from cluster analysis of semantic distance between terms, to neuro-psychological evidence.

Central to this effort is the view that common sense is rooted in a conceptualisation that is at its heart the result of our evolution. This conceptualisation is developed in order to deal with a dynamic and potentially dangerous environment. Our capacity to move, perceive and interact with reality has led to increasingly complex cognitive representations. These representations range from hard-wired abstraction in our perception system, such as the ability to perceive straight lines and angles at a mere 2 neurones away from the retina, via the inborn syntheses of perceptual input into basic properties, to – eventually – the representations accessible to conscious thought.

This range of increasingly abstract and complex representations of reality defines an axis that indexes organisms in successive stages of evolutionary development, e.g.

---

13Electronic Court: Judicial IT-based Management. e-Court was a European 5th Framework project, IST 2000-28199.

14For instance the classical distinction between abstract and concrete concepts in philosophy does not fit well with human understanding of the dynamics of the world in terms of physical causation, intention and life. The commonsense explanation of an event may involve all three categories.
from viruses and bacteria to multi-cell organisms, insects, mammals, primates and finally homo sapiens. In other words, the competencies of our genetic ancestors give insight in what the roots of our common sense are. These roots may well be hidden too deep to be accessible to conscious thought and introspection. Nonetheless, on the basis of insights in cognitive science we can make several basic distinctions that go well beyond a mere hunch. As most are of relevance to the LKIF Core ontology as well, they are briefly outlined below.

**A Cognitive Science Perspective**

Given that the physical environment is relatively stable – the notable exceptions being day-night cycles, changing weather conditions and the occasional geological perturbation – the perceptual apparatus of most organisms is tuned to register the slightest change occurring against this stable canvas. Of particular relevance is the ability to be aware of changes induced by other organisms. Firstly, the presence of another organism may present an opportunity for reproduction and sustaining metabolism (i.e. by eating). And secondly, their presence may signify a direct threat to an organism’s existence. The result is a prominent distinction in cognition between ‘background’ and ‘foreground’. In general our awareness is directed to the discontinuity of change rather than spatial arrangements of objects or historical continuity. The ability to perceive stability is enabled by episodic memory, though it requires some conception of the physical constraints underlying this stability. An example is the general rule that physical objects keep their position unless subjected to a change in the exertion of force. Changes occur against the canvas of temporal and spatial positions, and the speed of a change is indicative of whether it becomes foreground or remains in the background.

In LRI Core, the view that knowledge serves to interpret occurrences in the world was reflected by a contrast between concepts on the one hand, and individuals and their occurrence (instance) on the other. Furthermore, the cognitively primary distinction between static and dynamic elements in the (physical) world is reflected by differentiating objects and processes. Objects have extensions in space where processes have extensions in time, but are contained by objects. Processes reflect a causal explanation of change. The notions of space and time do not just play a role as the extension of objects and processes but can indicate positions as spatio-temporal referents. The spatio-temporal position of an object is not inherent: a change in position does not constitute a change in an object.

Very recent – at least in evolutionary terms – mammals developed the ability to attribute intentions to other animals. Because of the intentional interpretation of behaviour, change is no longer private to physical causation but to the actions of other agents as well. Actions are always intentional “under some description” [20]: performing an action comes down to the initiation of processes that bring about some intended change or state. Paradoxically enough, taking into account the mental state of other animals precedes the ability to consciously reflect on our own mind. And although this was long thought to be a skill exclusive to humans, it has been shown that our next of kin – chimpanzees, bonobo’s – have self awareness as well.

Because mental representation of other mental representations is a fairly recent accomplishment, the models we construct reuse many parts of the conceptualisation originally developed to deal with physical reality. We think and speak of mental processes and mental objects in the same terms we use to talk about the physical world. In other

\[\text{\textsuperscript{15}}\text{Also: mental actions, e.g. in (trying) to control one’s thoughts.}\]
words, these terms are metaphors of similar physical notions. Some, i.e. [50], even argue that the same mechanism is used to construct the highly abstract notions of mathematics.

Social awareness and self awareness are the most important prerequisites for complex social behaviour. First of all, they enhance the predictability of our environment by allowing us to take the possible intentions of other agents into consideration for planning and control. Secondly, it allows us to share plans with other agents. Repeated execution of such co-operative plans can render them institutionalised by a community. Plans make extensive use of roles to specify expected or default behaviour. The ability to play a certain role expresses a (recognised) competence, sometimes acknowledged as a social ‘position’. In LRI Core, roles played a central part in the construction of social structures.

The LRI Core ontology distinguished four ‘worlds’:

1. A physical world, divided by processes and objects, each containing matter and energy.
2. A mental world, containing mental objects and mental processes. The mental world is connected to the physical world through actions, which translate an intention into some physical change.
3. A social world, built from mental objects such as roles
4. An abstract world, which contains only formal, mathematical objects.

Because of its grounding in cognitive science and its explicit common sense perspective, the characteristics of the LRI Core ontology are relatively similar to those intended for the LKIF ontology – especially in comparison to the other ontologies we discussed so far. And it is indeed true that it can in many ways be seen as the direct precursor of the LKIF ontology.16

Nonetheless, the LRI Core ontology falls short in several important respects. Though it is specified in OWL DL, most concepts in the ontology are under specified. They are defined by subsumption relations, and are only sparsely characterised using more complex class restriction. Furthermore, the distinction between the different worlds in LRI Core is fairly absolute, and no theory is provided as to how they are connected and interact. Thirdly, apart from a relatively well-developed characterisation of roles, LRI Core is relatively underdeveloped with respect to the mental world. For instance, it emphasises physical objects, processes, energy and substance while remaining rather sketchy with respect to their mental counterparts. In part the limitations of LRI Core can be ascribed to an unprincipled top down methodology.

Concluding, although the perspective and main distinctions of LRI Core were used as inspiration in the construction of the LKIF ontology, it is not simply a specialisation of LRI Core. Not only is it built from the ground up, the methodology by which it is constructed forces a broader, more concrete, and more rigorous definition of concepts and relations. First, the scope of the ontology was determined by selecting a core set of basic concepts. These concepts were organised in modules, and formed the basis for a middle-out construction of the ontology.

16This is not really surprising as there exists an overlap between the developers of LRI Core and LKIF Core.
2. Methodology

The construction of LKIF Core is guided by a combination of methodologies for ontology engineering. Already in the mid-nineties, the need for a well-founded ontology development methodology was recognised, most notably by [36, 37, 74, 73] and later [25]. These methodologies follow in the footsteps of earlier experiences in knowledge acquisition, such as the CommonKADS approach [68] and others, but also considerations from naive physics and cognitive science, such as [41] and [49], respectively.

Hayes, in [41], describes an approach to the development of a large-scale knowledge base of naive physics. Instead of rather metaphysical top-down construction, his approach starts with the identification of relatively independent clusters of closely related concepts. These clusters can be integrated at a later stage, or used in varying combinations allowing for greater flexibility than monolithic ontologies. Furthermore, by constraining (initial) development to clusters, the various – often competing – requirements for the ontology are easier to manage.

Where the domain of Hayes’ proposal concerns the relatively well-structured domain of physics, the combination of common sense and law does not readily provide an obvious starting point for the identification of clusters. In other words, for LKIF Core, we cannot carve-up clusters from a pre-established middle ground of commonsense and legal terms. And furthermore, the field does not provide a relatively stable top level from which top-down development could originate.

In [74], who are the first to use the term ‘middle-out’ in the context of ontology development, it is stressed that the most ‘basic’ terms in each cluster should be defined before moving on to more abstract and more specific terms within a cluster. The notion of this basic level is taken from Lakoff [49], who describes a theory of categorisation in human cognition. Most relevant within the context of ontology engineering [74, 49, p. 12 and 13] are basic-level categorisation, basic-level primacy and functional embodiment. Categories are organised so that the categories that are cognitively basic are ‘in the middle’ of a taxonomy, generalisation proceeds ‘upwards’ from this basic level and specialisation proceeds ‘downwards’. Furthermore, these categories are functionally and epistemologically primary with respect to (amongst others) knowledge organisation, ease of cognitive processing and ease of linguistic expression. Basic level concepts are used automatically, unconsciously, and without noticeable effort as part of normal functioning. They have a different, and more important psychological status than those that are only thought about consciously. This approach thus fits well with the theoretical considerations of the previous section.

For the purpose of the LKIF ontology, we have made slight adjustments to the methodology of [41, 73]. We established design criteria for the development of the LKIF ontology based on [35, 73]. These criteria were implemented throughout the following phases: identify purpose and scope, ontology capture and coding, integration with existing ontologies and evaluation.

2.1. Ontology Capture

The LKIF Core ontology should contain ‘basic concepts of law’. However, it depends on the (potential) users what kind of vocabulary is aimed at, and which concepts are basic. We identified three main groups of users: citizens, legal professionals and legal scholars.
Table 1. Ten highest scoring terms for importance, abstractness, and legal relevance

<table>
<thead>
<tr>
<th>Importance</th>
<th>Abstractness</th>
<th>Legal Relevance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Law</td>
<td>Deontic operator</td>
<td>Civil law</td>
</tr>
<tr>
<td>Right</td>
<td>Law</td>
<td>Law</td>
</tr>
<tr>
<td>Jurisdiction</td>
<td>Norm</td>
<td>Legal consequence</td>
</tr>
<tr>
<td>Permission</td>
<td>Obligative Right</td>
<td>Legislation</td>
</tr>
<tr>
<td>Prohibition</td>
<td>Permissive Right</td>
<td>Obligation</td>
</tr>
<tr>
<td>Rule</td>
<td>Power</td>
<td>Right</td>
</tr>
<tr>
<td>Sanction</td>
<td>Right</td>
<td>Authority</td>
</tr>
<tr>
<td>Violation</td>
<td>Rule</td>
<td>Deontic operator</td>
</tr>
<tr>
<td>Power</td>
<td>Time</td>
<td>Duty</td>
</tr>
<tr>
<td>Duty</td>
<td>Anancastic Rule</td>
<td>Jurisdiction</td>
</tr>
<tr>
<td>Legal Position</td>
<td>Existential Initiation</td>
<td>Legal Fact</td>
</tr>
<tr>
<td>Norm</td>
<td>Existential Termination</td>
<td>Legal Person</td>
</tr>
<tr>
<td>Obligation</td>
<td>Potestative Right</td>
<td>Legal Position</td>
</tr>
<tr>
<td>Permissive Right</td>
<td>Productive Char.</td>
<td>Legal Procedure</td>
</tr>
<tr>
<td>Argument</td>
<td>Absolute Obl. Right</td>
<td>Liability</td>
</tr>
</tbody>
</table>

Although legal professionals use the legal vocabulary in a far more precise and careful way than laymen, for most of these terms there exists a sufficient common understanding to treat them more or less as similar, cf. [51]. Nonetheless, a number of basic terms have a specific legal-technical meaning, such as ‘liability’ and ‘legal fact’. Technical terms were included because they capture the ‘essential’, abstract meaning of terms in law. Furthermore, these terms can be used to structure the relations between more generally understood legal terms.

The Estrella consortium includes representatives of these three kinds of experts. Each partner was asked to supply their ‘top-20’ of legal concepts. Combined with terms we collected from literature (jurisprudence and legal text-books) we obtained a list of about 250 terms. As such a number is unmanageable as a basic set for modelling, we asked partners to assess each term from this list on five scales: level of abstraction, relevance for the legal domain, the degree to which a term is legal rather than common-sense, the degree to which a term is a common legal term (as opposed to a term that is specific for some sub-domain of law), and the degree to which the expert thinks this term should be included in the ontology. The relative position of a term with respect to these scales indicates its appropriateness for inclusion in the basic clusters. For instance, the higher a score for importance, legal relevance, and common legal term, the more appropriate a term is for inclusion in the ontology. On the other hand, a high or low score for abstractness and ‘common-sense vs. legal’ indicates that a term is not basic. For those scales, we were looking rather for terms that have an average score.

The resulting scores were used to select an initial set of 50 terms plus those re-used from other ontologies (see section 1), and formed the basis for the identification of clusters and the development of the LKIF Core ontology. To illustrate, table 1 shows the ten highest scoring terms for three of the scales.
3. Frameworks and Ontologies

Many of the terms we gathered are not necessarily suited for inclusion in the core ontology. Even if scores on the five scales indicate that a term is of the utmost importance, this does not immediately mean that a term, and the relations it has, is of ontological relevance. We adhere to a rather restrictive view on what an ontology should contain: terminological knowledge, i.e., intensional definitions of concepts, represented as classes with which we interpret the world. Definitions that do not meet this rule should not be part of an ontology, and in earlier papers we have consistently advocated a distinction between ontologies and so-called frameworks [12, 44].

The distinction between ontology and actual situations is often said to coincide with the contrast between terminological knowledge (T-Box) and assertional knowledge (A-Box). However, this distinction is not concise enough for our present purposes. As a rule, terminological knowledge is generic knowledge while assertional knowledge describes the (actual) state of some world: situations and events. Such states can be generalised to patterns typical to particular kinds of situations. To be sure, if some pattern re-occurs and has a justifiable structure, it might evidently pay to store its structure as generic description. For instance, it may capture a predictable course of events. The combination of the situations and events related to eating in a restaurant is a typical example, and served in the Seventies to illustrate the notion of knowledge represented by scripts [67] or ‘frames’ [57]. For representations of this kind of generic knowledge, we use the term framework. As we will see, representations of frameworks are structurally different from ontologies.

Whereas frameworks capture the systematic co-occurrence or the structural relations something has with other things, ontologies capture what things are in and of themselves. Where ontologies consist of definitions of terms that have ‘natural’ subsumption (‘is-a’) relations, frameworks describe such things as how activities are causally or intentionally related, or how objects are spatially and functionally configured.

Arguably, at a formal level the two are indistinguishable: every class in OWL is defined in relation to other classes, it cannot be otherwise. This distinction therefore is conceptual: it does not map easily onto representation formalisms. The conceptual distinction is analogous to the difference between intrinsic and accidental relations. We distinguish between those relations that make what a concept is and relations that place a concept in a particular frame of reference, or context. When a context is recurring and sufficiently stereotypical, it might well pay to represent its characteristic features. These features are not real ‘properties’, as they do not ‘define’ a concept, but merely enable its recognition.

Take for example a hammer, its composition of head and shaft is not by accident: it is this particular combination that allows the hammer to be used ‘as a hammer’. However, the mereological relation between the hammer as object, and its composites is not part of its ontological definition. Many different kinds of hammers exist, e.g., sledgehammers, mallets, conveniently shaped stones etc., each of which differs ever so slightly in the nature of its composites and the relations between them. But they are all hammers. It is therefore only the function of the hammer as an instrument that defines what a hammer is, its mereological properties are merely circumstantial evidence. Searle [70] stressed that the hammer-as-such cannot be substituted with the hammer-as-object; whether some object is a hammer is a social fact and depends on an attribution of the hammer func-
tion given a context of use. This attribution pinpoints exactly the conceptual difference between ontologies and frameworks. Where frameworks may incorporate the context in which this attribution holds by default and thus (over)generalise the typical physical features of a hammer-as-object to the hammer-as-such, ontologies should maintain an explicit distinction between the two.

From a methodological point of view, this allows us to introduce a rule of thumb: a combination of concepts and relations, as in e.g. a class restriction, is only to be considered part of an ontology when this particular combination is either systematic and independent of context, or when it makes the context explicit. A possible second consideration is inspired by the limitations imposed on admissible structures by the knowledge representation formalism. Of course this is a rather practical consideration, but nonetheless useful when considering the widespread use of OWL. If the primary task of ontology is to describe what things are, then a representation formalism specifically tailored to classification can be a fair benchmark for determining whether the kind of knowledge you want to express should be considered part of ontology or not. Given the limited expressiveness of OWL DL, especially because of its tree-model property [61], many frameworks are more easily represented using a rule formalism as they often require the use of variables. In fact, as we show in [43], OWL can be used to represent non-tree like structures commonly found in frameworks but only in a very round-about and non-intuitive manner. Furthermore, epistemic frameworks may define epistemic roles which can only be applied by reasoning architectures that go beyond the services provided by OWL DL reasoners (e.g. when they require meta-level reasoning). The limitations of OWL thus indicate a correlation between the conceptual distinction and representation formalisms. However, the two perspectives should not be confounded. Frameworks belong to the T-Box of any knowledge representation system, independent of whether it is based on a DL formalism or not.

We distinguish three types of frameworks:

3.1. Situational frameworks

Situational frameworks are most characteristic for the notion of framework in general, because of the strong emphasis on context and teleology they share with frames and scripts [57, 67]. They are the stereotypical structures we use as plans for achieving our goals given some recurrent context, such as making coffee. These plans are not necessarily private to a single agent, and may involve transactions in which more than one actor participates. For instance, the definition of Eating-in-a-restaurant is a structure consisting mainly of dependencies between the actions of clients and service personnel.

Another notable characteristic of situational frameworks is that they are not subclasses of the goal directed actions they describe. For instance, Eating-in-a-restaurant is not a natural sub-class of Eating but rather refers to a typical model of this action in the situational context of a restaurant. Furthermore, it usually does not make sense to define subclass relations between situational frameworks themselves. Although we can easily envisage a proliferation of all possible contexts of eating – Eating-at-home, Eating-with-family, etc. — but does eating in a French restaurant fundamentally differ from eating in a restaurant in general? [4, 10].

In the following all concepts will start with a capital letter, properties and relations will not
From a legal perspective, situational frameworks can be imposed on actual situations through articles of procedural (‘formal’) law. Although the stereotypical plans given in by custom, and the prescribed plans of law differ in their justification – rationality vs. authority – their representation is largely analogous. Similarly, legal norms combine generic situation descriptions with some specific state or action, where the description is qualified by a deontic term. For instance, the norm that “vehicles should keep to the right of the road” states that the situation in which a vehicle keeps to the right is obliged.

In short, situational frameworks are a fundamental part of the way in which we make sense of the world, and play a prominent role in both general and legal problem solving.

3.2. Mereological frameworks

Most entities, and objects and processes in particular, can be decomposed into several parts; they are composites. As we have seen in the hammer example, it can be tempting to incorporate a mereological view in the definition of a concept. A typical example is the definition of Car as having at least three, and usually four wheels, and at least one motor. However, a full structural description of a concept’s parts and connections goes beyond what it essentially is. Cars are designed with a specific function in mind, and although there are certainly some constraints on its physical properties relevant to its definition, these are limited to those constraints actually necessary for fulfilling that function. Concept descriptions that do iterate over all or most parts of the concept are mereological frameworks, and can appear under a large diversity of names: structural models, configurations, designs, etc. Mereological frameworks play a major role in qualitative reasoning systems (see e.g. [21, 40]).

Arguably, the line between the mereological framework and ontological definition of a term is sometimes very thin. For instance, if we want to describe a bicycle as distinct from a tricycle, it is necessary to use the cardinality of the wheels as defining properties as these are central to the nature of the bicycle. On the other hand, the number of branches a tree might have hardly provides any information as to what a tree is.

3.3. Epistemological frameworks

Inference in knowledge based reasoning seldom occurs in isolation. It is part of a larger structure of interdependencies between various steps in the reasoning process. Such epistemological frameworks focus on the epistemological status of knowledge, e.g. the use of knowledge in reasoning as hypothesis or conclusion. Typical examples are the problem solving methods (PSM) found in libraries of problem solving components [15, 62, 68]. What sets the epistemological framework apart from mereological frameworks is its characteristic specification of dependencies between distinct steps in a reasoning methodology. For instance, a problem solving method is not just a break-down of some problem, but it also specifies control over inferences by assessing success and failure in arriving at (sub)goals in problem solving. They invariably have at least two components: some method for selecting or generating potential solutions (hypotheses), and a method for testing whether the solutions hold. Whether a solution holds depends on whether it satisfies all requirements, or whether it corresponds with, explains, empirical data.
Figure 1. Dependencies between LKIF Core modules.

Not all epistemological frameworks are as specific as problem solving methods. For example, the Functional Ontology of Law of [75] is an epistemological framework describing the roles played by different types of knowledge in law as a system that controls society [14].

4. Ontology Modules

The preceding sections introduce the requirements for LKIF Core and specify a principled approach to (legal) ontology development. This approach is based on insights from cognitive science and uses a well established methodology. With these considerations in mind, the LKIF ontology was initially designed as a collection of eight ontology modules: expression, norm, process, action, role, place, time and mereology, cf. [9]. This collection was later extended with a top ontology, two more ontology modules (legal_action, legal_role), and two frameworks (time_modification and rules), see Figure 1, [44, 12]. Each of these modules contains a relatively independent cluster of concepts, represented using OWL DL in a middle-out fashion: for each cluster the most central concepts were represented first.18

We can distinguish three layers in the ontology: the top level (Section 4.1), the intentional level (Section 4.2) and the legal level (Section 4.3). These layers correspond to the different ‘stances’ one can adopt towards a domain of discourse, and are inspired by the work of Dennett [23] and Searle [70]. Dennet identified three stances we can adopt for explaining phenomena in the world: the physical stance, used for explaining behaviour in terms of of the laws of physics, the design stance, which assumes a system will behave according to its design, and the intentional stance, which can be adopted to explain the behaviour of rational agents, with beliefs, desires and intentions. The first two correspond roughly to the top level of LKIF Core, where the intentional stance is captured by the intentional level. The LKIF ontology thus adds a legal layer, containing concepts that are only sensible from a legal perspective. Accordingly, we represent social and legal

18The ontologies were developed using TopBraid Composer (http://www.topbraidcomposer.com) and Protégé 4.0a (http://protege.stanford.edu).
concepts as social constructs separate from the physical entities having such imposition, e.g. persons are said to play roles, objects fulfil functions [70].

The layers in LKIF Core should not be confused with stratified meta levels; each new layer introduces a straightforward extension of existing definitions. At each level, concepts are expressed in terms of concepts defined at a higher level of abstraction, adding new organising structures (such as properties) where necessary. This methodology ensures a modular set-up that improves reusability and allows extensions to commit to the ontology at any of the levels.

4.1. First Things First: The top-level

The description of any legally relevant fact, event or situation requires a basic conceptualisation of the context in which these occur: the backdrop, or canvas, that is the physical world. Fundamental notions such as location, time, part-hood and change are indispensable in a description of even the simplest legal account. The top level clusters of the ontology provide (primitive) definitions of these notions, which are consequently used to define more intentional and legal concepts in other modules. The most general categories of the LKIF ontology are based on the distinction between ‘worlds’ of LRI Core. We distinguish between mental, physical and abstract entities, and occurrences. Mental entities reside in the human mind, and only have a temporal extension. Physical entities exist independent of human experience, and have a spatial extension as well. Although these categories are superimposed on the concepts in the top level clusters, they were not directive in in their design.

Mereological relations allow us to define parts and wholes; they can be used to express a systems-oriented view on concepts. Examples are functional decompositions, and containment characteristic for many frameworks (Figure 2). Mereology lies at the basis of definitions for places and moments and intervals in time. The ontology for places is based on the work of [24] and adopts the Newtonian distinction between relative and absolute places. A relative place is defined by reference to some thing; absolute places

Figure 2. Place and Mereology related concepts.
are part of absolute space and have fixed spatial relations with other absolute places. A Location Complex is a set of places that share a common reference location, e.g. the locations of all furniture in a room. See figure 2 for an overview of concepts defined in the place module. Of the properties defined in this module, meet is the most basic as it is used to define many other properties such as abut, cover, coincide etc. See [12, 24] for a more in depth discussion of these and other relations. Currently this module does not define classes and properties that express direction and orientation.

Similar relations can be used to capture notions of time and duration. We adopt the theory of time of [1, 2], and distinguish between the basic concepts of Interval and Moment. Intervals have an extent (duration) and can contain other intervals and moments. Moments are points in time, they are atomic and do not have a duration or contain other temporal occurrences (see figure 3). Relations between these temporal occurrences can be used to express a timeline. [1] introduced the meet relation to define two immediately adjacent temporal occurrences. To discern between the temporal meets relation and its spatial counterpart [24], we dub it immediately_before. Where the spatial relation is unrestricted with respect to direction, the temporal meet relation is directed and asymmetric. It is used to define other temporal relations such as before, after, during. Locations and temporal entities are used to define the extension of mental and physical entities, e.g. the time when you had a thought, the location where you parked your car. They are occurrences and do not have extensions themselves, they are extensions.

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

Changes that occur according to a certain recipe or procedure, i.e. changes that follow from causal necessity are Processes. They thus introduce causal propagation and are said to explain the occurrence of change. Processes in LKIF Core are similar to the fluents of event calculus [77]. However, the ontology does not commit to a particular theory of causation and we consider the perspective generic enough to enable the definition of various disparate conceptions of causal relata. Contrary to changes, processes
are both spatially and temporally restricted. They extend through time – they have duration – and are located at some temporal and spatial position. We furthermore distinguish Physical Processes which operate on Physical Objects.

4.2. The Intentional Level

Legal reasoning is based on a common sense understanding that allows the prediction and explanation of intelligent behaviour. After all, it is only the behaviour of rational agents that is governed by law. The modules at the intentional level include concepts and relations necessary for describing this behaviour: Actions undertaken by Agents in a particular Role. Furthermore, it introduces concepts for describing the mental state of these agents, e.g. their Intention or Belief, but also communication between agents by means of Expressions.

The class of agents is defined as the set of things which can be the actor of an action: they may perform the action and are potentially liable for any effects caused by the action (see figure 4). Actions are processes, they are the changes brought about by some agent in realising his intentions. Agents are the medium of an action’s intended outcome: actions are always intentional. The intention held by the agent usually bears with it some expectation that the intended outcome will be brought about: the agent believes in this expectation. The actions an agent is expected or allowed to perform are constrained by the competence of the agent, sometimes expressed as roles assigned to the agent. Because actions are processes, they can play a role in of causal propagation, allowing us to reason backwards from effect to agent. Actions can be creative in that they initiate the coming into existence of some thing, or the converse terminate its existence. Also, actions are often a direct reaction to some other action (see figure 4).

We distinguish between persons, individual agents such as “Joost Breuker” and “Pope Benedict XVI”, and Organisations, aggregates of other organisations or persons which acts ‘as one’, such as the “Dutch Government” and the “Sceptics Society”. Artefacts are physical objects designed for a specific purpose, i.e. to perform some Function as instrument in a specific set of actions such as “Hammer” and “Atlatl”\textsuperscript{19}. Persons are

\textsuperscript{19}An atlatl is a tool that uses leverage to achieve greater velocity in spear-throwing, see http://en.wikipedia.org/wiki/Atlatl
physical objects as well, but are not designed (though some might hold the contrary) and are subsumed under the class of Natural Objects. Note that natural objects can function as tools or weapons as well, but are not designed for that specific purpose.

The notion of roles played an important part in recent discussions on ontology [72, 55, 38]. Roles specify standard or required properties and behaviour of the entities playing the role (see figure 5). However, they not only allow us to categorise objects according to their prototypical use and behaviour, they also provide the means for categorising the behaviour of other agents. They are a necessary part of making sense of the social world and allow for describing social organisation, prescribe behaviour of an agent within a particular context, and recognise deviations from ‘correct’ or normal behaviour. Indeed, roles and actions are closely related concepts: a role defines some set of actions that can be performed by an agent, but is conversely defined by those actions. The role module captures the roles and functions that can be played and held by agents and artefacts respectively, and focuses on social roles, rather than traditional thematic or relational roles.

A consequence of the prescriptive nature of roles is that agents connect expectations of behaviour to other agents: intentions and expectations can be used as a model for intelligent decision making and planning.20 It is important to note that there is an internalist and an externalist way to use intentions and expectations. The external observer can only ascribe intentions and expectations to an agent based on his observed actions. The external observer will make assumptions about what is normal, or apply a normative standard for explaining the actions of the agent.

The expression module covers a number of representational primitives necessary for dealing with Propositional Attitudes (viz. [19]). Many concepts and processes in legal reasoning and argumentation can only be explained in terms of propositional attitudes: a relational mental state connecting a person to a Proposition. However, in many applications of LKIF the attitude of the involved agents towards a proposition will not be relevant at all. For instance, fraud detection applications will only care to distinguish between potentially contradictory observations or expectations relating to the same propositional

---

20Regardless of whether it is a psychologically plausible account of decision making. Daniel Dennett’s notion of the Intentional Stance is interesting in this context (cf. [23]). Agents may do no more than occasionally apply the stance they adopt in assessing the actions of others to themselves.
content. Examples of propositional attitudes are Belief, Intention, and Desire. Each is a component of a mental model, held by an Agent.

Communicated attitudes are held towards expressions: propositions which are externalised through some medium. Statement, Declaration, and Assertion are expressions communicated by one agent to one or more other agents. This classification is loosely based on Searle (cf. [69]). A prototypical example of a medium in a legal setting is e.g. the Document as a bearer of legally binding (normative) statements.

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

Evaluative Attitudes express an evaluation of a proposition with respect to one or more other propositions, they express e.g. an evaluation, a value statement, value judgement, evaluative concept, etc. I.e. only the type of qualification which is an attitude towards the thing being evaluated, and not for instance the redness of a rose, as in [27] and others. Of special interest is the Qualification, which is used to define norms based on [7]. Analogous to the evaluative attitude, a qualification expresses a judgement. However, the subject of this judgement need not be a proposition, but can be any complex description (e.g. a situation).

4.3. The Legal Level

Legally relevant statements are created through public acts by both natural and legal persons. The legal status of the statement is dependent on both the kind of agent creating the statement, i.e. Natural Person vs. a Legislative Body, and the rights and powers attributed to the agent through mandates, assignments and delegations. At the legal level, the LKIF ontology introduces a comprehensive set of legal agents and actions, rights and
powers (a modified version of \cite{66, 65}), typical legal roles, and concept definitions which allow us to express normative statements as defined in \cite{7, 5, 6}.

The Norm is a statement combining two performative meanings: it is deontic, in the sense that it is a qualification of the (moral or legal) acceptability of some thing, and it is directive in the sense that it commits the speaker to bringing about that the addressee brings about the more acceptable thing (cf. \cite{64}), presumably through a sanction. These meanings do not have to occur together. It is perfectly possible to attach a moral qualification to something without directing anyone, and it is equally possible to issue a directive based on another reason than a moral or legal qualification (e.g. a warning).

A norm applies to (or qualifies) a certain situation (the Qualified situation), allows a certain situation – the Obliged situation or Allowed situation – and disallows a certain situation – the Prohibited or Disallowed situation, see Figure 7. The obliged and prohibited situation are both subsumed by the situation to which the norm applies.\footnote{Normatively qualified situations are analogous to the generic situations, or generic cases of \cite{75}.} Besides this, they by definition form a complete partition of the case to which the norm applies, i.e. all situation to which the norm applies are either a mandated case or a prohibited situation. This is true of the obligation and the prohibition: they are simply two different ways to put the same thing into words. The permission is different in that it allows something, but it does not prohibit anything. The logical complement of the mandated situation is here an opposite qualified situation, about which we know only that it cannot be obliged.

Where in other approaches, cf. CLO, a situation is the reification of some state of affairs, the normatively qualified situations in LKIF Core are instantiated by states of affairs: they are defined as class descriptions that represent a set of possible states of affairs. This means that a standard reasoner can infer whether some actual situation is subsumed under a generic situation, and thus whether norms exist that allows or disallows that situation. Similarly, a classifier will create an inferred hierarchy of situations, which enables a relatively straightforward resolution of \textit{lex specialis} exceptions between norms \cite{7}.

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{figure7.png}
\caption{Qualifications and Norms}
\end{figure}
The next section gives a brief example of how this mechanism can be applied to build representations of actual normative systems.

5. The Ontology in Practice

The LKIF ontology not only provides a theoretical understanding of the legal domain, but its primary use in practice is as a tool to facilitate knowledge acquisition, exchange and representation: i.e. to formalise pieces of existing legislation. The first, comprehensive evaluation of LKIF Core was done in the formalisation of EU Directive 2006/126 on driving licences\(^ {22} \) a relatively straightforward regulation in which at least two types of normative statement are recognisable – definitional and deontic. An example of a definitional statement from the directive is:

\[
\text{Art. 4(2) Category AM}
\]
Two-wheel vehicles or three-wheel vehicles with a maximum design speed of not more than 45 km/h.

The mereology module of the ontology along with a qualified cardinality restriction (available in OWL 2) allows us to express that AM vehicles have two or three wheels:

\[
\text{AM} \sqsubseteq \text{composed_of exactly } 2 \text{ Wheel} \sqcup \text{composed_of exactly } 3 \text{ Wheel}
\]

Modelling ‘design speed not more than 45 km/h’ is more challenging as it requires us to represent the domain of speeds and distances. Of course, one could introduce the datatype property \text{designSpeed} and require its value be expressed in km/h. Although the set of built-in datatypes of OWL 2 does not include km/h, we can define a custom datatype for speeds. However, the representation of speed using a concrete domain does not do justice to its conceptual complexity, as it involves several other notions: unit of measurement, number, designed speed, and a ‘no-more’ relation. Furthermore, a custom datatype can only impose a syntactic check on a value. As a speed-value is syntactically no different than the values covered by built-in datatypes such as \text{xsd:integer} and \text{xsd:float}, introducing a custom datatype does not extend expressive power.

In fact, for the current purposes, the concept can be represented by imposing a linear ordering relation less-than on the different (instances of the) subclasses of a class DesignSpeed. For any \( n \leq m \), we have that:

\[
\text{DesignSpeed-km-h-}n(a) \text{ less-than DesignSpeed-km-h-}m(b)
\]

with \( a, b \) instances. This allows us to define the class of those DesignSpeeds with a value not exceeding that of the class of speeds with the exact value ‘45’, i.e. less-than only DesignSpeed-km-h-45. The representation of definitional statements using the LKIF ontology is thus somewhat limited because of the broad range of domain concepts possibly involved in a definition. Of course, a complete coverage of all domains cannot be expected of a core ontology. Nonetheless, the representation of article 4(2) suggests the need for an extension of the LKIF ontology with a module that defines quantities, units of measurement, numbers, fractions, mathematical operations, etc. This is crucial not only for the EU Directive 2006/126, in which most definitional statements contain quantita-

\[^{22}\text{The text is available on-line at } \text{http://eur-lex.europa.eu/}.\]
tive features of vehicles (e.g. power, cylinder capacity), but quantities and calculations
play a central role in most, if not all legislative texts. We are currently evaluating possible
inclusion of readily available unit ontologies.23

On the other hand, deontic statements express norms in terms of the concepts intro-
duced by definitional statements. For instance,

\textbf{Art. 4(2)}

The minimum age for category $\textit{AM}$ is fixed at 16 years.

This article expresses an obligation whose logical form can be rendered by an implica-
tion:

If $x$ is driving a $\textit{AM}$ vehicle, then $x$ \textbf{must} be at least 16 years old.

Here, the antecedent is the \textit{context} to which the obligation applies; the consequent (mi-
nus the deontic operator \textbf{must}) is the \textit{content} of the obligation itself, i.e. that which the
obligations prescribes it ought to be the case. Consistent with this analysis, the LKIF
ontology defines obligations as classes (see Section 4.3).

The article allows situations that match the description $\text{DriverAM} \cap \text{DriverOlderThan16}$ and forbids $\text{DriverAM} \cap \neg \text{DriverOlderThan16}$. Given the presence of straight-
forward definitions of $\text{DriverOlderThan16}$ and $\text{DriverAM}$,24 we can define the obligation
that drives of $\text{AM}$ vehicles must be at least 16 years or older as follows:

\begin{align*}
\text{MinAgeAM} \sqsubseteq & \text{allows only } (\text{DriverAM} \cap \text{DriverOlderThan16}) \\
\text{MinAgeAM} \sqsubseteq & \text{allows some } (\text{DriverAM} \cap \text{DriverOlderThan16}) \\
\text{MinAgeAM} \sqsubseteq & \text{disallows only } (\text{DriverAM} \cap \neg \text{DriverOlderThan16}) \\
\text{MinAgeAM} \sqsubseteq & \text{disallows some } (\text{DriverAM} \cap \neg \text{DriverOlderThan16})
\end{align*}

Other deontic operators, such as permission or prohibition, can be accounted in an alike
manner (see [6]). Notwithstanding the parsimony of this type of definition, using the
LKIF ontology to model normative statements proves to be rather straightforward. Of
course, a specialised modelling environment for legislative drafters would need to pro-
vide a shorthand for the standard patterns used in OWL definitions of norms. The norm
module is currently used in the development of a legal assessment plugin to Protege 4.

Use of the ontology in practice not only called for more detailed domain modelling,
as was to be expected, but it also required us to explore the limits of OWL 2’s expres-
siveness, in particular in the representation of the complex structure of transactions and
situations [43]. Furthermore, as versioning and applicability play a central role in legal
reasoning, we evaluated the the ontology in the representation of intricate versioning of
definitions and norms in the Nomic game in [48].

---

23E.g. the ontology by Michel Dumontier, or a translation of the PHYSYS ontology in OWL DL. See \url{http://www.dumontierlab.com/}, \url{http://www.ksl.stanford.edu/software/ontolingua/}, respectively.

24The class $\text{DriverOlderThan16}$ can be defined by using a \textit{more-than} ordering relation, roughly along the same lines as the class \textit{less-than onlyDesSpeed-km-h-45}. 
6. Discussion

We presented a principled approach to ontology development in the legal domain, and described how it was applied in the development of the LKIF Core ontology. The initial requirements for this ontology, as part of LKIF, are that it should be a resource for special, legal inference, and that it provides support for knowledge acquisition. The legal domain poses additional requirements on the ontology in that it presupposes a common sense grounding. We discussed several other ontologies, argued where they fall short in providing this grounding, and presented an approach based on insights from cognitive science and knowledge engineering, rather than philosophy. This approach was further discussed as the basis for a comprehensive methodology for ontology development, including a principled distinction between frameworks and ontologies. The LKIF Core ontology was developed using this methodology, and is composed of fourteen modules, two of which represent frameworks. We briefly introduced the main concepts in these modules and gave an example of how the ontology can be used in modelling an EU directive.

As LKIF Core was developed by a heterogeneous group of people, we specified a number of conventions to uphold during the representation of terms identified in the previous phases [12]. One of these is that classes should be represented using necessary & sufficient conditions as much as possible (i.e. by means of owl:equivalentClass statements). Using such ‘complete’ class definitions ensures the ability to infer the type of individuals, which is not possible by means of partial class definitions (using only necessary conditions).

In retrospect, this convention turned out to pose severe problems for existing OWL 1 and OWL 2 DL reasoners as their performance is significantly affected by the generic concept inclusion axioms (GCI): axioms where the left-hand side of a rdfs:subClassOf statement is a complex class definition. These axioms are abundant when defining classes as equivalent to e.g. owl:someValuesFrom restrictions and in combination with lots of inverse property definitions, this creates a large completion graph for DL reasoners.25 As result of these findings, the LKIF ontology has underwent a significant revision since its initial release. However, reasoner performance remains an issue with large, complex ontologies such as LKIF Core. Accurate representations for the purpose of legal reasoning push the limits of OWL even further, and we expect some optimisation to be necessary before a legal assessment system based on this technology can be used in practice.

Using LKIF Core to represent regulations, as e.g. in the traffic example, furthermore points to the traditional knowledge representation threshold: for any formal representation of any domain, one needs to build formal representations of adjoining domains. As has been said, this can be largely overcome by including other generic or domain ontologies in a representation based on the LKIF ontology; provided that the quality of these ontologies is sufficient. Depending on availability we might consider providing a library of ‘compatible’ ontologies to users of LKIF Core. This will be of especial use when the ontology vocabulary is adopted for expressing the LKIF vendor models that are currently being developed within the Estrella project. Recent experience suggests that the ontology should be extended with not only a module for expressing measurements, but also frame-

25Thanks to Taowei David Wang for pointing this out, see http://lists.owldl.com/pipermail/pellet-users/2007-February/001263.html
works that allow us to adequately capture transactions and other more complex action structures.

With respect to coverage of the legal domain, the purpose of the study outlined in Section 2.1 is more ambitious than only the selection of the most basic terms for describing law, but time and effort constraints make it that we could only consider a small pool of referents. To achieve best results, the list of terms could be subjected to a more rigorous empirical study, e.g. by consulting groups of legal professionals (taking courses in legal drafting) and law students. This empirical study is planned in the sideline of Estrella. By applying statistical cluster analysis, we might be able to identify the properties of the scales used (e.g. are they independent?) and determine whether the statistical clusters have some resemblance to the clusters we have identified based on more theoretical considerations.

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

References


[60] Boris Motik, Peter F. Patel-Schneider, and Ian Horrocks. OWL 2 web ontology language: Syntax. Pre-working draft, W3C, 2008. (to be published, may be superseded).


Introducing pattern-based design for legal ontologies

Aldo Gangemi

Semantic Technology Lab, ISTC-CNR, Roma, Italy
aldo.gangemi@cnr.it

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

Keywords. legal knowledge engineering, ontology design, design pattern

1. Introduction

Semantic technologies are spreading out through the Semantic Web and Web of Data programmes [6,5]. Legal practice can take advantage from them for example in the form of semantic services [45,56], Linked Open Data [5], or knowledge-based intrawebs [31, 33], directed towards legal experts, citizens, institutions, and companies.

The core of semantic technologies is constituted by ontologies, which are logical theories that formalize the assumptions underlying domain knowledge, describing for example physical, social, and properly legal objects or relations, such as legal procedures, norms, roles, contracts, etc. Ontologies are usually expressed in first-order languages or fragments of them, although some typical modal and meta-level primitives are usually added to them, for example in description logics like OWL-DL [10].

Ontologies can be designed by means of methods such as those described in [34,13,60,49,11], encompassing top-down expertise elicitation from humans, bottom-up learning from documents, and middle-out application of design patterns, which can be specialized from domain-independent ontologies, extracted from best practices, existing ontologies or other knowledge resources, as well as learnt from conceptual invariances found in experts’ documents. This chapter focuses on the use of Content Ontology De-
sign Patterns (CP) [22] in the design of legal ontologies. A CP is defined as a reusable solution to a recurrent content-oriented (or “conceptual”) ontology design problem.1

The middle-out approach is useful for legal ontologies specially, because the legal domain is relatively complex compared to others: it involves knowledge of the physical and social worlds, as well as typical legal knowledge that creates a novel layer over the social world [54,46].

Due to the autonomy (on one hand) and dependence (on the other hand) of legal knowledge on both physical and social knowledge, legal reasoning tasks have evolved in a peculiar way, which include for example the norm structure based on conceptual schemes (that can be represented as CPs), like Requirement $\rightarrow$ Consequence (if the factual knowledge is $P$, then the legal knowledge is $Q$), Obligation $\rightarrow$ Right (if an agent $a$ has an obligation towards an agent $b$, then $b$ has a right towards $a$), Norm $\rightarrow$ Case (if a situation fulfils the conditions for violating a norm, it becomes a legal case), Crime Scenario (a crime is committed by a perpetrator and comes to the attention of authorities that pursue a criminal process), etc. We will show that the conceptual schemes that are assumed by legal experts can be formalized as CPs by specializing or composing other existing patterns that are more generally applicable to the social world.

This work introduces some use cases for legal ontologies;2 as well as some CPs that can be specialized to support ontology-driven solutions to those use cases. The CPs are based on contributions to an ontology design patterns repository [1,50,49], started within the NeOn project [3], and is currently evolving as a community-based effort. The CPs included in the Appendix are extracted from the Core Legal Ontology (CLO) [30,20], and JurWordNet [47], which are partly based on the DOLCE+DnS Ultralite ontology [21,42,23].

In section 1. some ontology design/engineering use cases in the legal domain are introduced. In section 2. pattern-based ontology design is presented. In section 3. the Core Legal Ontology is briefly summarized and some legal CPs are sketched with reference to a use case.

2. Legal Ontology Engineering: Functionalities and Techniques

In information and communication technology, ontology design is dependent on (ontology) engineering applications, which implement techniques and tools that help achieving typical functionalities. For a comprehensive framework of ontology design, and its relations to content, related data, formal languages, design patterns, social practices, organizations, teams, and functionalities, see [12,27]. The ontology encoding of a metamodel for describing collaborative ontology design activities and data can be found at [19].

Ontology engineering deals with designing, managing, and exploiting ontologies within information systems. Besides specific ontology design tools, ontology engi-

---

1There is a rather large literature on artifacts that have something in common with CPs, starting at least with Bartlett’s and Piaget’s schemata, and proceeding to Fillmore’s and Minsky’s frames, Shank’s scripts, Hayes’ clusters, Hay’s datamodel patterns, Clark and Porter’s knowledge patterns, Breuker’s epistemological frameworks [36] (see also this book), etc. Some details on the difference between CPs and related work are contained in [22,50,24]; however, the basic idea that makes CPs both a pragmatic and rigorous approach is their dependency on competency questions [34], which allows to define a method for pattern-based design.

2Use case is here associated with close notions such as requirement, task, etc.
neering applications are usually hybridized with other components in order to build semantically-explicit applications; for example, when ontologies are used jointly with:

- **theorem provers** [16], *consistency checking* can be performed to logically validate the set of assumptions encoded in an ontology
- **subsumption and instance classifiers** [8], as well as *materialization engines* [53] against a logical language of known and manageable complexity, like OWL (in the Lite and DL species) [61], *automatic inferences* can be derived for *taxonomical reasoning* as well as for *instance classification*; moreover, *materialization of assertions* can be inferred from for example inverse, symmetric and transitive relations (see [29], [30] for examples with legal knowledge bases)
- **computational lexicons** with NLP tools and *machine-learning algorithms* [11, 39], legal ontologies can enhance *information extraction* from semi-structured and non-structured data, adding a new dimension to knowledge management and discovery in Law [32]
- **planning algorithms** [44], ontologies can assist or automatize *negotiation or execution* for example for *contracts, regulations, services*, etc. An application of planning and legal ontologies in the IPR domain is [52]
- **case-based reasoners** [41], ontologies can *formalize case abstractions* within more general frameworks, or can *classify cases* according to pre-designed descriptions [17]. A legal example of integrating case-based reasoning with ontologies is [59]
- **rule-based engines** (e.g. SWRL, [38]), facts can be inferred for example for *causal responsibility assessment, conformity checking, conflict detection* and in general for *fact composition*. [29] applies ontologies with rules for reasoning over (static representations of) norm dynamics
- **query engines**, facts can be inferred similarly to what can be done with rule-based engines, such as with SPARQL *CONSTRUCT* queries [51]

All these components constitute a map of *reasoning ontology patterns* [50] that can be exploited to define the reasoning space of a legal semantic application (see also Fig. 2). In order to establish what reasoning patterns can be used with what content patterns (and logical languages) for what application, the domain requirements of an application must be specified. *Competency questions* [34] are usually quite efficient in encoding domain requirements. Competency questions can be coupled with pattern-based design by abstracting some “generic use cases” [22]. The main types of use cases that can be implemented or assisted by means of semantically-explicit applications in the legal domain (Fig. 1) are summarized in the following.

**Intersubjective agreement and meaning negotiation**

*Definition:* the task of getting consensus (or of discovering disagreement) about the intended meaning of a legal term, legal text unit, etc.

*Approach:* the formal representation of (part of) the intended meaning assumed by each of the parties involved in the task.

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

**Knowledge extraction**

*Definition*: the process of extracting concepts, relations, named entities, and complex knowledge patterns from a database, a document, or a corpus.

*Approach*: data- and text-mining, machine-learning, and NLP algorithms that can extract linguistic objects from a corpus, and semi-automated methods that match them to semantic objects.

*Issues*: this task is highly incremental, because the approaches need a training phase or an extensive data entry procedure, so that the extracted knowledge can be used to build a repository of patterns that can be used to improve further extraction processes. Best results can be achieved on very large corpora (for statistical reasons), or on well-delimited, possibly semi-structured corpora and tasks; for example, typical expressions that are found in legal drafting can be used to formalize expectation patterns in corpora consisting of homogeneous texts [9]. Additional issues for knowledge extraction in Law come from the possible heterogeneity of reusable resources (extracted text fragments, thesauri, datamodels, database schemas, etc.). A metamodel (called *semion*), and a procedure to apply reengineering ontology design patterns to this purpose is presented in [24].

**Conformity checking**

*Definition*: the task aimed at verifying if a social situation satisfies a legal description (norm, principle, regulation, maxim, contract, etc.). Also situations already known to be legally relevant for some reason (e.g. a crime situation) can be checked for conformity against another legal description (e.g. an appeal judgment procedure).

*Approach*: the formal representation of social or legal situations as well as of legal regulations. Reasoners to cluster/classify/abstract situations.

*Issues*: on the reasoning side, the typical inferences supported by semantic web engines for OWL [10] (e.g. FaCT++ [37]) currently include only concept subsumption and instance classification. The expressivity is also limited so that for example fact classification (also called materialization) is only performed by some engines (e.g. Pellet [58]). Moreover, compositions between facts cannot be inferred unless an additional rule or query engine is added (extensions supporting fact composition are provided by implementations of languages such as SWRL[38], SPARQL [51], F-Logic [40]). Moreover, clustering and abstraction of situations requires different reasoners, for example induction engines [9] and case-based reasoners [17]. Finally, approximate inferences [15] should also be supported in the generalized case of partial knowledge about situations. On the representation side, a homogeneous language to represent both situations and the constraints from a legal description is highly desirable, otherwise a higher-order logic would be required to express constraints on constraints on constraints etc. on situations.
The proposal in [30], shows a viable approach to represent both constraints and instance data in a same, partitioned first-order domain of quantification. See [23] for an extensive ontology of norms and plans that follows this approach.

**Legal advice**

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

*Approach:* subsumption/instantiation classification, rule-based or case-based reasoning. Most rule-based and case-based systems in the field of AI&Law have concentrated on this kind of reasoning patterns [55]. The hybridization with semantic technologies is not yet established though.

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

**Norm comparison**

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

*Approach:* approximate classification algorithms (i-ii), including legal text annotation and classification on large corpora (ii). [30] and [29] show simpler approaches to the classification of norm dynamics and conflicts within a finite set of norms after their first-order encoding. *Issues:* in Civil Law corpora, the task (ii) is sometimes relevant as much as in Common Law corpora, because of the stratification of laws that do not explicitly delete or even refer to previous ones (e.g. in the Italian legal system). In Common Law, implicit relations can be discovered more easily, because case abstraction has always a clear reference to a case, while in Civil Law, implicit relations appear mostly at the normative level.

**Norm rephrasing**

*Definition:* expression of norms’ content in different terms, which can be either translations in a different natural language, or in a different form within a same natural language, for example for the purpose of popularization.

*Approach:* translation between different languages requires a preliminary mapping between terms, like the EuroWordNet-oriented work performed in the LOIS project [48], classifications based on statistical NLP techniques, and subsumption classification for a close matching between content patterns and linguistic patterns. Popularization requires also the mapping between expert and naive content patterns.

*Issues:* Norm rephrasing across different legal systems has specific problems due to conceptual heterogeneity: different content patterns apply in different legal systems for a same domain, for example national application of European directives. An approach to deal with this heterogeneity is reported in [7].
**Contract management and execution**

Definition: a service assisting parties in the tasks of managing contract agreement and definition, and of following contract execution.

Approach: the semantic specification of contract content, as well as algorithms to manage the matching of parties’ constraints and preferences, and a planning algorithm for the generation of optimal obligations that parties could undertake [52].

**(Information) service matchmaking and composition**

Definition: operations carried on the description of services, in order to check for example if an offered service matches the requested service, or to orchestrate two services to get a more complex one.

Approach: in the legal domain, these tasks require the semantic specification of services with reference to the legal knowledge involved in the execution of the service. Appropriate reasoners and planners are required.

Issues: in a lightweight form, semantic matchmaking and composition of services and contracts are typical eGovernment tasks. However, applying legal semantic applications to eGovernment requires the integration of legal ontologies with governmental and social ontologies.

3. Content Ontology Design Patterns

Semantically-explicit applications in the legal domain present us with conceptual analysis and integration problems that require appropriately designed legal formal ontologies. Part of the design problems can be simplified by creating or extracting “Content Ontology Design Patterns” (CPs) [22,49], for a domain of application (see e.g. for legal and biomedical ontologies: [30] [25]). An intuitive characterization of CPs is provided here:

- A CP is a schema to represent, and possibly solve, a modelling problem. For example, a Norm ↔ Case CP (Fig. 4) facilitates the modelling of legal norms and cases (as well as their components and dependencies) in logical languages that require constraint reification. For example in OWL, relations with an arity > 2 are not allowed, therefore, as an approximate solution, OWL modelling requires a reification of those relations. A CP (the situation’ pattern [4]), extracted from the DnS ontology [23], provides a basic vocabulary for this type of reification vocabulary, and it is specialized in the Norm ↔ Case pattern.

- A CP can be extracted as a module [14] from a reference ontology, which in that case constitutes its “background”. For example in the Norm ↔ Case CP, a foundational distinction is reused from DnS, while the cardinalities for the relations are provided by the Core Legal Ontology (CLO, [30] [20]). DnS and CLO together form a background for the Norm ↔ Case CP. A CP can be used with reference to its background, or independently.

---

A. Gangemi / Introducing Pattern-Based Design for Legal Ontologies
A CP can be represented in any formal representation language, but its intuitive and compact visualization is an essential requirement. It requires a critical size, so that its diagrammatical visualization is aesthetically acceptable and easily memorizable. For example, the Norm → Case CP only includes eight classes, with several, systematic relations between them: this makes it self-connected, but manageable.

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

A CP should be intuitively exemplified, and should catch relevant, “core” notions of a domain. Independently of the generality at which a CP is defined, it must contain the central notions and best practices that “make rational thinking move” for an expert in a given domain for a given task. Typically, experts spontaneously develop schemata to improve their work, and to store relevant know-how. These schemata can be reengineered with appropriate methods (e.g. [26]).
• A CP can be similar to a database schema, but a CP has a general character, independently of system design. In this sense, it is closer to so-called data modelling patterns [35].

Content Ontology Design Patterns (CPs) are a resource and design method for engineering ontology content, currently exploited mainly over the Semantic Web [1,49].

A template (Fig. 3), also available for OWL ontologies specifically from [2]), can be used to annotate CPs as sub-theories of reference ontologies, in order to share them in pre-formatted documents, as well as to describe, visualize, and make operations over them appropriately.

The (general, not specific for OWL) CP template consists of:

• Two slots for the generic use case, and the local use cases, which include a description of context, problem, and constraints/requirements.
• Two slots for the addressed logic, and the reference ontologies used as a background for the pattern.
• Two slots for -if any- the specialized pattern and the composed patterns (by inheritance, and inverse inheritance, it’s possible to obtain the closure of specialized patterns).
• Two slots for the maximal relation that encodes the case space, and its intended axiomatization: a full first-order logic with meta-level is assumed here, but the slot can be empty without affecting the functionality of a CP frame.
• Two slots for explanation of the approach, and its encoding in the logic of choice.
• A last slot for a class diagram that visually reproduces the approach.

The high reusability of CPs and their formal and pragmatic nature make them suitable not only for isolated ontology engineering practices, but ideally in distributed, collaborative environments like intranets and the Web. CPs can also be used to generate intuitive, friendly UIs, which can present the user with only the relevant pattern diagram, avoiding the awkward, entangled graphs currently visualized for medium-to-large ontologies.
The advantages of CPs for ontology lifecycle are straightforward: firstly, patterns make ontology design easier for both knowledge engineers and domain experts (as when having a menu of pre-built, formally consistent components, pro-actively suggested to the modeller); secondly, pattern-based design makes it easier ontology mapping - perhaps the most difficult problem in ontology engineering. For example, the time-indexed participation presented in this paper requires non-trivial knowledge engineering competence to be optimally represented and adapted to a use case: a CP within an appropriate ontology management tool can greatly facilitate such representation.

The CP examples and the related frame and methods introduced in this paper have been applied for years in several administration, business and industrial projects, for example in fishery information systems [26], insurance CRM, biomedical ontology integration [25], anti-money-laundering systems for banking information systems [29], service-level agreements for service providers, biomolecular ontology learning (Ciaramita et al 2005), legal norms formalization [30,57], etc.

Current work focuses on building a tool that assists development, discussion, retrieval, and interchange of CPs over the Semantic Web, and towards establishing the model-theoretical and operational foundations of CP manipulation and reasoning. In particular, for CPs to be a real advantage in ontology lifecycle, the following functionalities are being developed in the NeOn project [3]:

- Categorization of CPs, based either on the use cases they support, or on the concepts they encode (having a task-based component library available).
- Pattern-matching algorithms for retrieving the CP that best fits a set of requirements, for example from a natural language specification, or from a draft ontology (having a component matcher available).
- Support for specialization and composition of CPs. A CP \( p_2 \) specializes another \( p_1 \) when at least one of the classes or properties from \( p_2 \) is a sub-class or a sub-property of some class resp. property from \( p_1 \), while the remainder of the CP is identical. A CP \( p_3 \) composes \( p_1 \) and \( p_2 \) when \( p_3 \) contains both \( p_1 \) and \( p_2 \). Notice that CPs –differently from “knowledge patterns” in [13], which are characterized as invariant under signature transformation and including some rule– are intended to be downward conservative under signature transformation, meaning that CP semantics is structure-preserving when a CP is specialized or composed, but this conservativeness holds only in the downward taxonomical ordering. On the contrary, logical ontology design patterns are conservative under signature transformation both down- and up-wardly [50] (having a CP processor available).
- Interfacing of CPs for visualization, discussion, and knowledge-base creation (having a user interface available for pattern-based design).
- Implementing a procedure that assists a designer in transforming use cases into unit tests that are then used to validate the pattern-based ontology being produced (having a built-in evaluation available at design time).

Knowledge patterns from [13] are therefore logical patterns, for which a specific vocabulary and at least one rule have been provided. However, the vocabulary of knowledge patterns is specified at the categorical level, therefore any transformation does not impact on downward conservativeness of CPs, because this is defined as a model-theoretic (not categorical) vocabulary.
4. The Core Legal Ontology and its Related Patterns

The need for an extended typology of legal entities is becoming a pressure, even from traditionally “bottom-up” approaches. For example, the need to pair case-based reasoning with an ontology of first-principles should be investigated in order to represent the two kinds of structures employed in reasoning: abstraction from cases, and satisfaction of constraint sets (e.g. norms) [17].

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

The Core Legal Ontology (CLO) [30] [20] is developed on top of DOLCE [42,21], the Social Roles extension of DOLCE [43], and Descriptions and Situations [23]. CLO allows for the representation of first principles (by means of a rich axiomatization), and granularity (by means of its reification vocabulary and axioms) in the legal domain.

The two pillars of CLO are: stratification and reification.

Based on the stratification principle, CLO provides types and relations for the heterogeneous entities from the legal domain, be it about the physical, cognitive, social, or properly legal worlds (cf. [46,30]). According to stratification, entities from different layers can be spatio-temporally co-located, yet being different and (mutually or one-way) dependent. For example, a physical person pertains to the physical world as a biological organism, but the properties of the organism are not sufficient to characterize it as a social person. On its turn, the properties of a social person are not sufficient to characterize it as a legal person. Clearly, there are dependencies among those properties, but if the properties from each layer are simply summed up in a same entity, an ontology designer can get undesirable results, for example it would be possible to infer that an organism (physical layer) can be “acting” after its death because of the legal existence (legal layer) of a person until its legal effects disappear.

In DOLCE, stratification is a general pattern, expressed as a disjointness axiom between the class of physical vs. social objects. Legally-relevant entities are mostly in the social realm (cf. the figures in the appendix, which also show some use of CLO for the JurWordNet lexical ontology [47,30]). Within one layer, disjointness axioms help carving the conceptual structure of a domain. Among social objects, CLO introduces further disjointness axioms: legal descriptions, concepts, situations, persons, and information objects are all mutually disjoint. Among legal descriptions, constitutive vs. regulative descriptions are distinguished from principles, rationales, modal descriptions (e.g. duties, powers, liabilities, etc.), as well as from mixed regulations such as contracts and bundles of norms. Among legal facts, natural, human, cognitive, and strictly legal cases are distinguished. Legally-relevant circumstances are further distinguished from legal facts as being ancillary to primary facts. Among legal persons, organizations, natural persons, and legal subjects are also distinguished.

Based on the reification principle, CLO enables to quantify either on legal rules or relations (type reification) [43,18], or on legal facts (token reification) [18,28]. CLO extends the Descriptions and Situations vocabulary for reification. For example, intensional specifications like norms, contracts, subjects, and normative texts can be represented in the same domain as their extensional realizations like cases, contract executions, agents, physical documents.
Another CP (called description, citeDESCP), also extracted from DnS, models the structure of an intensional specification, called a description in [43,23], as composed of its concepts and their internal dependencies. For example, the structure of a norm (a legal description) employs the description CP by specializing it. Another pattern models the matching between a description and its extensional realization, called a situation in [28,23], which can be described as the configuration of a set of entities according to the structure of a description. A legal application of this CP can be found in the dependencies among the rules in a contract, when they can be matched to a legal case (a legal situation or fact), or a contract execution. The matching is typically performed when checking if each entity in a legal fact is compliant to a concept in a legal description.

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

![Figure 3. The Norm ↔ Case CP: norms use tasks, roles, and parameters; legal cases conform to norms when actions, objects and values are classified by tasks, roles, and parameters respectively. Moreover, relations between legal roles, tasks and parameters correspond to relations between objects, actions and values. For example, an obligation for a role towards a task should correspond to a participation of an agent (object) in an action; a spatial parameter that is requisite for an object should correspond to an exact location of an object in a spatial value region that is classified by that parameter.](image)

In the following CP template (Tab. 1), the Limit ↔ ViolationCase CP is compactly introduced as a specialization of the generic Norm ↔ Case CP:

### 5. Conclusions

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

For some generic LKE use cases (conflict checking, knowledge extraction, etc.), some solutions and issues, both on the reasoning and (content) modelling sides, have been mentioned. While the reasoning side of LKE is a fast-moving target, with interesting solutions coming from for example hybridizing different inference engines and classifiers, the modelling side is far less developed, despite the huge literature that focuses on
A. Gangemi / Introducing Pattern-Based Design for Legal Ontologies

Slot: Value

Legal situations to be checked for (non)conformity to existing norms that establish limits.

Generic use case

Local use cases

Logics addressed

OWL, DL species

Specialized CP

Composed CPs

Table 1.
legal and jurisprudential content, let alone the work in formal ontology and beyond. This is not surprising. Currently, very few ontologies are actually reused, against the great expectations that have been grown in the field of reusability of semantic components.

In the past, the need for structures that grant systematicity to content modelling have mainly focused on so-called top-level ontologies, but the power of a small set of categories is not enough for realistic ontology projects like those presented to LKE. A top-level can even be a problem when its categories are brittle with respect to the domain task. A more sophisticated approach, which ensures a much higher level of cognitive interoperability, is constituted by foundational ontologies like DOLCE. Nonetheless, although their rich axiomatization makes foundational ontologies ideally suited for building a partial-order of reference ontologies, they require a substantial cognitive load to be accessed and then successfully reused.

A new dimension of reusability has been introduced in ontology engineering (and here extended to LKE) which revisits some good practices from AI (knowledge patterns) and databases (data model patterns), by providing a meta-model, some operations, and a generalized characterization to content ontology design patterns (CPs). A foundation to CPs is supposed to enhance the construction of tools for ontology design, as for example envisaged in the NeOn project [3], and to facilitate the collaborative and distributed negotiation of meaning across the members of a same or of sufficiently close communities. For LKE, CPs appear slightly more complex than in other domains, mostly because the use cases in LKE involve a layering of meaning (from the physical to the social, cognitive, and finally legal realms), which also require an extended reification of entities such as norms, contracts, cases, legal text corpora, etc. A complex network of dependencies between roles and tasks, agents, normative positions, validity parameters, assumed goals, cognitive attitudes, etc. makes the modelling task in LKE harder than elsewhere. A few steps toward the creation of a repository of legal CPs have been sketched, together with some possible directions for further development.

References


Appendix: some content patterns from the Core Legal Ontology

Figure 4. An excerpt of the CLO taxonomy: legal description types. Some classes are from the Jur-WordNet ontology [47][48].
Figure 5. The minimal LegalCase CP from CLO. The Norm $\rightarrow$ Case CP is composed by the LegalNorm and LegalCase CPs.

Figure 6. The JudgmentProcedure minimal CP. It allows to model the relation between a judgment procedure and a legal case that it tackles.
Figure 7. The LegalNorm minimal CP. I allows to associate a legal norm with the community adopting it, the legal principle that is implemented in the norm, and the legal text that expresses that norm.

Figure 8. The LegalDocument CP. It allows to model legal documents as realizations of legal texts, which have an appropriate legal quality for their validity.
Figure 9. The LegalInstitution CP. It allows to associate a legally-recognized institution with the constitutive norm that introduces it, and the power-conferring rule that describes it specifically.

Figure 10. The Legislation CP. It specializes the collection CP, and allows to model the relation between a legal description (law, norm, etc.), and the legal corpus to which it belongs.
This page intentionally left blank
Information Extraction of Legal Texts
Ontology learning from Italian legal texts

Alessandro LENCI\textsuperscript{a}, Simonetta MONTEMAGNI\textsuperscript{b}, Vito PIRRELLI\textsuperscript{b} and Giulia VENTURI\textsuperscript{b}

\textsuperscript{a} Dipartimento di Linguistica – Università di Pisa (Italy)
\textsuperscript{b} Istituto di Linguistica Computazionale (CNR, Pisa, Italy)

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

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

Introduction

Ontology building is nowadays a very active research field, as witnessed by the fast growing literature on the topic and the increasing number of Knowledge Management applications based on automated routines for ontology navigation and update. This task, however, requires harvesting domain–specific knowledge on an unprecedented scale, by tapping and harmonizing knowledge sources of highly heterogeneous conception, format and coverage, ranging from foundational ontologies and structured databases to electronic text documents. As electronic texts still represent the most accessible and natural repositories of specialised information worldwide, there is a reasonable expectation that the increasingly growing demand for ontologically–interpreted knowledge can eventually be met by making automatically–interpreted text information more and more available.

Different methodologies have been proposed to automatically extract information from texts and provide a structured organisation of extracted knowledge in as diverse domains/sectors as bio–informatics, health–care, public administration and company document bases. The situation in the legal domain is in line with this general trend.

The work illustrated in this paper reports the results of a case study carried out in the legal domain to automatically induce ontological knowledge from texts with an ontology learning system, hereafter referred to as T2K (Text–to–Knowledge), jointly designed and developed by the Institute of Computational Linguistics (CNR) and the Department of Linguistics of the University of Pisa. The system offers a battery of tools for Natural Language Processing (NLP), statistical text analysis and machine language learning,
which are dynamically integrated to provide an accurate representation of the content of vast repositories of unstructured documents in technical domains [8]. Text interpretation ranges from the acquisition of lexical and terminological resources, to advanced syntax and ontological/conceptual mapping, that are annotated as XML metadata to offer growing interoperability with automated content management systems for personalised knowledge profiling. Prototype versions of T2K are currently running on public administration portals and have been used for indexing E–learning and E–commerce materials. Customised versions of the T2K tool have been used to extract terminological and ontological knowledge from Italian legislative text collections. In what follows, we report the ontology learning results of two case studies carried out respectively on Italian legislative texts belonging to the environmental domain [21,14] and on EC texts on consumer law; the second case study has been performed in the framework of the DALOS (DrAfting Legislation with Ontology–based Support) European project 1.

1. Background

The last few years have seen a growing body of research and practice in constructing legal ontologies and applying them to the law domain. A number of legal ontologies have been proposed in different research projects: yet, most of them focus on a upper level of concepts and were mostly hand–crafted by domain experts (for a survey of legal ontologies, see [20]). It goes without saying that applications with realistically large knowledge bases in the legal domain need more comprehensive ontologies incorporating up–to–date knowledge: ontology–learning from texts could be highly effective in this direction. To our knowledge, relatively few attempts have been made so far to automatically induce legal domain ontologies from texts: this is the case, for instance, of [13], [19] and [24].

According to [6], the task of ontology learning is primarily concerned with the definition of concepts and relations (both taxonomical and non-hierarchical) holding between them. Ontology learning also includes the acquisition of linguistic knowledge about the terms used to designate a given concept in running texts as well as about synonym relations among terms. The various steps of ontology learning can be arranged in a layer cake of increasingly complex subtasks, as illustrated in Figure 1 from [6].

![Figure 1. Ontology Learning Layer Cake from [6]](http://www.dalosproject.eu/)

1http://www.dalosproject.eu/
All ontology learning experiments carried out in the legal domain are mainly focused on concept extraction as a primary step of the ontology development process. There are different methods illustrated in the ontology-learning literature used to learn concepts and relations from legal texts.

One approach exploits the frequency of definitions in legal texts as a valuable source of information from which it is possible to learn domain-relevant concepts as well as (both taxonomic and non-taxonomic) relations. This is the method followed by [24] whose focus is on definitions in a large collection of German court decisions. The extraction and analysis of definitions are carried out through a rule-based approach operating on 6000 German court decisions within the field of environmental law and are shown to have a significant impact on the ontology learning process.

Another methodology typically followed in the ontology-learning literature focuses on term extraction as a fundamental prerequisite for the identification of concepts and relations in normative texts. The basic assumption is that domain-specific concepts are typically phrased as terminological units. Semantic relations holding among concepts are conveyed through syntactic relations linking domain terms. In [13], the extraction of terms, concepts and (lexico-)semantic relations is carried out on a corpus of French legal texts through natural language processing techniques combined with statistical measures. Starting from the results of the terminology extraction step, the final ontology is built organising and structuring the knowledge implicitly contained in legal documents [12]. A similar approach is followed by [19], in which NLP tools are used to identify legal concepts and their defining properties. The new domain ontology bootstrapped from texts is then integrated and merged with an externally defined upper foundational legal ontology, with the result of creating a new domain ontology combining low-level concepts with top-level ones. The work reported in [23] belongs to the same line of research, where a combination of machine learning and natural language techniques are resorted to to extract domain-relevant terms, ontological concepts, instances and relations from Spanish legal texts. In particular, [23] focuses on the customisation effort carried out to specialise the Text2Onto ontology learning system [7] to support the Spanish legal ontology learning task in the framework of the European SEKT Project².

The present paper contributes to this latter line of research: first, relevant domain terminology is extracted, then both taxonomical and non-taxonomical relations (e.g. synonymy) between terms are acquired. This is done by combining NLP technologies with statistical techniques. It is interesting to note that in all cases surveyed so far natural language processing techniques play a central role in the process of legal knowledge extraction and organisation. In what follows we will focus on this specific issue.

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

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

²http://ontoware.org/projects/text2onto/
chains, complex domain-specific terms and the ubiquitous surface variability of phraseological expressions require the operation of a battery of disambiguating constraints, i.e. a set of interface rules mapping the underlying conceptual organization of a domain onto the surface language. With no such constraints in place, text becomes a slippery ground of unstructured, strongly perspectivised and combinatorially ambiguous information bits.

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

To implement this scenario, a few NLP ingredients are required. Preliminary term extraction presupposes pos-tagged texts, where each word form is assigned the contextually appropriate part-of-speech and a set of morpho-syntactic features plus an indication of lemma. Whenever more information about the local syntactic context is to be exploited, it is advisable that basic syntactic structures are identified. As we shall see in more detail below, we use chunking technology to attain this level of basic syntactic structuring. NLP requirements become more demanding when identified terms need to be organised into larger conceptual structures and connected through long-distance relational information. For this purpose syntactic information must include identification of dependencies among lexical heads.

The approach to ontology learning adopted by T2K differentially exploits all these levels of linguistic annotation of texts in an incremental fashion. Term extraction operates on texts annotated with basic syntactic structures (so-called “chunks”, see below). Identification of conceptual structures, on the other hand, is carried out against a dependency-annotated text. In what follows, the general architecture of the Italian parsing system underlying T2K (henceforth referred to as AnIta, [3]) is briefly illustrated.
Il presente decreto stabilisce le norme per la prevenzione ed il contenimento dell’inquinamento da rumore [...] ‘this decree establishes the rules for prevention and control of noise pollution [...]’

Figure 2. A sample of chunked text

2.1. An outline of AnIta

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

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

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

Text chunking is carried out through a battery of finite state automata (CHUG–IT, [10]), which takes as input a morphologically analysed and lemmatised text and segments it into an unstructured (non-recursive) sequence of syntactically organized text units called “chunks” (e.g. nominal, verbal, prepositional chunks). Chunking requires a minimum of linguistic knowledge; its lexicon contains no other information than the entry’s lemma, part of speech and morpho–syntactic features. A chunk is a textual unit of adjacent word tokens sharing the property of being related through dependency relations (e.g. pre–modifier, auxiliary, determiner, etc.). A chunked sentence, however, does not give information about the nature and scope of inter–chunk dependencies which are identified during the phase of dependency analysis (see below). Morpho–syntactic disambiguation is performed simultaneously to the chunking process.

To be more concrete, the sentence fragment reported in Figure 2 is segmented into eight chunks, each including a sequence of adjacent word tokens mutually related through dependency links of some kind. For example, the first nominal chunk (N_C) covers three word tokens, il presente decreto ‘the present deliberation’: the noun head decreto ‘deliberation’, the adjectival premodifier presente ‘present’ and an introducing definite article. Although the representation is silent about the relationship between stabilire ‘establish’ and le norme ‘the rules’, this is not to entail that such a relationship cannot possibly hold: simply, the lexical knowledge available to this parsing component makes it impossible to state unambiguously how chunks relate to each other and the nature of this relationship. This is the task for further analysis steps.

Dependency parsing is aimed at identifying the full range of syntactic relations (e.g. subject, object, modifier, complement, etc.) within each sentence: syntactic relations are
Figure 3. A sample of dependency–parsed text

represented as dependency pairs between lexical heads. It is carried out by IDEAL [2],
a finite state compiler for dependency grammars. The IDEAL general grammar of Ital-
ian is formed by ca. 100 rules covering the major syntactic phenomena. The grammar
rules are regular expressions (implemented as finite state automata) defined over chunk
sequences, augmented with tests on chunk and lexical attributes. The rules are organized
into two major modules: structurally–based rules and lexically–based rules, the latter
accessing a syntactic lexicon of more than 26,000 subcategorization frames for nouns,
verbs and adjectives. A “confidence value” (PLAUS) is associated with identified depen-
dency relations, to determine a plausibility ranking among competing analyses. Figure 3
reports the dependency representation of the same sentence.

The output consists of binary relations between content words, typically a head and
a dependent. There may be features associated with both participants in the relation con-
veying other types of information such as the semantic type of a dependent (ROLE) or the
preposition introducing a certain relation (INTRO). The sentence fragment is described
by seven dependency relations including subject, object as well as other modification
relations: for instance, decreto has been identified as the subject of the verb
stabilire and
norme as its direct object.

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

3. T2K architecture

T2K is a hybrid ontology learning system combining linguistic technologies and statisti-
cal techniques. T2K does its job in two basic steps:
1. extraction of domain terminology, both single and multi-word terms, from a document base, including term variants;
2. organization and structuring of the set of acquired terms into proto-conceptual structures, namely
   (a) fragments of taxonomical chains and
   (b) clusters of semantically related terms.

Figure 4 illustrates the functional architecture of T2K:

The two basic steps are represented in the central part of Figure 4, showing the interleaving of NLP and statistical tools. Acquired results are structured in the ontology box on the right-hand-side of the diagram, whose stratified organization is reminiscent of the hierarchical cascade of knowledge layers in the “Ontology Learning Layer Cake” by [6] (see Figure 1), going from terminological information to proto-conceptual structures corresponding to taxonomical and non-hierarchical relationships among terms. Acquired knowledge is also used for document indexing, on the basis of extracted terms and acquired conceptual structures. In what follows we focus on the ontology learning process.

3.1. Term extraction

Term extraction is the first and most-established step in ontology learning from texts. Terms are surface realisations of domain-specific concepts and represent, for this reason, a basic prerequisite for more advanced ontology learning tasks. In principle, they need to be recognized whatever the surface form they show in context, irrespectively of morpho-syntactic and syntactic variants. For our present purposes, a term can be a common noun as well as a complex nominal structure with modifiers (typically, adjectival and prepositional modifiers). Term extraction thus requires some level of linguistic pre-processing of texts.
T2K looks for terms in syntactically chunked texts such as those illustrated in Section 2.1 (Figure 2). Candidate terms may be one word terms (“single terms”) or multi-word terms (“complex terms”). The acquisition strategy differs in the two cases.

Single terms are identified on the basis of frequency counts in the chunked source texts, after discounting stop-words. The acquisition of multi-word terms, on the other hand, follows a two-stage strategy. First, the chunked text is searched for on the basis of a set of chunk patterns. Chunk patterns encode syntactic templates of candidate complex terms and cover the main types of modification observed in complex nominal terms: i.e. adjectival modification (e.g. organizzazione internazionale ‘international organisation’), prepositional modification (e.g. tutela del territorio ‘protection of the territory’), including more complex cases where different modification types are compounded (e.g. incenerimento dei rifiuti pericolosi ‘incineration of dangerous waste’). The set of chunk patterns used to identify candidate terms has been tailored to meet the specific needs of the legal domain, characterised by the frequent use of deep PP-attachment chains including a high number of embedded prepositional chunks (see [22]). Secondly, the list of acquired potential complex terms is ranked according to their log-likelihood ratio [9], an association measure that quantifies how likely the constituents of a complex term are to occur together in a corpus if they were (in)dependently distributed, where the (in)dependence hypothesis is estimated with the binomial distribution of their joint and disjoint frequencies. We tested the log-likelihood ratio against other association measures such as mutual information, chi-square etc., log-likelihood scoring consistently better than the others. Moreover this measure is known to be less prone to assigning high scores to very sparse pairs. It should be recalled that the log-likelihood ratio is commonly used for discovering collocations. Hence, we are treating complex terms as though they belonged to the more general class of collocations. However, T2K uses the log-likelihood ratio in a somewhat atypical way: instead of measuring the association strength between adjacent words, T2K measures it between the lexico-semantic heads of adjacent chunks. The main and often underestimated advantage of defining co-occurrence patterns over syntactic structures is that we can broaden our search space (the text window) in a controlled way, by making it sure that there is a syntactic pattern linking two adjacent lexical heads.

So far, acquisition of potential complex terms has involved chunk pairs only (bigrams). In T2K recognition of longer terms is carried out by iterating the extraction process on the results of the previous acquisition step. This means that acquired complex terms are projected back onto the original text and the acquisition procedure is iterated on the newly annotated text. The method proves helpful in reducing the number of false positives consisting of more than two chunks [4]. Interestingly, the chunk patterns used for recognition of multi-word terms need not necessarily be the same across different iteration stages. In fact, it is advisable to introduce potentially noisy patterns only at later stages. This is the case, for instance, of coordination patterns.

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

For each acquired term, either single or complex, the variants attested in the document base are also extracted. Term variation represents a crucial issue which is usually under-discussed and rarely accounted for in terminological processing (see [16]). Quite the contrary, it deserves specific attention, from both the theoretical and applicative points of view. From the theoretical point of view, the range of detected term variants can help shedding light on the nature of the term, e.g. whether it is a frozen multi-word term or it presents itself as a flexible linguistic construction (for instance, if it allows for the insertion of modifiers). On the applicative front, term variants can be usefully exploited to improve indexing and retrieval results. T2K treats as term variants all instantiations of the same (either single or complex) term showing the same abstract representation at the chunking level, and covering at least 5\% of the occurrences of the term in the acquisition corpus. Term variants acquired in this way cover a wide spectrum of linguistic phenomena, ranging from orthographic to morphological and syntactic variation. A typology of acquired term variants follows:

1. orthographic variants, e.g. *tasso d’interesse / tasso di interesse* ‘interest rate’;
2. inflectional variants:
   (a) singular vs plural form, e.g. *convertitore catalitico / convertitori catalitici* ‘catalytic converter(s)’;
   (b) simple vs articulated prepositions, e.g. *bonifica di aree inquinate / bonifica delle aree inquinate* ‘drainage of polluted zones / drainage of the polluted zones’;
3. structural variants, e.g. *fornitore per il servizio finanziario / fornitore di servizi finanziari* ‘supplier for the financial service/of the financial service’;
4. variants including modifiers, e.g. *adempimento del detto obbligo / adempimento dell’obbligo* ‘fulfilment of the said obligation / fulfilment of the obligation’;
5. variants combining different types of variation, e.g. *contratto di fornitura / contratti di fornitura / contratti per la fornitura* ‘supply contract / supply contracts / contracts for the supply’.

3.2. Term organization and structuring

In the second extraction step, proto-conceptual structures involving acquired terms are identified. The basic source of information is no longer a chunked text, but rather the dependency analysis exemplified in Figure 3, with the original text containing an explicit indication of the multi-word terminology acquired at the previous extraction stage.

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

Moreover, T2K clusters semantically-related terms by using CLASS, a distributionally-based algorithm for building lexico-semantic classes [1]. According to CLASS, two terms are semantically related if they can be used interchangeably in a statistically sig-
significant number of syntactic contexts. The starting point for the CLASS algorithm is provided by a dataset of dependency triples – \(<T, C, s>\) –, where \(T\) is a target linguistic expression, \(C\) is a linguistic context for \(T\), and \(s\) is the particular syntagmatic dependency relation between \(T\) and \(C\). For our present concerns, variables are interpreted as follows:

1. \(T\) corresponds to an acquired term in the TermBank;
2. \(s\) stands for either a subject or a direct object dependency relation;
3. \(C\) corresponds to a verb with which \(T\) is attested to co–occur as a subject or a direct object.

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

For all terms (both single and complex) in the TermBank, we extract from the dependency–annotated text the best verb/subject and verb/object pairs. CLASS then computes the degree of semantic relatedness between two terms \(T_1\) and \(T_2\) by measuring the degree of overlapping between \(BVT_1\) and \(BVT_2\), according to the metric described in [1]. This corresponds to the assumption that the semantic similarity between two terms is a function of the possibility for the entities denoted by the terms to be involved in similar events, where the latter are expressed by the term best verbs. The cluster of terms semantically related to a target term \(T\) is finally ordered by decreasing similarity scores with respect to \(T\). For each term, the user can define the maximum number of related terms to be returned by the system; this parameter can be set on the basis of the user’s needs (it should be kept in mind that going down in the ranked list of related terms the semantic distance from \(T\) increases; therefore, it becomes more likely to find spurious associations).

4. Ontology learning from legislative texts

In this section we summarise the results of two case studies carried out on two different corpora of Italian legal texts.

4.1. The corpora

Two different sets of experiments were carried out with T2K on two legal corpora, differing at the level of the regulated domains (i.e. the environmental and the consumer protection domains) and of the releasing agencies (i.e. European Union, Italian national state and Piedmont local authority). These two corpora will be henceforth referred to as Environmental Corpus and Consumer Law Corpus respectively.

The Environmental Corpus consists of 824 legislative, institutional and administrative acts concerning the environmental domain, for a total of 1,399,617 word tokens, coming from the BGA (Bollettino Giuridico Ambientale) database edited by the Pied-
Table 1. An excerpt of the automatically acquired TermBank

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

The Consumer Law Corpus was built in the framework of the DALOS project by legal experts within European normative sources on consumer law and includes Directives, Regulations and case law: the resulting corpus is composed of the Italian version of 16 Directives and 42 Case Law texts, for a total of 292,609 word tokens. Unlike the Environmental Corpus, it represents a homogeneous document collection, including Italian European law texts only.

The corpora were converted into a suitable format and then preprocessed with a view to a) extending the general purpose morphological lexicon with missing lexical items, and b) tailoring the multi-word term recognition grammar to the main linguistic peculiarities of legal and administrative acts identified in a contrastive corpus study (whose results are reported in [22]).

4.2. The acquired TermBank

Table 1 contains a fragment of the TermBank automatically acquired from the Environmental Corpus. For each selected term, the TermBank reports its prototypical form (in the column headed “Term”, and corresponding to the term form most frequently attested in the corpus), its frequency of occurrence in the whole document collection, and the lemma of the lexical head(s) of the chunk(s) covering the term (see column “Lemmatised headwords”). Finally, the column headed “Stop” is used to mark potentially noisy terms: whereas the NULL value applies to those terms for which a variety of different surface realisations have been recorded, in the case of potentially noisy terms the “Stop” column records the preposition always introducing the candidate term in the acquisition corpus (see below).
Table 2. An excerpt of the automatically acquired TermBank with potentially noisy terms

<table>
<thead>
<tr>
<th>ID</th>
<th>Term</th>
<th>Freq</th>
<th>Lemmatised headwords</th>
<th>Stop</th>
</tr>
</thead>
<tbody>
<tr>
<td>12</td>
<td>sensi</td>
<td>691</td>
<td>senso</td>
<td>ai</td>
</tr>
<tr>
<td>42</td>
<td>sensi dell’art.</td>
<td>332</td>
<td>senso art.</td>
<td>ai</td>
</tr>
<tr>
<td>380</td>
<td>sensi dell’articolo</td>
<td>35</td>
<td>senso articolo</td>
<td>ai</td>
</tr>
<tr>
<td>377</td>
<td>sensi della direttiva</td>
<td>37</td>
<td>senso direttiva</td>
<td>ai</td>
</tr>
</tbody>
</table>

The choice of representing a domain term through its prototypical form rather than the lemma (as typically done in ordinary dictionaries) follows from the assumption that a bootstrapped glossary should reflect the actual usage of terms in texts. In fact, domain-specific meanings are often associated with a particular morphological form of a given term (e.g. the plural form). This is well exemplified in Table 1 where the acquired terms headed by *acqua* ‘water’ can be parted into two groups according to their prototypical form: either singular (e.g. *acqua potabile* ‘drinkable water’) or plural (e.g. *acque superficiali* ‘surface runoff’). It should be noted, however, that reported frequencies are not limited to the prototypical form, but refer to all occurrences of the abstract term.

Due to the specific stylistic conventions of the legal language characterised by the massive use of formulaic expressions, the set of acquired terms also includes wrongly interpreted domain relevant terms such as *sensi della legge* lit. ‘sense of the law’ or *fini della protezione del consumatore* lit. ‘end of consumer protection’ and the like, corresponding to the multi-word prepositions *ai sensi di* ‘under (the law, clause, etc.)’ and *ai fini di* ‘in order to’. In order to avoid these multi-word prepositions to be wrongly acquired as candidate domain relevant terms, T2K keeps track of the preposition introducing the term in all its textual occurrences: when the term is always introduced by the same preposition, it is very likely that the candidate term is a multi-word preposition or adverb. Since with low frequency thresholds this may also occur with real terms, it was decided to keep the candidate noisy term in the resulting TermBank and to mark it as potentially noisy by recording the preposition introducing it in the “Stop” column. This is well exemplified in Table 2, where a TermBank excerpt with potentially noisy terms is reported. On the basis of this information type, potentially noisy terms can be removed from the TermBank, either automatically or – more appropriately – through manual inspection and validation.

As already pointed out in Section 3.1, acquired information about terms also includes significantly attested term variants. In fact, the selection of the prototypical form of a term does not represent an uncontroversial choice; it is often the case that the same term appears in different terminological resources under different prototypical forms, differing, for instance, in number (e.g. *accordo di programma* vs *accordi di programma* ‘programmatic agreement’) or at the level of the linking prepositions (e.g. *acquisizione dati* vs *acquisizione di dati* ‘acquisition of data’, *abbandono di rifiuti* vs *abbandono dei rifiuti* ‘waste abandon’, *contratto di fornitura* / *contratti di fornitura* / *contratti per la fornitura* ‘supply contract / supply contracts / contracts for the supply’). These simple examples show that term variation represents an important issue which needs to be adequately accounted for. According to [11], on average approximately one third of the occurrences of a term appear to be variants; moreover, they show up not only in text, but also in controlled, manually built terminological resources. From this it follows that automatic term recognition cannot be restricted to the identification of the most likely
T2K deals with two different types of term variation:

1. lexical variation, i.e. genuine lexical synonyms, acronyms and abbreviations;
2. orthographic, morphological and structural variation.

Variants of the first type are acquired in the term structuring step (see Section 3.2); in particular, acquired clusters of semantically related terms can include lexical synonyms, acronyms as well as abbreviations, a very frequent form of term variant attested in technical sublanguages, which in some cases represents the unmarked usage of terms (e.g. \textit{PBC} for \textit{polichlorobifenile} ‘polychlorinated biphenyl’ or \textit{l.r.} for \textit{legge regionale} ‘regional law’). Variants of the second type are extracted during the term extraction phase (see Section 3.1), based on their sharing the same abstract syntactic representation. Table 3 exemplifies acquired information for each identified variant. Each raw represents a term variant: the column “\text{var\_id}” contains the variant identifier; the column “\text{term}” specifies the term the variant being described refers to; the column “\text{variant\_form}” records the term variant, and the “\text{freq}” column its frequency of occurrence in the whole document collection.

\subsection{Evaluation of the acquired TermBanks}

As expected from the specific nature of processed documents, in both case studies the acquired TermBanks include both legal and regulated domain terms, environmental and consumer protection terms respectively. Since the two classes of terms show quite dif-
ferent frequency distributions in the acquisition corpora, different extraction experiments were carried out by setting different thresholds (see Section 3.1). With the Environmental Corpus, a standard threshold configuration yielded a TermBank of 4,685 terms (both single and multi-word terms): minimum frequency of both single and multi-word terms was set to 7, while the percentage of selected terms from the ranked lists was 10% for single terms and 70% for multi-word terms. Environmental terms, however, turned out to scarcely be represented, due to their low frequency. Since the focus of our interest was on both types of terminology, we carried out two further acquisition experiments, by setting the minimum frequency threshold to 5 and 3 respectively. The expected increase in the number of acquired environmental terms (a minimum frequency threshold set to 3 raises the number of extracted candidate terms to 11,103) was attained at the cost of getting a larger number of spurious candidate terms. With the Consumer Law Corpus, the best configuration resulted to be the following one: the selected minimum frequency threshold for both single and multi-word terms was 5, the percentage of selected terms from the ranked lists was 20% in the case of single terms and 70% for multi-word terms. With this configuration a TermBank of 1,443 terms (both single and multi-word terms) was obtained.

4.3.1. Evaluation criteria

Evaluation of acquired TermBanks was carried out by comparing T2K results against selected gold standard resources.

Different types of matches were taken into account. Besides the full match between the T2K term and the term in the reference resource, different types of partial matches were also considered, i.e.:

1. the same term appears both in the T2K TermBank and in the gold standard resource but under different prototypical forms: this is the case, for instance, of the term accordi di programma ‘programmatic agreement’ which appears in the plural form in T2K and in the singular form in the legal reference resources taken into account. At this level, two terms may also differ for the prepositions linking the nominal headwords of a complex term, as in the case of acquisizione dati vs acquisizione di dati ‘acquisition of data’ or abbandono di rifiuti vs abbandono dei rifiuti ‘waste abandon’;

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

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

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

Evaluation of extracted terms was carried out against the selected gold standard resources by taking into account both full and partial matches. In particular, the reliability of the bootstrapped Term-Banks was measured in terms of type precision, which was calculated as the percentage of correctly acquired term types with respect to all acquired
term types. On the other hand, an evaluation in terms of type recall (calculated as the percentage of correctly acquired term types with respect to all term types in the gold standard lexicon) was not easily applicable since this measure does not permit to discriminate between effectiveness and reliability of the term extraction system and the coverage of the reference resource with respect to the acquisition corpus. In the specific context of ‘legal–environmental’ TermBank, the selected reference resources could not be used for this specific purpose due to their wider coverage, not circumscribed to the environmental domain. A partial recall evaluation could instead to be carried out for what concerns the consumer law TermBank, with respect to a subset of domain relevant concepts 4.3.3.

4.3.2. Evaluation of the legal–environmental TermBank

Evaluation was based on the TermBank of 4,685 terms (with minimum frequency threshold equal to 7). Due to the highly heterogeneous nature of the terms in the glossary, belonging to both the legal–administrative and the environmental domains, two different resources were taken as a gold standard: the Dizionario giuridico (Edizioni Simone) available online4 was used as a reference resource for what concerns the legal domain (henceforth referred to as Legal_RR), and the Glossary of the Osservatorio Nazionale sui Rifiuti (Ministero dell’Ambiente) available online5 for the environmental domain (henceforth referred to as Env_RR), which contain respectively 6,041 and 1,090 terminological entries recorded in their prototypical form.

The results of the evaluation carried out on the basis of the criteria described in Section 4.3.1 can be summarised as follows: in 51% of the cases a match, either full or partial, was found between the T2K glossary and the reference resources. In particular, 89% of identified matches were found to be legal terms and 34.5% environmental ones, meaning that 23.5% terms are present in both reference resources. A natural question to be answered at this point is whether the remaining 49% of terms for which no match was found was represented by errors and noisy terms or were domain–specific terms not appearing in the selected reference resources. To address this issue, we selected two additional resources available on the Web: the list of keywords used for the online query of the Archivio DoGi (Dottrina Giuridica)6 for the legal domain, and the thesaurus EARTH (Environmental Applications Reference Thesaurus)7 for the environmental domain, against which a manual evaluation was carried out for 25% of the automatically acquired T2K glossary. Results are encouraging: by including these two richer reference resources, the percentage of matching terms increases to 75.4%. This percentage grows up to 83.7% if we also include terms which, however absent in the selected reference resources, were manually evaluated as domain–relevant terms: this is the case, for instance, of the terms anidride carbonica ‘carbon dioxide’ for what concerns the environmental domain or beneficiari ‘beneficiary’ for the legal one. The percentage of manually detected errors is 21.1%, which includes some of the terms for which a partial match was detected.

4http://www.simone.it/cgi-local/Dizionari/newdiz.cgi?index.5.A
5http://www.osservatorionazionalerifiuti.it/ShowGlossario.asp?L=Z
6http://nir.ittig.cnr.it/dogiswish/dogiConsultazioneClassificazioneKWOC.php
7http://uta.iaa.cnr.it/earth.htm#EARTH%202002
Table 4. An example of acquired taxonomical chain

<table>
<thead>
<tr>
<th>Term</th>
</tr>
</thead>
<tbody>
<tr>
<td>applicazione</td>
</tr>
<tr>
<td>applicazione dei paragrafi</td>
</tr>
<tr>
<td>applicazione dell’ articolo</td>
</tr>
<tr>
<td>applicazione della direttiva</td>
</tr>
<tr>
<td>applicazione della legge</td>
</tr>
<tr>
<td>applicazione della tariffa</td>
</tr>
<tr>
<td>applicazione delle disposizioni</td>
</tr>
<tr>
<td>applicazione delle sanzioni</td>
</tr>
<tr>
<td>applicazione delle sanzioni amministrative</td>
</tr>
<tr>
<td>applicazione delle sanzioni previste</td>
</tr>
<tr>
<td>applicazione del presente decreto</td>
</tr>
<tr>
<td>applicazione del regolamento</td>
</tr>
<tr>
<td>applicazioni di quarantena</td>
</tr>
</tbody>
</table>

4.3.3. Evaluation of the consumer law TermBank

In spite of the fact that the evaluation results illustrated above are to be considered quite satisfactory and in line with state of the art term acquisition systems, more promising results were obtained with the Consumer Law Corpus [15]. The consumer law TermBank automatically acquired by T2K was evaluated both against the Archivio DoGi (Dottorina Giuridica) and against JurWordNet [18], containing respectively 9,127 keywords recorded in their prototypical form and 5,353 lemmata. In particular, by considering both full and partial matches, the observed precision of the bootstrapped TermBank improved to 85.38%, while the cases of non-matching terms decreased from 21.1% to 14.62%. Moreover, the percentage of manually detected errors dropped to 6.1%.

A partial recall evaluation was carried out in this context with respect to a subset of 56 of the European Union Legal Concepts (EULG concepts) from LOIS (see [17] and [15] for the complete list) with encouraging results: 80.69% of the terms in the subset were acquired by T2K, of which 52.17% was represented by full matching terms and 47.83% by partially matching ones.

4.4. Proto-conceptual organisation of terms

A first step towards the conceptual organization of terms in the TermBank consists in building taxonomical chains. This is to say that single and multi-word terms are structured in vertical relationships providing fragments of taxonomical chains such as the one reported in Table 4, where the acquired direct and indirect hyponyms of the term *applicazione* ‘enforcement’ are reported. In this example, it can be noticed that terms sharing the head only are the direct hyponyms of the root term. Further hyponymy levels can be detected when two or more multi-word terms share not only the head but also modifiers, as in the case of the *applicazione delle sanzioni amministrative* ‘enforcement of administrative sanctions’ with respect to the more general term *applicazione delle sanzioni* ‘enforcement of sanctions’.

In the Environmental Corpus, with minimum frequency threshold set to 7 the number of extracted hyponymic relations is 2,181 referring to 272 hyperonym terms; with the threshold set to 3, identified hyponymic relations increase to 6,635 regarding 454 hyperonym terms. Concerning the Consumer Law Corpus, the number of extracted hyponymic
Table 5. Examples of acquired clusters of semantically related terms

<table>
<thead>
<tr>
<th>Term</th>
<th>Related Terms</th>
</tr>
</thead>
<tbody>
<tr>
<td>disposizioni 'provision'</td>
<td>norme, disposizioni relative, decisione, atto, prescrizioni</td>
</tr>
<tr>
<td>legge 'law'</td>
<td>regolamento, protocollo, accordo, statuto, amministrazioni comunali</td>
</tr>
<tr>
<td>inquinamento 'pollution'</td>
<td>danno ambientale, inquinamento marino, effetti nocivi, conseguenza, inquinamento atmosferico</td>
</tr>
<tr>
<td>impatto ambientale 'environmental impact'</td>
<td>esposizione, danno, esigenze, conseguenza, pericolo</td>
</tr>
</tbody>
</table>

relations is 623 referring to 229 hypernym terms; due to the smaller size of the corpus, these results were achieved with with the minimum frequency threshold set to 5.

The second structuring step performed by T2K consists in the identification of clusters of semantically related terms which is carried out on the basis of distributionally–based similarity measures (see Section 3.2). In Table 5, clusters of semantically related terms are exemplified. For each target term, the set of the first 5 most similar terms is returned, ranked for decreasing values of semantic similarity. In the Environmental Corpus, the number of identified related terms is 3,448 referring to 665 terminological headwords (this result was obtained with the minimum frequency threshold set to 7). With the minimum frequency threshold set to 5, the number of related terms identified from the Consumer Law Corpus is 1,258 referring to 279 terminological headwords.

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

5. Conclusions and further directions of research

We reported the results of applying an automatic ontology learning system, T2K, on two different corpora of Italian legislative texts belonging to the environmental and consumer protection domains.
The elements of novelty of the T2K approach are at least twofold. First of all, the incremental interleaving of robust NLP and machine–learning technologies allows us to successfully tackle what we termed above as the "acquisition paradox". Secondly, T2K is able to acquire both fairly abstract pieces of conceptual knowledge in the form of semantic associations between terms, and low-level term properties, such as the variants in which a term can appear in texts. Although these types of information lie at the opposite extremes of a continuum spanning from abstract knowledge to its actual realization in texts, it is essential that they can be simultaneously targeted by a term acquisition system.

T2K is also easily customizable, as shown by its successful adaptation to the legal domain. Actually, our work shows that linguistic peculiarities of legal texts need to be investigated and taken into account for any higher–level content analysis. By bootstrapping base domain–specific knowledge from texts through knowledge–poor language tools we can incrementally develop more and more sophisticated levels of content representation. In the end, the purported dividing line between language–knowledge and domain–specific knowledge proves to be untenable in language use, where language structures and bits of world–knowledge are inextricably intertwined.

There is an enormous potential for this bootstrapping technology. Acquired TermBanks can be transformed into semantic networks linking identified legal and environmental entities. Current lines of research in this direction include a) semi–automatic induction and labeling of ontological classes from the proto–conceptual structures identified by T2K, and b) the extension of the acquired ontology with concept–linking relations (first steps in this direction are reported in [21]).

Our experiments also highlighted some interesting open issues which need to be tackled in the near future. As pointed out in Section 4.3, running T2K on a corpus of legislative and administrative acts results in a two–faced terminological glossary, which includes terms belonging to both the legal–administrative and regulated domains. Establishing the domain relevance of each acquired term represents a central issue when dealing with legal–administrative texts. Some preliminary experiments have already been carried out in order to semi–automatically identify the domain–relevance of each acquired term. In particular, terminology acquisition was carried out with T2K on thematically different legislative corpora. By comparing the TermBanks automatically extracted from different corpora, we could classify the terms belonging to their intersection as belonging to the legal–administrative lexicon. This is in line with the contrastive approach to term extraction proposed by [5]. Similarly, the relevance of regulated domain (e.g. environmental) terms will be validated by running terminology extraction on domain–specific literature.

References


Definitions in Court Decisions – Automatic Extraction and Ontology Acquisition

Stephan WALTER and Manfred PINKAL
Universität des Saarlandes
66123 Saarbrücken, Germany

Abstract. This paper deals with the use of computational linguistic analysis techniques for information access and ontology learning within the legal domain. We present a rule-based approach for extracting and analysing definitions from parsed text and evaluate it on a corpus of about 6000 German court decisions. The results are applied to improve the quality of a text based ontology learning method on this corpus.

Keywords. Definition Extraction, Computational Linguistics, Legal Information Systems, Ontology Learning

1. Motivation

Methods like ontology based knowledge management and information access through conceptual search have become active research topics, with practical applications in many areas. However the use of IT in legal practice (at least in German speaking countries) is up to now mainly restricted to document preparation and management or Boolean keyword search on full-text collections. Legal ontologies have been proposed in various research projects, but they focus on an upper level of concepts and are, with only a few exceptions, small knowledge repositories that were hand-made by experts (for a summary of existing legal ontologies, cf. [1]).

It is clear that realistically large knowledge-based applications in the legal domain will need more comprehensive ontologies incorporating e.g. up-to-date knowledge from court decisions. For this purpose an expert-based approach has to be supplemented by automatic acquisition methods. The same is true for large-scale advanced information access: Extensive conceptual indexation of even a fraction of all court decisions published in one year seems hardly possible without automatic support. However there has been relatively little research on the use of natural language processing for this purpose (exceptions are [2] and [3]).

In this paper we look at the use of computational linguistic analysis techniques for information access and ontology learning within the legal domain. We present a rule-based method for extracting and analyzing definitions from parsed text, and evaluate this method on a corpus of about 6000 German court decisions within the field of environmental law. We then report on an experiment exploring the use of

1 This paper describes research within the project CORTE funded by the German Science Foundation, DFG PI 154/10-1 (http://www.coli.uni-saarland.de/projects/corte/)
our extraction results to improve the quality of text-based ontology learning from noun-adjective bigrams. We will start however with a general discussion of the role that definitions play in legal language.

2. Definitions in Legal Language

Two central kinds of knowledge contained in the statutes of a code law system are normative knowledge, connecting legal consequences to descriptions of certain facts and situations, and terminological knowledge, consisting in definitions of some of the concepts used in these descriptions ([4]). Normative content is exemplified by (1), parts of section 29 of the German Federal Water Act. The legal consequence consisting in the responsibility for the maintenance of waters is connected to the precondition of ownership of the respective waters:

\[(1.1) \text{Die Unterhaltung von Gewässern} \ldots \text{obliegt den Eigentümern der} \begin{array}{l}
\text{Gewässer} \\
\text{\ldots the responsibility for maintaining waters shall lie with the owners of waters} \\
\end{array} \]

Article 1 of the German Federal Water Act e.g. captures a specific terminological sense of waters as follows:

\[(1.2) \text{Dieses Gesetz gilt für folgende Gewässer:} \\
\begin{array}{l}
1. \text{das ständig oder zeitweilig in Betten fließende oder stehende (..) Wasser} \\
\text{\ldots permanently or temporarily flowing or standing (..) waters confined within a bed} \\
\end{array} \]

If the definitions contained in statutes would fully specify how the relevant concepts are to be applied, cases could be solved (once the relevant statutes have been identified) by mechanically checking which of some given concepts apply, and then deriving the appropriate legal consequences in a logical conclusion. However, discussion in courts (and consequently texts that document court decisions) is largely devoted to pinning down whether certain concepts are to be applied or not. Controversies often arise because not all relevant concepts are defined at all within statutes, and because the terms used in legal definitions are often in need of clarification themselves. Many evaluative concepts (such as e.g. significant value) cannot be captured by general descriptive definitions at all. However even relatively concrete descriptive concepts, such as waters in (1), often need to be supported with further definitions in courts’ decisions.

The definitions in (2) are quoted from a decision by the Higher Administrative Court of Hamburg. Sentence (2.1) fixes what generally counts as the bed of a body of water (a concept that is used in (1) but not defined in the Federal Water Act), and (2.2) states precisely how this definition is to be applied regarding the specific case of tubed ditches.
(2.1) Unter einem Gewässerbett ist eine äußerlich erkennbare natürliche oder
künstliche Begrenzung des Wassers in einer Eintiefung an der Erdoberfläche zu
verstehen (vgl. BVerwG, Urt. v. 31.10.1975, BVerwGE Bd. 49 S. 293, 298;
Beschl. v. 17.2.1969, Buchholz 445.4 § 1 WHG Nr. 3, m.w.N.).

[By a bed of a body of water is to be understood: the natural (…) confines of wa-
ter within a cavity in the surface of the earth. (cf. BVerwG, Urt. v. 31.10.1975,
BVerwGE Bd. 49 S. 293, 298; Beschl. v. 17.2.1969, Buchholz 445.4 § 1 WHG Nr.
3, m.w.N.).]

(2.2) Von einem derartigen Bett kann u.a. dann nicht mehr gesprochen werden,
Wenn ein Graben vollständig verrohrt wird.

Such a bed of a body of water (…) can no longer be assumed if a ditch is fully

tubed.

On the one hand - as can be seen from the extensive amount of citation sources
mentioned in brackets in (2.1) – such definitions are frequently re-used and remain
binding beyond the case at hand. On the other hand they are generally open for
later amendment and modification. The semantics of legal concepts is thus subject
to constant adaptation and revision within use.

At the heart of this phenomenon lies the fact that statutes are written in natural lan-
guage, not in a formalized or a strongly restricted specialized language. It is widely
assumed in the philosophical literature that most natural language concepts do not
lend themselves to definitions fixing all potential conditions of applicability a pri-
ori. Additionally, reality is complex and constantly changing, and these changes
also pertain to the applicability of formerly clear-cut concepts. While this is espe-
cially true of social reality, rather physical concepts may be affected as well. An
often cited legal example is a case where the German Reichsgericht had to decide
whether electricity was to be counted as a thing.

From the point of view of legal theory this open-textured character of natural lan-
guage concepts is often seen as essential for the functioning of any legal system
(the term open texture was introduced into this discussion by [5]).

The application of legal terminological knowledge thus involves a continuous re-
adjustment of the balance between precision and openness, which is brought about
by the use of explicit defining statements. Easy access to definitions in decisions is
therefore of great importance to the legal practitioner. Section 3 discusses a de-
tailed linguistic analysis of various definition types and Section 4 shows how com-
putational linguistic analysis helps answering this need by enabling an accurate
search for definitions in a large collection of court decisions.

Definitions are however not only of direct value as a source of information in legal
practice. They also provide contexts that contain particularly much relevant termi-
nology, and are therefore a good place to search for concepts to be integrated in a
domain ontology. Given the importance and frequency of definitions in legal text,
such an approach seems particularly promising for this domain. Section 5 de-
scribes how automatically extracted definitions improve the results of a standard
ontology learning method.
3. Linguistic Analysis of Definitions

Our current work is based on a collection of more than 6000 verdicts in environmental law. As a starting point however we conducted a survey based on a random selection of 40 verdicts from various legal fields (comprising 127 349 tokens in 3757 sentences, “pilot study corpus”), which contained 126 definitions. Inspection of these definitions has on the one hand shown a range of different functional types and formulation patterns, which on the other hand share various common structural elements.

3.1. Types and Structures

The prototypical structure dealt with in most literature on definitions looks as follows (so called ‘Aristotelian’ definition scheme):

An A is a B which C

This construction also occurs in court decisions, but there it is by no means the only way used to make defining statements. Firstly, the relation between the defined term and the defining phrase can be established by means of appositive constructions instead of clause predicates. Through appositive constructions, either the defined term or the defining phrase is marked as background information. Such constructions therefore often serve to remind the reader of definitions that are presupposed or have been given explicitly somewhere else in the text.

Moreover, a considerable number of different predicates can be used in clause-based definitions apart from copular be. One important reason for this high degree of variability in formulation lies in the specific role of definitions in court decisions. Scientific and technical terminology is often built up using more or less context-free general definitions assigning new terms to places within a given taxonomy. In contrast, defining statements in verdicts are parts of coherent texts and do not only serve as specifications of terms, but also as arguments for or against their application in a specific case. Example (2) (in Section 2) exhibits a typical macro-structure of such ‘definitional arguments’: A general core definition (2.1) is elaborated by an additional, more specific statement (2.2) serving an argumentative goal with respect to the given case. Such elaborations may be given dialectically, for and against the use of the concept (here: against the classification as bed of water). Additionally, background information – not conclusively supporting, but ‘pointing to’ the (non-) applicability of the respective concept – may be attached to each of the elaborating statements in such a sequence.

Out of the 126 definitions found in our pilot study, 36 were realized by appositive means, while 90 include at least one predicate-based core-statement, with an additional 18 elaboration and 21 background statements (1.5 sentences per definition on average). These definitions use 52 different predicates that mostly fall into the following four classes:

1. expressing a classification: e.g. copular sein-be, classificatory verbs such as fallen unter-fall under or (less neutrally) gelten als – be considered as
2. meta-linguistic, used to speak directly either about word meaning or conditions of applicability (e.g. *bedeuten* – mean or *vorliegen* – ‘be existent’)

3. referring to aspects important to the process of legal interpretation (e.g. *fordern* – require, *darstellen* – constitute)

4. naming a specific type of feature used in the respective definition (e.g. *dienen zu* – serve as, *schützen* - protect)

The diagram in Figure 1 shows that the relatively general classificatory and meta-linguistic predicate types are most frequent in core statements. Within the two types of additional statements, legal interpretational questions are at issue. This is reflected by the relative prominence of the class of interpretation related predicates. Finally, background statements tend to give quite specific information on the defined concept, and therefore contain feature specific predicates more often than the other two statement types.

3.2. Structural Elements

Inspection of the definitions in our pilot study corpus has further shown a range of common structural elements, and has allowed us to identify typical linguistic realizations of these. We will illustrate this with the example definition given in (3):

(3) [4 Bei einem Einfamilienreihenhaus] [3 liegt] ein [1 mangelhafter Schallschutz] [5 dann] [3 vor, wenn] [2 die Haustrennwand einschalig errichtet wurde] (#).

(One-family row-houses have insufficient noise insulation if the separating wall is one-layered.)

This definition contains:

1. The definiendum, i.e. the element that is defined (*unzureichender Schallschutz* - insufficient noise insulation).
2. The definiens, i.e. the element that fixes the meaning to be given to the definiendum (die Haustrennwand einschalig errichtet wurde - the separating wall is one-layered).

Apart from these constitutive parts, it contains:
3. A connector, indicating the relation between definiendum and definiens (liegt...vor, wenn, have.... if).
4. A qualification specifying a domain area of applicability, i.e. a restriction in terms of the part of reality that the regulation refers to (bei Einfamilienreihenhäusern - one-family row-houses).
5. Signal words that cannot be assigned any clear function with regard to the content of the sentence, but serve to mark it as a definition (dann - α).

The connector normally contains at least the predicate of the main clause, often together with further material (subjunction, relative pronoun, determiner). It not only indicates the presence of a definition. It also determines how definiens and definiendum are realized linguistically and often contains information about the type of the given definition (core, elaboration or background; full, partial, by examples...). The linguistic realization of definiendum and definiens depends on the connector. One common pattern realizes the definiendum as the subject, and the definiens within a subclause. The domain area is often specified by a prepositional phrase (PP) introduced by bei (“in the field of”, for), as seen in the example. Further possibilities are other PPs or certain subclauses. Signal words are certain particles (dann in the example), adverbs (e.g. begrifflich - conceptually) or nominal constructions containing the definiendum (e.g. der Begriff des..., the concept of...).

Of course many definitions also contain further structural elements that are not present in Example (3). For instance certain adverbials or modal verbs modify the force, validity or degree of commitment to a definition (e.g. only for typical cases). The field of law within which the given definition applies is often specified as a PP containing a formal reference to sections of statutes or simply the name of a statute, document, or even a complete legal field (e.g. Umweltrecht - environmental law). Citation information for definitions is standardly included in brackets as a reference to another verdict by date, court, and reference number.
4. Automatic extraction of definitions

The corpus based pilot study discussed in the last section has on the one hand shown a broad linguistic variation among definitions in reasons for judgement. No simple account, for instance in terms of keyword spotting or pattern matching, will suffice to extract the relevant information from a significant amount of occurrences. On the other hand our survey has shown a range of structural uniformities across these formulations. This section discusses computational linguistic analysis techniques that are useful to identify and segment definitions based on these uniformities.

4.1. Linguistic Analysis

Our current work is based on a collection of more than 6000 verdicts in environmental law that were parsed using the Preds-parser (Preds stands for partially resolved dependency structure), a semantically-oriented parsing system that has been developed in the Saarbrücken Computational Linguistics Department within the project COLLATE. It was used there for information extraction from newspaper text ([6], [7]). The Preds-parser balances depth of linguistic analysis with robustness of the analysis process and is therefore able to provide relatively detailed linguistic information even for large amounts of syntactically complex text.

It generates a semantic representation for its input by a cascade of analysis components. Starting with a topological analysis of the input sentence, it continues by applying a phrase chunker and a named entity recognizer to the contents of the topological fields. The resulting extended topological structure is transformed to a semantic representation (called Preds, see above) by a series of heuristic rules. The Preds-format encodes semantic dependencies and modification relations within a sentence using abstract categories such as deep subject and deep object. This way it provides a common normalized structure for various surface realizations of the same content (e.g. in active or passive voice).

The Preds-parser makes use of syntactic underspecification to deal with the problem of ambiguity. It systematically prefers low attachment in case of doubt and marks the affected parts of the result as default-based. Later processing steps are
enabled to resolve ambiguities based on further information. But this is not necessary in general. Common parts of multiple readings can be accessed without having to enumerate and search through alternative representations. Figure 2 shows the parse for the definition in Example (3). The parser returns an XML-tree that contains this structure together with the full linguistic information accumulated during the analysis process.

4.2. Search and processing

The structures produced by the *Preds* parser provide a level of abstraction that allows us to turn typical definition patterns into declarative extraction rules. Figure 4 shows one such extraction rule. It specifies (abbreviated) XPath-expressions describing definitions such as Example (3). The field *query* contains an expression characterising a sentence with the predicate vorliegen and a subclause that is introduced by the subjunction *wenn* (*if*). This expression is evaluated on the *Preds* of the sentences within our corpus to identify definitions. Other fields determine the locations containing the structural elements (such as definiendum, definiens and domain area) within the *Preds* of the identified definitions.

```xml
<pattern>
  description=liegt vor + wenn-Nebensatz
  query=sent/parse/preds/word[@stem="vorliegen" and INDPRES and WENN]
  filters=definite
  definiendum=DSUB
  definiens=WENN/arg/word
  area=PPMOD{PREP%bei}
</pattern>
```

Figure 4. Extraction rule.

The field *filters* specifies a set of XSLT-scripts used to filter out certain results. In the example we exclude definienda that are either pronominal (because we do not presently resolve anaphoric references) or definite (because these are often also anaphoric, or indicate that the sentence at hand is valid for that particular case only). Figure 3 shows how the definition in Example (3) is analyzed by this rule.

4.3. Evaluation

Based on our observations on the pilot study corpus described in Section 3 and on domain expert knowledge (in cooperation with a trained jurist), we specified two sets of such extraction rules:
Pattern set 1 consists of 33 rules. They were designed with a focus on precise extraction and therefore incorporate a number of generic and pattern specific filtering rules (e.g. constraints on determiners and ordering of definition subparts, stopword lists).
Pattern set 2 incorporates modifications and further rules (it consists of a total of 59 extractors) based on observations about performance problems on the pilot study data. This section evaluates two extraction experiments with the pattern sets just described.

4.3.1. Evaluation on pilot study corpus

In a first experiment we used pattern set 1 to extract definitions from our pilot study corpus (i.e. the data used for their development) and performed a detailed analysis of the reasons for recall failures on that data. One obvious reason is the relatively small number of patterns in our extractor set 1. As mentioned above, pattern set 2 therefore contains 26 additional patterns. Our observations further allowed us to correct errors and to introduce generalizations within some patterns. Additionally, the filtering approach of patterns 1 turned out to be too restrictive, and therefore all in-built filtering mechanisms were removed in pattern set 2. While patterns 1 retrieve 27 of the total verified 93 predicate-based core definitions in the pilot study data, patterns 2 extract 42 of them.

Precision scores varied between 1 and about 0.1 for the different patterns. The diagram in Figure 5 plots both recall and cumulative precision scores over the total number of hits, where all hits are sorted according to the precision of the retrieving pattern. Pattern set 2 is consistently superior to patterns 1 also in terms of precision according to this evaluation. This is quite surprising given the considerable advance in recall achieved by patterns 2 over patterns 1, and we assume that the effect is incidental on the particular data in our pilot study corpus. A comparison of both pattern sets on the full environmental law corpus supports this assumption.

Figure 5: Precision and recall scores of both pattern sets on pilot study corpus
4.3.2. Evaluation on full corpus

In a second experiment, we applied both pattern sets to all of the 6000 documents in our environmental law corpus. Pattern 1 extracted a total of 2364 sentences. A double annotation was performed on 546 of these by one graduate law student and one computational linguist. The sentences to be annotated were selected randomly, but balanced over the various patterns. Annotation consisted in a yes-no-decision. As a rough guideline, the annotators were asked to state for each sentence whether it was likely to be useful for deciding about the application of the definiendum in a reasonable number of cases. The overall agreement of the annotators’ judgements was relatively high in spite of the vagueness of this instruction, with $\kappa=0.71$ on the full 546 annotated sentences.\footnote{Cohen’s $\kappa$ coefficient is a measure of inter-annotator agreement that also accounts for the probability of chance agreement. A value of 1 corresponds to full agreement while a value $\leq 0$ indicates that the agreement is no larger than would be expected by chance.}

15389 sentences were retrieved by patterns 2. 3816 of these were checked by a single (law student) annotator. Of these 3816 sentences, 219 were also amongst the ones judged by annotator 2 in our first experiment. The inter-annotator agreement on this overlapping segment was $\kappa=0.62$.

Based on the annotated hits, precision estimates were calculated for each pattern. As in experiment 1, precision scores for the different patterns varied considerably. Figure 6 gives an overview of the estimated maximum precision achievable for an increasing number of retrieved hits. Like Figure 5, it was obtained by sorting all
extraction results according to the precision estimate for their respective retrieving pattern (using the judgements of annotator 1 for the double annotated items), and then calculating the overall precision within each top-n-segment of this ordering.\footnote{This also means that the precision values that can be read off the diagram are estimates that are based on more data from left to right. The rightmost point corresponds to the total precision score for all of the annotated results and is thus the most reliable estimate.}

The precision scores for patterns 1 this time still lie above those for patterns 2 on around the first 1400 hits and then drop below that line. The total yield of pattern set 2 is much higher than that of pattern set 1. The total number of sentences retrieved by patterns 1 is 2364, compared to 15389 sentences retrieved by patterns 2. The median of the precision (i.e. the precision at which half of all hits can be retrieved) achieved by patterns 1 is still slightly higher than that of patterns 2 on the same number of hits (0.71 compared to 0.68). However while the precision score drops considerably (to 0.48) for the full 2364 hits retrievable by patterns 1, it remains relatively stable (at 0.61) for the same number of hits by patterns 2.

Unlike in our pilot study, we do not have markup of all definitions in our full corpus and can therefore only resort to an estimation to assess the recall performance of our search patterns: Out of the 3509 sentences of the relevant parts of the 40 decisions in the pilot study, a total of 93 sentences were predicate-based core definition statements (cf. Section 3). Taking this rate of about 2.7\% as an orientation, one would expect to find about 6300 such sentences in the corpus used in the extraction experiment just discussed. The price for the relatively high precision scores achieved by our extractor pattern set 1 is thus clearly a very poor recall performance.

Figure 7: Precision scores for rankings according to setting (a), (b) and (c)
4.4. Ranking

For the evaluation of the experiments discussed in the previous section, we used the precision scores that were calculated based on the available annotations for ordering all hits by the precision of their retrieving pattern. The same data was then used to evaluate the ranked results. This procedure is useful as a first approximation (and it was without alternative given the relatively small amount of annotated data available for the experiment with pattern set 1). However it does not cleanly separate training and test data. Generalizations from the evaluation are therefore only possible under the optimistic assumption that the precision scores used for sorting are valid estimates of the performance on unseen data.

Furthermore, the formulation patterns encoded in our extraction rules are most likely not the only linguistic indicators of definitionhood. Presumably there is also a number of additional, more fine grained formulation features that are more or less indicative of definitions.

In this section, we discuss an experiment that looks at rankings for the results of pattern set 2 based on precision scores as well as various other features, and uses a strict separation between test and training data in the evaluation. The experiment was based on the 3816 annotated hits for this pattern set, which were partitioned into training- and test-sets of about the same size in four different ways (randomly, but balanced over the various patterns). The reported results are averaged over the four training/test-splits.

4.4.1. Experiments

We compare three different settings:
(a) A precision-based ranking (such as in Section 4.3), in which the training data in each of the four splits was used to estimate pattern precisions and the corresponding test data was sorted accordingly.
(b) A feature based ranking, in which we used the annotated data to train weights for a linear combination of various linguistic features (using the linear regression module from the WEKA toolkit, cf. [8]). These included lexical (such as the occurrence of certain stop- or boost-words, bag-of-word similarity to a set of known definitions), structural (e.g. embedding level, ordering of surface elements) and domain specific (e.g. occurrence of citations) attributes of each hit.
(c) A combination of settings (a) and (b).

The diagram in Figure 5 compares the precision scores for all top-n segments of the rankings produced by settings (a), (b) and (c), averaged over the four training/test-splits. Setting (b) performs only slightly worse than the ordering based on precision estimates, and the combined setting (c) improves considerably over (b) as well as (a).

4.4.2. Discussion

The relatively good performance of setting 1 (no access to precision estimates) suggests that the weights assigned to our hand selected features by the regression analysis do in fact reflect the importance of these features quite accurately. This in turn allows us to draw conclusions about typical linguistic features of definitions that are missed out by a strict focus on typical definition predicates and their syntactic environment.
Table 1 shows those features that were assigned the strongest positive respectively negative weights in setting 1, when using all annotated instances as training data. Some of the features used in our experiments are specific to definitions that introduce a superconcept (‘Aristotelean definitions’, cf. Section 2), or can only be determined if the main subparts of a (candidate) definition could be identified. The training data was therefore portioned into three groups (with superconcept, no superconcept and no subparts identified) on which linear regression was performed separately. Table 1 reflects this categorization. In the categories no superconcept and no subpart there was only one feature that was given a negative weight.

These results largely conform to the expectations we had when selecting the respective features. References to the concrete case are dispreferred, and in fact sentences with such references are only seldom general enough to qualify as definitions. Negations as well as embeddings get penalized. This presumably reflects the fact that such statements often assert the factual inapplicability of an assumed definition. The feature hit within previous three sentences was given positive weight in all categories (although it only occurs in the top five in the category no subparts). As our annotation pilot study suggests there are sometimes accumulations of definitions, where e.g. one sentence introduces applicability conditions that are then developed further in the subsequent context. A further feature that was weighted positively in all three categories but does not figure in the top five is sentence length in tokens, suggesting that definitions are often particularly long sentences.

However some of the weightings shown in Table 1 are counter-intuitive. It seems surprising that the absence of a legal quote should be strongly indicative of definitions in the category no subparts, but this may be a result of imperfections of the legal quote-recognition component. The fact that the absence of adjectival stopwords from the hand compiled collection is penalized in the with superconcept-category is probably due to the relatively ad-hoc choice of adjectives that was used when that stop-word-list was compiled. A more principled method for stopword compilation, based e.g. on the comparison of texts from different genres, would be desirable.

5. Ontology Extraction

Occurrence of a concept within a definition is likely to indicate that the concept is important for the text at hand. Moreover in court decisions, a great deal of the important (legal as well as subject domain) concepts will in fact have at least some occurrences within definitions. This can be assumed because legal argumentation (as discussed in Section 2) characteristically proceeds by adducing explicit definitions for all relevant concepts. Definition extraction therefore seems to be a promising step for identifying concepts, in particular within legal text. This section discusses how extracted definitions can be used to improve the quality of text-based ontology learning from court decisions. For this purpose we first examine the results of a standard method – identification of terms and potential class-subclass relations through weighted bigrams – and then look at the effect of combining this method with a filter based on occurrence within definitions identified by our extraction system.
Table 1: Features with strong positive or negative weights

<table>
<thead>
<tr>
<th>With superconcept</th>
<th>No superconcept</th>
<th>No subparts</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Positive</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Definiendum precedes super-concept</td>
<td>Definition is top level-clause</td>
<td>Definition is top level-clause</td>
</tr>
<tr>
<td>Legal quote in concept</td>
<td>Definiendum is not a negated NP</td>
<td>Definition contains no legal quote</td>
</tr>
<tr>
<td>Definiendum precedes condition</td>
<td>Definiendum is not a definite NP</td>
<td>Definition is not negated at sentence level</td>
</tr>
<tr>
<td>Definition is top level-clause</td>
<td>Definition is not negated at sentence level</td>
<td>Definition contains modifier</td>
</tr>
<tr>
<td>No nominal stopword as definiendum</td>
<td>Definiendum is anaphoric</td>
<td>At least one hit within previous three sentences</td>
</tr>
<tr>
<td><strong>Negative</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Definition contains concrete case-stopword</td>
<td>Definition contains concrete case-stopword</td>
<td>Definition contains concrete case-stopword</td>
</tr>
<tr>
<td>Concept contains no adjectival stopword from hand-compiled list</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Definition contains no condition</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

5.1. Bigram Extraction

Adjective-noun-bigrams are often taken as a starting point in text based ontology extraction because in many cases they contain two concepts and one relation (see e.g. [9]). The nominal head represents one concept, while adjective and noun together represent another concept that is subordinate to the first one. There are however obvious limits to the applicability of this concept-subconcept-rule:

1. It may happen that the bigram or even already the nominal head on its own do not correspond to relevant concepts, i.e. that one or both of the denoted classes are of no particular relevance for the domain.
2. Not all adjective-noun-bigrams refer to a subclass of the class denoted by the head noun. Adjectives may e.g. be used redundantly, making explicit a part of the semantics of the head noun, or the combination may be non-compositional and therefore relatively unrelated to the class referred to by the head noun.

For these reasons, extracted bigrams generally need to be hand-checked before corresponding concepts can be integrated into an ontology. This time-intensive step can be facilitated by providing a relevance-ranking of the candidates to be inspected. Such rankings use association measures known from collocation discovery (like $\chi^2$, pointwise mutual information or log-likelihood-ratios). But while the elements of a collocation are normally associated in virtue of their meaning, they do not necessarily correspond to a domain concept just by this fact. Moreover, many collocations are non-compositional. An association based ranking therefore cannot solve Problem (2) just mentioned, and only partially solves Problem (1). However it seems likely that the definiendum in a definition is a domain concept,
and for the reasons discussed in Section 2, it can be assumed that particularly many concepts will in fact occur within definitions in the legal domain. In order to investigate this hypothesis, we extracted all head-modifier pairs with nominal head and adjectival modifier from all parsed sentences in our corpus. We then restricted this list to only those bigrams occurring within at least one identified definiendum, and compared the proportion of domain concepts following the concept-subconcept-rule on both lists.

5.2. Unfiltered Extraction and Annotation

We found a total of 165422 bigram-occurrences of 73319 types (in the following we use bigrams to refer to types, not to occurrences) within the full corpus. From this list we deleted combinations with 53 very frequent adjectives that are mainly used to establish uniqueness for definite reference (e.g. vorgenannt – mentioned above). All types with more than 5 occurrences were then ranked by log-likelihood of observed compared to independent occurrence of the bigram elements. The resulting list contains 4371 bigrams on 4320 ranks. Each bigram on the first 600 ranks of this list (601 bigrams, two bigrams share rank 529) was assigned one of the following five categories:

3. Environmental domain: Bigrams encoding concepts from the environmental domain (e.g. unsorted construction-waste). These occur because our corpus deals with environmental law.

4. Legal domain: Bigrams encoding concepts from the legal domain. These range from concepts that are more or less characteristic of environmental law (e.g. various kinds of town-planning schemes) to very generic legal concepts (such as statutory prerequisite)

5. No subconcept: Bigrams that would be categorized as 1. or 2., but (typically for one of the reasons explained above) do not encode a subconcept of the concept associated with the head noun. An example is öffentliche Hand (“public hand”, i.e. public authorities – a non-compositional collocation).

6. No concept: All bigrams that - as a bigram - do not stand for a domain concept (although the nominal head alone may stand for a concept).

7. Parser error: Bigrams that were obviously misanalysed due to parser errors.

Figure 8 shows the distribution of categories among the 600 top-ranked bigrams, as well as within an additionally annotated 100 ranks towards the end of the list (ranks 3400 - 3500). For selecting the two categories of central interest, namely those of legal and environmental concepts to which the concept-subconcept rule applies, the ranking is most precise on the first few hundred ranks, and looses much of its effect on lower ranks. The percentage of such concepts decreases from 56% among the first 100 ranks to 51% among the first 200, but is roughly the same within the first 500 and 600 ranks (with even a slight increase, 45.6% compared to 46.8%). Even the segment from rank 3400 to 3500 still contains 39% of relevant terminology. There are no bigrams of the “no subconcept” category within this final segment. The explanation for this fact is probably that such bigrams (especially the non-compositional

---

4 The ranking was calculated by the Ngram Statistics Package described in [10]
ones) are mostly established collocations and therefore show a particularly high degree of association.

It must be noted that the results of our annotation have to be interpreted cautiously. They have not yet been double-checked and during the annotation process there turned out to be a certain degree of uncertainty especially in the subclassification of the various categories of concepts (1, 2 and 3). A further category for concepts with generic attributes (e.g. *permissible*, combining with a whole range of one-word terms) would probably cover many cases of doubt. The binary distinction between concepts and non-concepts in contrast was less difficult to make, and it is surely safe to conclude about general tendencies based on the annotation.

5.3. Filtering and Combined Approach

By selecting only those bigrams that occur within *definienda* from the definitions identified by our extractor pattern set 1 (cf. 4.3), the 4371 items on the original list were were reduced to 227 (to allow for comparison, these were kept in the same order and annotated with their ranks as on the original list). Figure 9 shows how the various categories are distributed within the items selected from the top segments of the original list, as well as within the complete 227 filtering results. The proportion of interesting concepts reaches about 80% and is higher than 60% on the complete selection. This is still well above the 56% precision within the top 100-segment of the original list. However the restriction to a total of 227 results on our filtered list (of which only 145 are useful) means a dramatic loss in recall. This problem can be alleviated by leaving a top segment of the original list in place (e.g. the top 200 or 500 ranks, where precision is still at a tolerably high level) and supplementing it with the lower ranks from the filtered list until the desired number of items is reached. Another option is to apply the filtering to the complete list of extracted bigrams, not only to those that occur more than 5 times. We assume that a concept that is explicitly defined is likely to be of particular relevance for the domain regardless of its frequency. Hence our definition-based filter should still work well on concept candidates that are too infrequent to be considered at all in a
log-likelihood ranking, and allow us to include such candidates in our selection, too.

We investigated the effect of a combination of both methods just described. For this purpose, we first extracted all noun-adjective bigrams occurring within any of the identified definienda, regardless of their frequency within the corpus. After completing the annotation on the 627 resulting bigrams they were combined with various top segments of our original unfiltered list.

Figure 10 shows the distribution of the annotated categories among the 627 bigrams from definienda, as well as on two combined lists. Cutoff 200/750 is the result of cutting the original list at rank 200 and filling up with the next 550 items from the filtered list. For cutoff 500/1000 we cut the original list at rank 500 and
filled up with the following 500 items from the filtered one. The distribution of
categories among the original top 200 is repeated for comparison.

Precision among the 627 filtering results is higher than among the original top 200
(almost 56% compared to 51%), and only slightly smaller even for the 1000 results
in the cutoff 500/1000 setting. Using definition extraction as an additional knowl-
edge source, the top 1000 results retrieved are thus of a quality that can otherwise
only be achieved for the top 200 results.

6. Conclusion

In this paper we argued that definitions are an important element of legal texts and
in particular of court decisions. We provided a classification of definition types as
well as a structural segmentation scheme for definitions and discussed a method of
applying computational linguistic analysis techniques for their text-based extrac-
tion and automatic segmentation. We showed that a large number of definitions
can in fact be extracted at high precision using this method, but we also pointed
out that there is still much room for improvement in terms of recall, e.g. through
the inclusion of further definition patterns. Moreover we showed how an automati-
cally acquired ranking based on a combination of various features can improve the
results of a strictly rule based approach to definition extraction.

Our future work in this area will focus on the integration of extraction results
across documents (e.g. recognizing and collecting complementary definitions for
the same concept) and on a user interface for structured access to this data. For this
work we have access to a corpus of several million verdicts provided to us by the
company juris GmbH, Saarbrücken.

We also demonstrated how the identification of definitions can improve the results
of text-driven ontology learning in the legal domain. When looking for noun-
adjective bigrams encoding relevant concepts, it leads to a considerable increase in
precision to restrict the search to definienda only. This method is more precise than
selecting the top ranks of a log-likelihood ranking. Its great disadvantage is the
very low total number of results, leading to poor recall. However by combining a
log-likelihood ranking with definition-based concept extraction, recall can be im-
proved while still achieving better precision than with a log-likelihood ranking
alone. Moreover this combined method also retrieves concepts that are too infre-
quent to be included at all in a log-likelihood ranking.

There is however another, maybe even more relevant reason to look for definitions
in ontology learning. Definitions in legal text often very explicitly and precisely
determine all kinds of relational knowledge about the defined concept. For in-
stance they specify explicit subordinations (as in the classical definitio per genus et
differentiam), introduce restrictions on roles inherited from a superconcept, deter-
mine the constitutive parts of the definiendum, or contain information about its
causal relations to other concepts. As one focus of our future work we plan to in-
vestigate how such rich ontological knowledge can be extracted automatically.
References

This page intentionally left blank
Automatic Argumentation Detection and its Role in Law and the Semantic Web

Raquel MOCHALES PALAU, Marie-Francine MOENS

Department of Computer Science
Katholieke Universiteit Leuven
B-3001 Heverlee, Belgium

Abstract. The automatic detection of arguments in text regards a relatively new area at the intersection of Natural Language Processing, Information Retrieval and Legal Information Systems. This paper presents some fundamental issues when processing texts that contain argumentation. Furthermore, our research bridges different areas, including the legal field and the Semantic Web, where argumentation detection and reconstruction could be beneficial. Finally, it analyses several methodologies to accomplish this task, providing results from different experiments done over several kinds of texts, specially legal reports.

Keywords. Argumentation detection, Argumentation structure, Semantic Web.

Introduction

In the last decades the study of argumentation has evolved as a field of interest in its own right. Several associations, academic journals, books, conferences and courses are devoted to argumentation. The growing interest in argumentation is an international phenomenon and it is not restricted to one discipline. Originally grounded in philosophy, it abuts onto cognitive science, linguistics and communication theory. Argumentation embraces the sciences of debate, dialogue, conversation and persuasion. It is concerned primarily with reaching conclusions through reasoning, that is, claims based on premises. Argumentation is used in daily life, e.g. discourses, news, books, opinion sections, political speeches or blogs, and specially in the legal domain, e.g. in trials, presenting arguments to a court, or testing the validity of certain kinds of evidence.

The World Wide Web is considered as our largest library of information, therefore we expect to find many argumentative texts on the Web. The Semantic Web is an evolving extension of the World Wide Web. It creates a Web where content is not only expressed in natural language, but also in a formalized language that can be read and used by the machine - in particular by software agents -, thus permitting the agents to find, share and integrate information more easily, and eventually make their decisions while reasoning with this formal knowledge. Arguments pro or contra a certain statement influence the reasoning process, hence their detection in natural language texts is a relevant research topic.

In this paper we use the more general term "argumentation" for any act or process of forming reasons, drawing conclusions, and applying them to the case that is subject of
the discussion. Following [17] we use the more specific term "argument" when referring to a specific discoursive section of this argumentation that is intended to persuade or convince. We describe an argument as a set of textual clauses, consisting of a number of premises, a number of inferences, and a conclusion. Furthermore, an argument is said to have the following property: if the premises are true, then the conclusion is highly likely to be true.  

The main goal of this paper is to draw the attention to the importance of automatic argumentation recognition in legal text and place this recognition in the context of the Semantic Web given the few studies on this matter. We also want to outline the main steps that are needed to achieve the automatic detection of argumentation and its structure, and the problems that come along with this task. Furthermore, even if there is still a need for further research, we already present different experiments that prove it is currently possible to achieve some important tasks on the automatic detection of argumentation.

This paper is organized as follows. First we present an overview of the main characteristics and problems argumentation must deal with. This is followed by the influence of automatic argument detection in the frame of the legal Semantic Web and the Semantic Web in general. The final part focuses on our work in automatic argument detection, the problems encountered and the results obtained so far.

1. Background: Argumentation Fundamentals

Argumentation can be defined as a verbal and social activity of reason aimed at increasing (or decreasing) the acceptability of a controversial standpoint for the listener or reader, by putting forward a constellation of propositions (i.e. arguments) intended to justify (or refute) the standpoint before a rational judge [23]. Each argument of the argumentation has some propositions, i.e. premises, to arrive to a conclusion, sometimes implicit, that justify the standpoint of the argumentation. Argumentation is familiar to all of us. Both oral and written argumentation are indeed integral parts of our daily routines. Furthermore, it also has an important role in philosophy, logic, linguistics, discourse analysis, sociology, rhetoric, political science, law and many other disciplines.

Argumentation for or against a standpoint can exhibit a simple or a more complex structure, depending on how the arguer organizes the defense of his standpoint in view of doubts and criticisms. The simplest case is that of a single argumentation, consisting of just one argument pro or contra the standpoint. Single argumentation is considered to display the basic structure of argumentation, because complex argumentation can always be analyzed as a combination of single argumentations. In typical single argumentation the basic argument contains one conclusion and two premises, one explicit and one unexpressed [23], e.g. "You are a chauvinist, because you are a man" where "Men are chauvinists" is unexpressed but a necessary element of the argument.

In more complex, structured argumentation, different arguments are advanced pro or contra the same standpoint. These arguments can be unconnected or interdependent [23]. In the first case, the argumentation consists of alternative defenses of the same standpoint and can be structurally characterized as a multiple argumentation. In the second case, a compound argumentation is created consisting of a chain of arguments that reinforce

---

1 Our research on argument detection should not be confused with the research on argument detection in Natural Language Processing, where argument refers to constituents or adjuncts to a predicate in a sentence.
each other. The single arguments constituting the chain can be "connected in parallel" or "connected in series" [23]. Arguments connected in parallel are a combined attempt to defend a single standpoint; the argumentation can be structurally characterized as coordinately compound argumentation. On the other hand, arguments connected in series, i.e. the one supporting the other, create a subordinatively compound argumentation. Whereas the arguments connected in parallel relate directly to the standpoint, the first argument of the arguments connected in series relates directly to the standpoint and the second to the reason presented in the first argument, which serves now as a "substandpoint" and so forth.

A major problem is the analysis of complex argumentation structure. Explicit indications of the argumentation structure in the discourse are generally scarce. As a rule, the human analyst will have little more to fall back on than the connectives and other clarifying linguistic expressions that are used in the text. Another complication is that various types of argumentation structures may occur together in the discourse and be coordinatively or subordinatively combined (see Table 1).

As a consequence, we expect the automated analysis of argumentation in text to be a difficult task. In addition, each single argument, of which the argumentation is composed, might be characterized by its own complex internal organization. An argument can have a simple structure, containing just two elements, i.e. an explicit premise and a conclusion. It can, also, present a more complex structure, but in contrast with the complex argumentation structure, the argument complexity can only be due to coordinate premises, i.e. one premise alone does not imply the conclusion. An argument with subordinate premises would be seen as two arguments where the premise of the first is the conclusion of the second. An argument with multiple premises, non-coordinated, would be seen as two arguments sharing a conclusion.

Furthermore, as shown by examples in Table 2, an argument can be formulated using different writing styles. The typical case presents argumentative markers such as words like "but", "hence", "so", "since" or "therefore" that ease the argument recognition and analysis. However, in ordinary discourse such indicators can be normally confused with rhetorical markers or even be frequently absent. Thus, it is also usual to present an argument by statements consisting simply of impersonal indicative sentences or even by questions. Also, most arguments have implicit premises or conclusions. The fewer the number of argumentative markers and the higher the implicit information, the more it will be necessary to make use of the context, type of speech and background knowledge to detect the different arguments of the discourse.

Another important point in the analysis of argumentation is that it is known that people rely on more or less ready-made argumentation schemes; conventionalized ways of displaying a relation between that which is stated in the premises and that which is stated in the conclusion, and how all the involved arguments relate between themselves [23]. Over the years there have been different theories to describe these argumentation schemes, to determine how they operate, and to assess their acceptability [22], [25], [13]. These schemes try to define the possible functions fulfilled by each clause inside an argument and also by each argument inside the whole argumentation, e.g. a premise can be a rebuttal or a claim while the argument can be based on argumentation by ignorance or by precedent.

The study of argumentation has so far not resulted in a universally accepted theory. The state of the art can therefore not be explained by describing one leading theory. It
Table 1. Complex argumentation structures

| Coordinatively structure | Arg1 | Doyle is not expected to win, so his defeat would be no skin off the prime ministerial nose. |
| Arg2 | But if Doyle did better than expected, Howard, who is soaring high, would receive some credit. |
| Standpoint | For Howard, Victoria is a no-lose situation |

| Multiple structure | Arg1 | The decision to quit drugs cannot be enforced from the outside. Drag workers talk about a "circuit-breaker", the event that triggers a person’s decision to quit - a long-suffering partner walking out, a valued job being lost. The addict realises what they stand to lose if they don’t stop using. This realisation is an internal one. |
| Arg2 | Locking up addicts against their will would not only violate their rights; it would be likely to prove counter-productive. Addiction means the person has lost control over their drug use. The addiction cannot be “cured” by reducing still further that person’s control over their own lives and circumstances, as Watters is suggesting. |
| Standpoint | Watters suggestion is disturbing, unworkable and potentially dangerous. |

| Subordinatively structure | Arg1 | If the South African deficit is lower than the deficits European economies struggle to stay below, South Africa can be less stringent about its deficit. |
| Arg2 | European economies struggle to hold their deficits at below 3% ours will grow to 2.4% over the next two years. |
| Standpoint | One can also protest that there is scope for less stringency on the budget deficit. |

| Mixed structure | Arg1 | He went to Cincinnati and he gave a speech in which he said, “We will plan carefully. We will proceed cautiously. We will not make war inevitable. We will go with our allies.” He didn’t do any of those things. They didn’t do the planning. They left the planning of the State Department in the State Department desks. They avoided even the advice of their own general. |
| Arg2 | The terrorism czar, who has worked for every president since Ronald Reagan, said, “Invading Iraq in response to 9/11 would be like Franklin Roosevelt invading Mexico in response to Pearl Harbor.” |
| Standpoint | The president is not able to plan and organize war operations |

is characterized by the coexistence of a variety of approaches, differing considerably in conceptual breadth, scope of horizon, and degree of theoretical refinement. As a consequence, automatic argument detection is dependent on the theories that are used to analyse the argumentation. Even though, automatic argument detection can be divided in two
main tasks, (a) argumentation structure detection and (b) argument structure detection. Argumentation structure detection consists of detecting the arguments in a text and the relations between them. Argument structure detection focuses on a single argument and analyses its inner structure, i.e. the distribution of premises and conclusions inside it.

2. Automatic Argumentation Detection: The Reasons

Argumentation based techniques have a wide range of applications in both theoretical and practical branches of artificial intelligence and computer science [18] ranging from non-monotonic reasoning [18,4], multi-agent systems’ communication and negotiation [14,15] to knowledge engineering [2] or argumentation-support systems [7]. The automatic detection of argumentation can help in most of these research areas and others. For example, in information systems, where information searches comprise a substantial amount of time, it is interesting to visualize the argumentation structure of a text, as is done in Araucaria [19] or ArguMed [24]. These tools assist in drafting this structure by allowing to manually drag and enter text into a graph or tree structure that represents the argumentation. The manual structuring of an argumentative text as is done in these projects is a very costly job. The automatic detection of all the arguments, their structure and the general argumentation structure of the discourse will reduce the amount of time involved in each analysis.

Automatic argumentation detection could also help in the indexing of texts for retrieval purposes. Indexes could be generated that reflect the standpoints defended in each text or by the type of scheme followed in each argument used to defend it. Also the summarization of documents could benefit from the detection of all the elements of an argument. Identifying which sentences relate to each standpoint of the document can help to extract the main elements of the argumentation, reducing the document length. This can be compared to the idea of [21] where roles, such as "Background" or "Own", are given to all the sentences of a scientific text to be able to summarize the groups of sentences depending on their role in the document.

Legal texts could be compared taking into account the arguments presented in them or how these arguments interact between themselves, i.e. its general argumentation structure. Recognizing the argument structure or scheme in cases is also a valuable step in automatically qualifying arguments and describing them with conceptual factors from a legal ontology that play a role in the decision, and reasoning with these factors in case-
based reasoning systems. Also a European project such as ALIS 2, which aims to facilitate compliance with existing laws and regulations of governmental actions and the prevention of conflict, could benefit from the automatic detection of legal arguments when searching for legal documents in their vast databases. Governments, bureaucracies and companies, as well as individuals and groups, could also benefit of the automatic argument detection when developing tools or adapting processes to meet the aspirations of e-democracy, e-participation or e-business. The study of arguments written in blogs or newsgroups could support the development of these new areas.

Furthermore, the Semantic Web could also benefit of the automatic detection of arguments. Rahwan et al. [16] state a long-term vision where one queries documents of the Web by asking a question like "List all arguments that support the War on Iraq on the basis of expert assessment that Iraq has Weapons of Mass Destruction (WMDs)". Then, one will be presented with various arguments ordered by strength (calculated using the number and quality of its supporting and attacking arguments). One of these arguments can be a blog entry, with a semantic link to a CIA report claiming the presence of WMDs. Therefore, one could inspect the counterarguments to the CIA reports and find an argument that attacks them by stating that "CIA experts are biased". One could also inspect this attacking argument and find a link to a BBC article discussing various historical examples of the CIA’s alignment with government policies, and so on. However, [16] base this long-term vision on a large-scale Web of inter-connected arguments posted by individuals on the World Wide Web in a structured manner, which is far from being the case. The automatic detection of arguments could go further and automatically extract the argument structures from already existing Web data, overcoming the need to force individuals to write in a non-natural structured way.

Table 3. Examples sentences Araucaria [19] corpus

<table>
<thead>
<tr>
<th>Text type</th>
<th>Argumentative Sentence</th>
<th>Non-Argumentative sentence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Discussion fora</td>
<td>&quot;On this occasion, however, I shall not vote for any individual or party but will spoil my paper.&quot;</td>
<td>&quot;I have been voting since 1964 and at one time worked for my chosen party.&quot;</td>
</tr>
<tr>
<td>Legal judgments</td>
<td>&quot;He is aware of the risks involved, and he should bear the risks.&quot;</td>
<td>&quot;Let there be any misunderstanding one point should be clarified at the outset.&quot;</td>
</tr>
<tr>
<td>Newspapers</td>
<td>&quot;Labor no longer needs the Liberals in the Upper House.&quot;</td>
<td>&quot;The independents were a valuable sounding board for Labor’s reform plans.&quot;</td>
</tr>
<tr>
<td>Parliamentary records</td>
<td>&quot;I have accordingly disallowed the notice of question of privilege.&quot;</td>
<td>&quot;Copies of the comments of the Ministers have already been made available to Dr. Raghuvansh Prasad Singh.&quot;</td>
</tr>
<tr>
<td>Weekly magazines</td>
<td>&quot;But for anyone who visits Rajasthan’s Baran district, the apathy of the district administration and the failure of the Public Distribution System (pds) is clear to see&quot;</td>
<td>&quot;This time in Rajasthan.&quot;</td>
</tr>
</tbody>
</table>

2www.alisproject.com
3. Automatic Argument Detection: Methodology

In real life, there are no easy mechanical rules to identify arguments, and persons usually have to rely on the context in order to determine which are the premises and the conclusions. However, it is true that sometimes the presence of certain indicators facilitates the detection of the premises or the conclusion. For example, if someone makes a statement, and then adds “this is because”, then it is quite likely that the first statement is a conclusion, supported by the statements that come afterwards. The skill of distinguishing arguments from non-arguments is sophisticated and requires training: it is a typical learning outcome of an undergraduate critical thinking course. The analysis of arguments, including the categorization of text by an argumentation scheme and so recognizing a complete argumentation, is more challenging yet, and faces the additional problem that multiple analyses may be possible [11]. Our research on how to automatically extract argumentation and its structure, and to classify it in different schemes has only started. Until now, we have focused on the detection of argumentative sentences and more in concrete on the detection of premises and conclusions. Up until now we studied three methodologies, presented in the next sections, but there should be other different approaches to solve the different problems of automatic argumentation and argument detection.

3.1. Classification Algorithms

A simple approach to automatically detecting argumentative sentences is to consider this task like in a classification problem, where the classifier is trained on examples that are manually annotated. For instance, if we want to classify whether a sentence contains a premise or a conclusion, each example sentence in our training set, which is manually classified by a human, is represented as a vector of features that serve as input for state-of-the-art pattern recognition algorithms, such as a multinomial naive Bayes classifier or a maximum entropy model. The features are generic and are easily extracted from the texts, such as statistical measurements or syntactical characteristics. A classification model is learned and can then be applied to new, unseen sentences that are represented with the same features.

In [11] we presented some tests using this methodology over the Araucaria corpus, a collection of English argumentative sentences extracted during a six week period in 2003 from different sources, during which time a weekly regime of data collection scheduled regular harvest of one argument from each of 19 newspapers (from the UK, US, India, Australia, South Africa, Germany, China, Russia and Israel, in their English editions where appropriate), from 4 parliamentary records (in the UK, US and India), from 5 court reports (from the UK, US and Canada), from 6 magazines (UK, US and India), and from 14 further online discussion boards and "cause" sources such as Human Rights Watch and GlobalWarming.org. These sources were selected because they offered (a) long-term online archive of material; (b) free access to archive material; (c) reasonable likelihood of argumentation. Each week, the first argument encountered in each source was identified and analysed by hand (Table 3).

The Araucaria corpus comprises 1899 sentences that contain an argument and 1854 sentences without argument, which we used for our experiments. The sentences of the Araucaria corpus were studied in isolation, i.e. out of context, and were classified as containing or being an argument or not. The statistical classifiers used were a multinomial
naive Bayes classifier and a maximum entropy model. We compute here the accuracy of the detection as the number of correctly classified sentences divided by the number of sentences that were classified, and average the values obtained in a ten-fold cross-validation. A classification by type of source text was also done. The features used can be seen in Table 4. These experiments showed that with this technique it was possible, in the Araucaria corpus, to detect when a sentence was or contained an argument with nearly 73% accuracy. We also observed that legal texts showed good characteristics for the automatic detection of argumentative sentences. This simple experiments thus provided promising results, encouraging new studies using more complex features and different texts. In Table 5 we present different sentences that were wrongly classified together with the main reason for the misclassification.

<table>
<thead>
<tr>
<th>Feature</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unigrams</td>
<td>Each word in the text.</td>
</tr>
<tr>
<td>Bigrams</td>
<td>Each pair of successive words.</td>
</tr>
<tr>
<td>Trigrams</td>
<td>Each three successive words.</td>
</tr>
<tr>
<td>Adverbs</td>
<td>Detected with a POS tagger.</td>
</tr>
<tr>
<td>Verbs</td>
<td>Detected with a POS tagger.</td>
</tr>
<tr>
<td>Modal auxiliary</td>
<td>Binary feature that indicates if a modal auxiliary (<em>must, need to, have to, should, ought to, had better, can, could, may, might, will, shall and would</em>) is present.</td>
</tr>
<tr>
<td>Word Couples</td>
<td>All possible combinations of two words in the text.</td>
</tr>
<tr>
<td>Text Statistics</td>
<td>Sentence length, average word length and number of punctuation marks.</td>
</tr>
<tr>
<td>Punctuation</td>
<td>Different patterns of successive punctuation marks.</td>
</tr>
<tr>
<td>Keywords</td>
<td>286 words suggested by [8] WHAT TYPES OF WORDS ARE THEY/ DEFINE SHORTLY?</td>
</tr>
<tr>
<td>Parse features</td>
<td>Depth of the tree and number of subclauses of each tree.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Sentence</th>
<th>Arg/Non-Arg.</th>
<th>Classified as?</th>
</tr>
</thead>
<tbody>
<tr>
<td>&quot;So many dead people were lying in the area. We do not know who is who, but the raid was a success,&quot; interim government spokesman Abdinhaman Dimari told AFP news agency.</td>
<td>Non-Arg.</td>
<td>Arg.</td>
</tr>
<tr>
<td>The government of Angola, which yesterday joined the Organization of Petroleum Exporting Countries (OPEC), should publicly account for how it spends the countries massive oil wealth instead of harassing citizens who criticize corruption.</td>
<td>Non-Arg.</td>
<td>Arg.</td>
</tr>
<tr>
<td>The baby, whose stepfather denies murder, was examined by pathologist Dr. James Grieve shortly after her death.</td>
<td>Arg.</td>
<td>Non-Arg.</td>
</tr>
<tr>
<td>Yet participants cautioned against too much finger pointing.</td>
<td>Arg.</td>
<td>Non-Arg.</td>
</tr>
<tr>
<td>What I have heard my Republican colleagues on the other side of the aisle say is that they need more time to look at this.</td>
<td>Arg.</td>
<td>Arg.</td>
</tr>
<tr>
<td>This maniac has got what he deserves.</td>
<td>Non-Arg.</td>
<td>Non-Arg.</td>
</tr>
</tbody>
</table>
3.2. Argument Semantics

It is often the case that we need additional semantic information to precisely detect the inter-sentential relations. An example of this is presented in the following sentences: “The government should apply in my case the same pension rights than on a married woman case. We were never married, but I lived with him a long time as his wife and I called myself Mrs.”. Here, the second sentence would be classified as a simple explanation and not as an argument, when the first sentence is not taken into account. How can we capture the information relating to all the sentences that are part of an argument? Pronouns, generalizations, specifications, negations, cue-phrases, synonyms or antonyms could help in identifying a hidden relation structure between all the sentences of a text. Therefore, a possible solution could be to check all the semantic links between the different sentences inside an argument, to see how they relate between themselves. In the previous example it is possible to find the following semantic relations: “married - never married”, “married woman - wife”, “married woman - Mrs.” and “married - lived with”. However, to solve these coreferring expressions is not an easy task and has been studied during many years resulting in different approaches such as [3] or [12]. But to deal with synonyms/antonyms or negative relations has a major difficulty due to the vast context and semantic complexity involved in their identification. Furthermore, there is a lack of complete tools able to extract list of detailed word relations. In the previous example the relations “married woman - Mrs.” or “married - lived with”, could be quite hard to identify. These relations need world knowledge to detect their synonymy, so a semantic thesaurus would be needed, e.g., WordNet. However, this kind of resource needs a regular updating due to a world that is in constant change, and it is not always easy to correctly disambiguate the meaning of a word, so that the correct synonyms can be found. In conclusion, a lexico-semantic resource such as WordNet could be helpful to identify inter-relations between sentences, but should rather be seen as an assisting resource. Therefore, we did not extend this approach with experimental work.

3.3. Argument Structure

Even if argumentation theorists have not arrived at a consensus over which is the unique argumentation theory, most of them agree there is a clear structure behind each argumentative text and more specifically behind each argument. Therefore, it could be possible to use this structure to improve the detection of argumentation and its arguments.

3.3.1. Related Research: Rhetorical Structure Theory

In the past some efforts have been made to construct formalized discourse models, however none of these studies specifically explores the identification of argumentation in text. These models can be easily implemented in natural language processing applications and one of the most successful theories is Rhetorical Structure Theory (RST). RST describes what parts or segments texts have and what principles of combination can be found to combine parts into entire texts. It assumes a text to have a hierarchical organization based on asymmetrical nucleus-satellite relationships. This means that pairs of adjacent elementary units combine into parent units or text spans, which are again recursively merged until at a certain point one unit spans the entire text. The constituent halves of any text span are linked together by a text structuring relation, which typically holds
between a semantically more central unit, the nucleus, and a more peripheral one, the
satellite (although there also exists a small set of multi-nuclear relations which consist of
two or more equivalent units). Rhetorical relations are not mutually exclusive according
to Hovy [5] as he states that it is possible to assign different relations to the same text
fragments. Although in theory the set of rhetorical relations is open, it is generally
assumed that the number of relevant rhetorical relations is relatively small. The classifica-
tion designed by Mann and Thompson [9] seems to have been accepted by some part of
the discourse analysis community. The result of a complete rhetorical analysis is a tree
structure, every node of which represents a rhetorical relation. However, the nucleus-
satellite distinction does not always reflect the relations of the corresponding text units
correctly, the authors of [27] argue and prove with a set of 135 texts that trees cannot
adequately represent many coherence structures in natural language texts. The discourse
structure of these texts contains various kinds of crossed dependencies as well as nodes
with multiple parents. Neither phenomenon can be represented by using trees. This pa-
per pleads that a chain graph\(^3\) representation is better than a tree. In a labeled coherence
chain graph, an ordered array of nodes represents the discourse segments; the order in
which the nodes occur reflects the temporal sequencing of the discourse segments. La-
beled directed or undirected edges represent coherence relations that hold among the dis-
course segments. RST may be very successful in providing a formalization of discourse
structure, but it nevertheless has some inherent shortcomings that make it not so good for
detecting argumentative relations.

\(^3\)A chain of a graph \(G\) is an alternating sequence of vertices and edges \(x_0, e_1, x_1, e_2, \ldots, e_n, x_n\), beginning
and ending with vertices in which each edge is incident with the two vertices immediately preceding and
following it.

Figure 1. Example of an RST sentence analysis [1]
For example:

- The theory has been primarily designed for analyzing monologue discourse [9] while argumentative texts have some inherent dialogue properties (e.g. the alternation of the plaintiff’s and the defendant’s arguments in a case text).
- Arguments are often constructed in a highly parallel arrangement in which the most applicable RST relation for many inter-premise and inter-argument spans is “joint” (which functions in RST as a catch-all when other relations are inapplicable).
- RST relations can fail to capture argumentative relations and moves. Inferential structures (like modus tollens or disjunctive syllogism) and rhetorical figures (such as ad hominem or ad populum arguments) are simply functioning at a different level to that captured by RST.

3.3.2. Joining Argument Structure and Automatic Argument Detection

As stated in the previous sections, several argumentation models have been developed by different authors. It is not the purpose of this paper to make a comparison study between them or to decide which is the right one. The key issue for us is that all of them agree on the existence of an argumentation structure, consisting on a set of arguments, which in turn also have a inner-structure. For example, one of the most relevant studies is the structure proposed by Toulmin [22], and more recently the interest of the research community has focused on the structure behind argumentation schemes, best explained in a book written by Walton [25].

We have performed preliminary experiments to determine how the argumentation structure, and also the argument structure, could be used for the automatic detection of arguments. Given the current interest on argumentation schemes in the legal community we have chosen to take this model as our starting point. Walton [25] gathers together twenty-five different argumentation schemes, which capture common, stereotypical patterns of reasoning which are nondeductive and nonmonotonic. For example, the following instance of a presumptive, defeasible argument is presented as an instantiation of a particular scheme, the argument from waste: “A PhD student, Susan, has spent more than five years trying to finish her thesis, but there are problems. Her advisers keep leaving town, and delays are continued. She contemplates going to law school, where you can get a degree in a definite period. But then she thinks: Well, I have put so much work into this thing. It would be a pity to give up now”. Argumentation schemes are the forms of argumentation that enable one to identify and evaluate common types of arguments in everyday discourse [20]. The list of presumptive argumentation schemes given in [25] is not complete, but identifies many of the most common forms of defeasible argumentation. In further work Walton and other authors have defined other schemes and found more distinctive kinds of arguments [26,13,6], which we did not study yet.

In these argumentation schemes the only argumentative elements interacting are premises and conclusions. In our research [10] the schemes were used to determine which parts of the text are part of an argument, and how they relate between themselves and with the other arguments. The experiments detected which sentences are argumentative, where sentences were considered in a window of $N$ elements in order to capture contextual features. We represented a sentence as a vector of features and trained a maximum entropy classifier on examples that were manually annotated. The features
Argumentation schemes in Araucaria [19] are generic and can be easily extracted from the texts (Table 4). The feature vectors of the training examples serve as input for the classification algorithm. For these tests we did not divide the sentences depending on their argumentation scheme as we are just determining if a sentence contains an argument (or part of it) or not, more specifically we determine whether the sentence contains a premise or a conclusion, or is a non-argumentative sentence. We did not study the interaction between the different schemes, or test if they are interrelated. We left this for further investigation, that will focus on argument interaction and behaviour.

The corpus used during these tests comprise an English structured set of legal documents, admissibility reports and legal cases. The data was collected from the European Court of Human Rights (ECHR), concretely from the collections of August 2006 and December 2006. This source was selected because it offers (a) a long-term online and cd archive of material; (b) clean and structured legal texts; (c) a reasonable likelihood of legal argumentation. We randomly selected 54 documents composed of 25 legal cases and 29 admissibility reports, that have similar discourse structures, divided by facts, complaints, the law and a final conclusion from the judges. Some of the cases also present a partly dissenting opinion of one or more judges, included at the end of the case. The language of these texts is formal, written mainly in the third person with some direct references to what each party stated in the court.

Those documents contain between 40 and 300 sentences, having an average size of 145 sentences per document. Each sentence contains an average of 49 tokens, that made
a final total average of more than 6,900 tokens per document. We would like to remark that an average of almost 50 tokens per sentence represents a corpus with quite long sentences. This is in accordance with general knowledge about legal texts that are characterized by long and complex statements. Table 6 shows some examples of sentences extracted from the ECHR corpus.

Table 6. Examples sentences ECHR corpus

<table>
<thead>
<tr>
<th>Non argumentative</th>
<th>Premise</th>
<th>Conclusion</th>
</tr>
</thead>
<tbody>
<tr>
<td>The applicant is a Greek national born in 1947.</td>
<td>The Commission notes that the applicant’s conviction involved his writings.</td>
<td>Consequently, it is essentially the exercise of the applicant’s freedom of expression with which there has been an interference.</td>
</tr>
<tr>
<td>By telefax of April 1994, the applicant’s representatives informed the Commission that the applicant had been killed four days before.</td>
<td>However, while basing his case on the narrower domestic criminal law provision of Article of the Greek Penal Code, the applicant claimed the right to use the words “Turk” or “Turkish” to identify the Moslems of Western Thrace.</td>
<td>The purported remedies were thus ineffective for the following reasons:</td>
</tr>
</tbody>
</table>

During four weeks the documents were manually analysed by two lawyers. The lawyers were given a set of documents, where each document was divided in two sections. The top section of each document contains a full case from the ECHR, so the lawyer could read it and analyse the case. The lower section contains the same case structured by grammatical clauses. For each clause we asked to annotate if it is part of an argument. In the affirmative case we asked to determine from which argument the clause originates, which scheme is followed, and which role, i.e. premise or conclusion, of the argument is expressed. Note that a premise or a conclusion can be formed by more than one clause, therefore, also the enumeration of the premises and conclusions inside the argument is maintained.

We asked the two lawyers to annotate the same corpus in order to compute inter-annotator agreement. In the final corpus there are a total of 12,904 analysed sentences, 10,133 non-argumentative and 2,771 argumentative, 2,355 premises and 416 conclusions. The presence of non-argumentative sentences is considerably higher due to the fact that even in legal texts the presentation of facts, explanations and summaries are more frequent than the argumentation itself. It is also evidenced that the amount of premises was much higher than the amount of conclusions. This can be easily explained by the need of different premises, that are in favour or against one conclusion.

The results with the maximum entropy classifier, over the ECHR corpora, confirm that a general study of a sentence context, such as a simple bag-of-words of previous/next $N$ sentences, helped to increase the base accuracy of detecting the argumentativity of the sentence, obtained just with basic linguistic features (see Table 4) from the given sentence. Furthermore, it is possible to detect with accuracy higher than 90% which sentences are non-argumentative in a normal legal text. However, working on a sentence level does not deal with inter-sentence relations, i.e. between the sentence clauses, be-
ing impossible to classify sentences that contain at the same time a premise and a conclusion. Furthermore, the current approaches show poor results on the distinction between premises and conclusions, given the short number of examples compared to non-argumentative sentences, specially for conclusions with nearly 55% precision and only 16% recall but also for premises, with 76% precision and 60% recall.

4. Conclusions

The work done till now has proved that automatically detecting the difference between an argumentative sentence and a non-argumentative sentence in a text is possible and this task can be accomplished with good accuracy based on the simple linguistic features of Table 4 and a bag-of-words from $N$ previous sentences or next sentences. In addition, we are already able to detect whether an argumentative sentence contains a premise or a conclusion. Nevertheless, these results have not been as good as expected, therefore our work will focus on the improvement of the distinction between premises and conclusions. Furthermore, automatic argument detection is still a new research area and several argumentation schemes that were theoretically developed could be discovered in the texts and evaluated upon their validity as argumentation representations. We have demonstrated that there are many applications for argumentation detection in the frame of Semantic Web technology, and more specifically with regard to the legal domain, therefore, more research on this topic is necessary.

References

This page intentionally left blank
Legal Knowledge Applications

Applying Semantic Web Technologies
This page intentionally left blank
Matching Law Ontologies using an
Extended Argumentation Framework
based on Confidence Degrees

Cassia TROJAHN\textsuperscript{a,1}, Paulo QUARESMA\textsuperscript{a} and Renata VIEIRA\textsuperscript{b}

\textsuperscript{a} Departamento de Informática, Universidade de Évora, Évora, Portugal

\textsuperscript{b} Faculdade de Informática, Pontifícia Universidade Católica do Rio Grande do Sul, Porto Alegre, Brazil

Abstract. Law information retrieval systems use law ontologies to represent semantic objects, to associate them with law documents and to make inferences about them. A number of law ontologies have been proposed in the literature, what shows the variety of approaches pointing to the need of matching systems. We present a proposal based on argumentation to match law ontologies, as an approach to be considered for this problem. Argumentation is used to combine different techniques for ontology matching. Such approaches are encapsulated by agents that apply individual matching algorithms and cooperate in order to exchange their local results (arguments). Next, based on their preferences and confidence, the agents compute their preferred matching sets. The arguments in such preferred sets are viewed as the set of globally acceptable arguments. We show the applicability of our model matching two legal core ontologies: LKIF and CLO.

Keywords. Ontology matching, law ontologies, argumentation framework

Introduction

Law ontologies provide a formal description of the objects and their relations in the legal domain. Law information retrieval systems, such as question answering systems, use this knowledge to represent semantic objects, to associate them with law documents and to make inferences about them. Law ontologies covering different aspects of the law domain have been proposed in the literature.

Regarding the fact that overlapping ontologies cover complementary aspects of the law domain, a primary problem to solve in order to obtain interoperability between the systems is the ontology matching. The matching process takes two ontologies as input and determines as output correspondences between the semantically related entities of those ontologies. There are many different approaches to the matching problem. Whereas lexical approaches consider measures of lexical similarity; semantic ones consider semantic relations usually on the basis of semantic oriented linguistic resources. Other approaches consider term positions in the ontology hierarchy. Indeed, taxonomies of

\textsuperscript{1}Corresponding Author: Rua Romão Ramalho, 59, Évora, Portugal; E-mail: cassia@di.uevora.pt
the different matching approaches have been proposed in the literature, see for example [17][18] and [19]. However, the use of a single technique for a large variety of schemes is unlikely to be successful[5]. Since these approaches are complementary to each other their combination should lead to high matching accuracies than those provided by each one individually.

This paper presents a proposal based on argumentation to combine ontology matching approaches. We use the abstract argumentation framework[6] to combine matching approaches. In previous work [25] we extended a state of art argumentation framework, namely Value-based Argumentation Framework (VAF)[3], in order to represent arguments with confidence degrees. The VAF allows to determine which arguments are acceptable, with respect to the different audiences represented by different agents. We then associate to each argument a confidence degree, representing how confident an agent is in the similarity of two ontology terms. Our agents apply different matching approaches and cooperate in order to exchange their local results (arguments). Next, based on their preferences and confidence of the arguments, the agents compute their preferred matching sets. The arguments in such preferred sets are viewed as the set of globally acceptable arguments. The idea is not to provide an improved matching technique, but allowing all techniques to compete, as there are many different matching situations, a competing (complementary) approach is evaluated as a way to select the right strategy to apply in each situation.

We show the applicability of our model in a law match case using two legal core ontologies, LKIF and CLO.

This paper is structured as follows. Section 1 introduces the ontology matching approaches. Section 3 comments on argumentation framework. Section 4 presents our argumentation model. Section 5 presents the law matching case. Section 6 presents the main related work. Finally, section 7 presents the final remarks and future work.

1. Ontology Matching

Ontology matching is the process of finding relationships or correspondences between entities of different ontologies [8]. If two concepts correspond, they mean the same thing, or closely related things. The approaches for ontology matching vary from lexical (see [22][16]) to semantic and structural levels (see [10]). In the lexical level, metrics to compare string similarity are adopted. One well-known measure is the Levenshtein distance or edit distance [13], which is given by the minimum number of operations (insertion, deletion, or substitution of a single character) needed to transform one string into another. Other common metrics can be found in [16], [21], and [7].

The semantic level considers the semantic relations between concepts to measure the similarity between them, usually on the basis of semantic oriented linguistic resources. The well-known WordNet² database, a large repository of English semantically related items, has been used to provide these relations. This kind of matching is complementary to the pure string similarity metrics. It is common that string metrics yield high similarity between strings that represent completely different concepts (i.e, the words “score” and “store”). Moreover, semantic-structural approaches have been explored [4][10]. In this

²http://www.wordnet.princeton.edu
case, the positions of the terms in the ontology hierarchy are considered, i.e., terms more generals and terms more specifics are also considered as input to the matching process.

It is assumed that the approaches are complementary to each other and combining different ones reflect better solutions when compared to the solutions of the individual approaches. Heuristics to combine different approaches for ontology matching have been proposed in the literature (see, for example, [14], [5], [9]). Our proposal is to use argumentation to combine such approaches. Different approaches are encapsulated by agents that cooperate in order to exchange their local results (arguments). Based on their preferences and confidence of the arguments, the agents compute their preferred matching sets. The arguments in such preferred sets are viewed as the set of arguments globally acceptable (objectively or subjectively).

2. Argumentation Framework

Our argumentation model is based on the Value-based Argumentation Frameworks (VAF)[3], a development of the classical argument system of Dung [6]. First, we present the Dung’s framework, upon which the VAF rely. Next, we present the VAF and our extended framework.

2.1. Classical argumentation framework

Dung [6] defines an argumentation framework as follows.

Definition 2.1.1 An Argumentation Framework is a pair \( AF = (AR, \text{attacks}) \), where \( AR \) is a set of arguments and \( \text{attacks} \) is a binary relation on \( AR \), i.e., \( \text{attacks} \subseteq AR \times AR \). An attack \( (A,B) \) means that the argument \( A \) attacks the argument \( B \). A set of arguments \( S \) attacks an argument \( B \) if \( B \) is attacked by an argument in \( S \).

The key question about the framework is whether a given argument \( A, A \in AR \), should be accepted. One reasonable view is that an argument should be accepted only if every attack on it is rebutted by an accepted argument [6]. This notion produces the following definitions:

Definition 2.1.2 An argument \( A \in AR \) is acceptable with respect to set arguments \( S(\text{acceptable}(A,S)) \), if \( (\forall x)(x \in AR) \) & \( (\text{attacks}(x,A)) \rightarrow (\exists y)(y \in S) \) & attacks\((y,x)\)

Definition 2.1.3 A set \( S \) of arguments is conflict-free if \( \neg(\exists x)(\exists y)(x \in S) \& (y \in S) \) & attacks\((x,y)\)

Definition 2.1.4 A conflict-free set of arguments \( S \) is admissible if \( (\forall x)(x \in S) \rightarrow \text{acceptable}(x,S) \)

Definition 2.1.5 A set of arguments \( S \) is a preferred extension if it is a maximal (with respect to inclusion set) admissible set of \( AR \).
A preferred extension represent a consistent position within AF, which can defend itself against all attacks and which cannot be further extended without introducing a conflict.

The purpose of [3] in extending the AF is to allow associate arguments with the social values they advance. Then, the attack of one argument on another is evaluated to say whether or not it succeeds by comparing the strengths of the values advanced by the arguments concerned.

2.2. Value-based argumentation framework

In Dung’s frameworks, attacks always succeed. However, in many domains, including the one under consideration, arguments lack this coercive force: they provide reasons which may be more or less persuasive [12]. Moreover, their persuasiveness may vary according to their audience.

The VAF is able to distinguish attacks from successful attacks, those which defeat the attacked argument, with respect to an ordering on the values that are associated with the arguments. It allows to accommodate different audiences with different interests and preferences.

Definition 2.2.1 A Value-based Argumentation Framework (VAF) is a 5-tuple \( VAF = (AR, \text{attacks}, V, \text{val}, P) \) where \((AR, \text{attacks})\) is an argumentation framework, \(V\) is a nonempty set of values, \(\text{val}\) is a function which maps from elements of \(AR\) to elements of \(V\) and \(P\) is a set of possible audiences. For each \(A \in AR\), \(\text{val}(A) \in V\).

If \(V\) contains a single value, or no preference between the values has been defined, the AVAF becomes a standard AF. If each argument can map to a different value, a Preference Based Argumentation Framework is obtained [1].

Definition 2.2.2 An Audience-specific Value Based Argumentation Framework (AVAF) is a 5-tuple \(VAF_a = (AR, \text{attacks}, V, \text{val}, \text{valpref}_a)\) where \(AR, \text{attacks}, V\) and \(\text{val}\) are as for a VAF, \(a\) is an audience and \(\text{valpref}_a\) is a preference relation (transitive, irreflexive and asymmetric) \(\text{valpref}_a \subseteq V \times V\), reflecting the value preferences of audience \(a\). \(\text{valpref}(v_1, v_2)\) means \(v_1\) is preferred to \(v_2\).

Definition 2.2.3 An argument \(A \in AR\) defeats \(A_a\) (or successful attacks) an argument \(B \in AR\) for audience \(a\) if and only if both \(\text{attacks}(A, B)\) and not \(\text{valpref}(\text{val}(B), \text{val}(A))\).

An attack succeeds if both arguments relate to the same value, or if no preference value between the values has been defined.

Definition 2.2.4 An argument \(A \in AR\) is acceptable to audience \(a\) (\(\text{acceptable}_a\)) with respect to set of arguments \(S\), \(\text{acceptable}_a(A, S)\) if \((\forall x)(x \in AR \& \text{defeats}_a(x, A)) \rightarrow (\exists y)((y \in S) \& \text{defeats}_a(y, x)).\)

Definition 2.2.5 A set \(S\) of arguments is conflict-free for audience \(a\) if \((\forall x)(\forall y)((x \in S \& y \in S) \rightarrow (\neg\text{attacks}(x, y) \lor \text{valpref}(\text{val}(y), \text{val}(x)) \in \text{valpref}_a))\).

Definition 2.2.6 A conflict-free set of arguments \(S\) for audience \(a\) is admissible for an audience \(a\) if \((\forall x)(x \in S \rightarrow \text{acceptable}_a(x, S)).\)
Definition 2.2.7 A set of arguments $S$ in the VAF is a preferred extension for audience $\alpha$ (preferred$_\alpha$) if it is a maximal (with respect to set inclusion) admissible for audience $\alpha$ of AR.

In order to determine the preferred extension with respect to a value ordering promoted by distinct audiences, [3] introduces the notion of objective and subjective acceptability.

Definition 2.2.8 An argument $x \in AR$ is subjectively acceptable if and only if $x$ appears in the preferred extension for some specific audiences but not all. An argument $x \in AR$ is objectively acceptable if and only if, $x$ appears in the preferred extension for every specific audience. An argument which is neither objectively nor subjectively acceptable is said to be indefensible.

2.3. An extended value-based argumentation framework (E-VAF)

We extend the VAF in order to represent arguments with confidence degrees. Two elements have been added to the VAF: a set with confidence degrees and a function which maps from arguments to confidence degrees. The confidence value represents the confidence that an individual agent has in some argument. Any matching tools actually output mappings with a confidence that reflects the confidence degree they have in the similarity of the entities involved in the correspondence. These confidence degrees are usually derived from the similarity assessments made during the ontology matching process, e.g. from an edit distance measure between labels, or a overlap measure between instance sets. So, the confidence degrees is a criteria which must be considered when combining matching approaches.

Definition 2.3.1 An Extended Value-based Argumentation Framework (E-VAF) is a 7-tuple $E$-VAF = $(AR, attacks, V, val, P, C, valC)$ where $(AR, attacks, V, val, P)$ is a value-based argumentation framework, $C$ is a nonempty set of values representing the confidence degrees, $valC$ is a function which maps from elements of AR to elements of $C$. $valC \subseteq C \times C$ and $valC_{ pref}(c_1, c_2)$ means $c_1$ is preferred to $c_2$.

Definition 2.3.2 An argument $x \in AR$ defeats$_\alpha$ (or successful attacks) an argument $y \in AR$ for audience $\alpha$ if and only if $attacks(x,y) \land (valC_{ pref}(valC(x), valC(y)) \lor \neg valC_{ pref}(val(y), val(x)) \lor \neg valC_{ pref}(valC(y), valC(x)))$.

An attack succeeds if (a) the confidence degree of the attacking argument is greater than the confidence degree of the argument being attacked; or if (b) the argument being attacked does not have greater preference value than attacking argument (or if both arguments relate to the same preference values) and the confidence degree of the argument being attacked is not greater than the attacking argument.

Definition 2.3.3 A set $S$ of arguments is conflict-free for audience $\alpha$ if $(\forall x)(\forall y) ((x \in S \land y \in S) \implies \neg attacks(x, y) \lor \neg valC_{ pref}(valC(x), valC(y)) \land (valC_{ pref}(val(y), val(x)) \lor valC_{ pref}(valC(y), valC(x))))$. 
It is important to distinguish the difference between values and confidence. There are different types of agents representing different matching approaches. Each approach represent a value and each agent represents an audience, with preferences between the values. The values are used to determine the preference between the different agents. Moreover, each agent generates arguments with a confidence, based on the confidence returned by the matching technique. So, we extended the VAF in order to define a new notion of argument acceptability which combines values (related with the agent’s preference) and confidence (confidence degree of an argument). If our criterion was based only on the confidence of the arguments, a Preference Based Argumentation Framework could be used [1].

3. E-VAF for Ontology Matching

In our model, dedicated agents encapsulate different matching approaches. In this paper we consider three values: lexical (L), semantic (S), and structural (E) (i.e. \( V = \{L, S, E\} \), where \( V \in \text{E-VAF} \)). These values represent the matching approach used by the agent. Each audience has an ordering preference between the values. For instance, the lexical agent represents an audience where the value L is preferred to the values S and E. Our idea is not to have an individual audience with preference between the agents (i.e., semantic agent is preferred to the other agents), but it is to try accommodate different audiences (agents) and their preferences. So, using only the strengths is not sufficient to the problem. The idea is to obtain a consensus when using different matching techniques, which are represented by different preference between values.

3.1. Argumentation generation

First, the agents work in an independent manner, applying the matching approaches and generating matching sets. The matching result will consist of a set of all possible correspondences between terms of two ontologies. A matching \( m \) can be described as a 5-tuple \( m = (t_1, t_2, h, R, c) \), where \( t_1 \) corresponds to a term in the ontology 1, \( t_2 \) corresponds to a term in the ontology 2, \( h \) is one of \(-, +\) depending on whether the matching does or does not hold, \( R \) is the mapping relation resulting from the matching between these two terms, and \( c \in C \) is the confidence degree associated to that matching (certainty or uncertainty, as it will be commented below). In an initial setting, the agents are able to return equivalence value to \( R \). Each mapping \( m \) is encapsulated in an argument. The arguments can be defined as follows:

**Definition 4.1** An argument \( \in AR \) is a 2-tuple \( x = (m, a) \), where \( m \) is a mapping; \( a \in V \) is the value of the argument, depending of the agent generating that argument (i.e, lexical, semantic or structural);

The confidence degree is defined by the agent when applying the specific matching approach. Here, we assumed \( C = \{\text{certainty, uncertainty}\} \), where \( C \in \text{E-VAF} \). Table 1 shows the possible values to \( h \) and \( c \), according to the agent’s value. The agents generate their arguments based on rules from Table 1.
Table 1. \( h \) and \( c \) to values.

<table>
<thead>
<tr>
<th>( h )</th>
<th>( c )</th>
<th>Lexical</th>
<th>Semantic</th>
</tr>
</thead>
<tbody>
<tr>
<td>+ certainty</td>
<td>1</td>
<td>synonym</td>
<td></td>
</tr>
<tr>
<td>+ uncertainty</td>
<td>( 1 &gt; r &gt; t )</td>
<td>related</td>
<td></td>
</tr>
<tr>
<td>- certainty</td>
<td>( 0 &lt; r \leq t )</td>
<td>unknown</td>
<td></td>
</tr>
<tr>
<td>- uncertainty</td>
<td>0</td>
<td>unknown</td>
<td></td>
</tr>
</tbody>
</table>

3.1.1. Lexical agent

The output of lexical agents \( (r) \) is a value from the interval \([0,1]\), where 1 indicates high similarity between two terms. This way, if the output is 1, the lexical agent generates an argument \( x = (m,L) \), where \( m = (t_1,t_2,+,equivalence,\text{certainty}) \). If the output is 0, the agent generates an argument \( x = (m,L) \), where \( m = (t_1,t_2,-,equivalence,\text{certainty}) \). A threshold \( (t) \) is used to classify the output in uncertain categories. The threshold value can be specified by the user.

3.1.2. Semantic agent

The semantic agents consider semantic relations between terms, such as synonym, antonym, holonym, meronym, hyponym, and hypernym (i.e., such as in WordNet database). When the terms being mapped are synonymous, the agent generates an argument \( x = (m,S) \), where \( m = (t_1,t_2,+,equivalence,\text{certainty}) \). The terms related by holonym, meronym, hyponym, or hypernym are considered related and an argument \( x = (m,S) \) is generated, where \( m = (t_1,t_2,+,equivalence,\text{uncertainty}) \) (the terms have some degree of equivalence); when the terms can not be related by the WordNet (the terms are unknown for the WordNet database), an argument \( x = (m,S) \), where \( m = (t_1,t_2,-,\text{uncertainty},equivalence) \), is then generated.

3.1.3. Structural agent

The structural agents consider the super-classes (or sub-classes) intuition to verify if the terms can be mapped. First, it is verified if the super-classes of the compared terms are lexically similar. If not, the semantic similarity between they is used. If the super-classes of the terms are lexically or semantically similar, the terms are considered equivalent to each other. The argument is generated according to the lexical or semantic comparison. For instance, if the super-classes of the terms are not lexically similar, but they are synonymous (semantic similarity), an argument \( x = (m,E) \), where \( m = (t_1,t_2,+,equivalence,\text{uncertainty}) \), is generated. If the structural agent finds similarity between the super-classes of the compared terms, it is because they can be mapped, but it does not mean that the terms have lexical or semantic similarity, then the confidence for the mapping is uncertainty. For instance, for the terms “Publication/Topic” and “Publication/Proceedings”, the structural agent indicates that the terms can be mapped because they have the same super-class, but not with certainty because it is not able to indicate that the terms are equivalent at all.

Moreover, it is pointed out that the semantic agent does not explore any kind of hierarchial propriety, as done by the structural agent. The semantic agent is based on the analysis of synsets and it does not use the structural information available on WordNet.
After generating their set of arguments, the agents exchange with each other their arguments. When all agents have received the set of arguments of the each other, they generate their attacks set. An attack (or counter-argument) will arise when we have arguments for the mapping between the same terms, but with conflicting values of $h$. For instance, an argument $x = (m_1, L)$, where $m_1 = (t_1, t_2, +, certainty, equivalence)$, have as an attack an argument $y = (m_2, E)$, where $m_2 = (t_1, t_2, -, uncertainty, equivalence)$. $m_1$ and $m_2$ refer to the same terms in the ontologies. The argument $y$ also represents an attack to the argument $x$.

As an example, consider the mapping between terms “Abstract” (from generic core ontology SUMO - Figure 1) and “Abstract-Entity” (from the LKIF ontology - Figure 2), and the lexical and structural agents.

The lexical agent generates an argument $x = (m, L)$, where $m = (Abstract, Abstract-Entity, +, certainty, equivalence)$; and the structural agent generates an argument $y = (m, E)$, where $m = (Abstract, Abstract-Entity, -, uncertainty, equivalence)$. For both lexical and structural agents, the set of arguments is $AR = \{x, y\}$ and the attacks $= \{(x, y), (y, x)\}$. However, the relations of successful attacks will be defined according to specific audience (see Definition 2.3.2). The argument $x$ successfully attacks the argument $y$, because $x$ has greater confidence than $y$.

When the set of arguments and attacks have been produced, the agents need to define which of them must be accepted. To do this, the agents compute their preferred extension, according to the audiences and confidence degrees. A set of arguments is globally subjectively acceptable if each element appears in the preferred extension for some agent. A set of arguments is globally objectively acceptable if each element appears in
the preferred extension for every agent. The arguments which are neither objectively nor subjectively acceptable are considered indefensible.

In the example above, considering the lexical (L) and structural (E) audiences, where L ≻ E and E ≻ L, respectively, for the lexical audience, the argument y successful attacks the argument x, while the argument x does not successful attack the argument y for the structural audience. Then, the preferred extension of both lexical and structural agents is composed by the argument y, which can be seen as globally objectively acceptable.

4. Experiments and Evaluation

Let us consider that three agents need to obtain a consensus about mappings that link corresponding class names in two different ontologies. We had used two legal core ontologies: LKIF and CLO. Table 2 shows the number of classes and attributes of the two ontologies.

<table>
<thead>
<tr>
<th>Ontology</th>
<th>Classes</th>
<th>Attributes</th>
</tr>
</thead>
<tbody>
<tr>
<td>LKIF-Core</td>
<td>154</td>
<td>84</td>
</tr>
<tr>
<td>CLO</td>
<td>107</td>
<td>69</td>
</tr>
</tbody>
</table>

Three agents were considered: lexical (L), semantic (S), and structural (E). The agents were implemented in Java 5.0, and the experiments ran on Pentium(R) 4, UCP 3.20GHz, 512MB.

The lexical agent was implemented using the edit distance measure (Levenshtein measure). We used the algorithm available in the API for ontology alignment (INRIA)\(^3\) (EditDistNameAlignment). The semantic agent has used the JWordNet API\(^4\), which is an interface to the WordNet database. For each WordNet synset, we retrieved the synonymous terms and considered the hypernym, hyponym, member-holonym, member-meronym, part-holonym, and part-meronym as related terms. The structural agent was based on super-classes similarity.

The threshold used to classify the matcher agents output was 0.8. This value was defined based on previous analysis of the edit distance values between the terms of the ontologies used in the experiments. The terms with edit distance values greater than 0.8 have presented lexical similarity.

The evaluation of law ontology matching lacks well established benchmarks. Therefore our choices on evaluation were based on the manual analysis of the positive mapping (h = +) returned by our model, when using the ontologies described above. So, we compute the precision of the automatic positive mappings. Table 3 shows the number of correct and incorrect mappings, together the precision, when considering the mappings with certainty and the mappings with uncertainty.

As shown in Table 3, the results are better when considering only the mapping with certainty. This setting is specially useful when the matching system is used without user feedback (i.e., communication between agents where the mappings are computed on the fly). In the cases where the systems is used to help users in a manual matching

---

\(^3\)http://alignapi.gforce.inria.fr

\(^4\)http://jwn.sourceforge.net (using WordNet 2.1)
### Table 3. Matching results.

<table>
<thead>
<tr>
<th>Ontology entity</th>
<th>Mapping</th>
<th>Total</th>
<th>Correct</th>
<th>Incorrect</th>
<th>Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>Attribute</td>
<td>certainty</td>
<td>9</td>
<td>7</td>
<td>2</td>
<td>0.77</td>
</tr>
<tr>
<td></td>
<td>uncertainty</td>
<td>245</td>
<td>29</td>
<td>216</td>
<td>0.11</td>
</tr>
<tr>
<td>Class</td>
<td>certainty</td>
<td>20</td>
<td>14</td>
<td>6</td>
<td>0.70</td>
</tr>
<tr>
<td></td>
<td>uncertainty</td>
<td>929</td>
<td>59</td>
<td>870</td>
<td>0.06</td>
</tr>
</tbody>
</table>

process, could be interesting to retrieval a larger set of mapping, i.e., considering also the mappings with uncertainty.

Although further evaluation is needed, the values of precision are promising, for both mappings with classes and instances.

## 5. Related Work

In the field of ontology argumentation few approaches are being proposed. Basically, the closer proposal is from [12][11], where an argument framework is used to deal with arguments that support or oppose candidate correspondences between ontologies. The candidate mappings are obtained from an Ontology Mapping Repository (OMR) – the focus is not how the mappings are computed – and argumentation is used to accommodate different agent’s preferences. In our approach mappings are computed by the specialized agents described in this paper, and argumentation is used to solve conflicts between the individual results.

We find similar proposals in the field of ontology negotiation. [23] presents an ontology to serve as the basis for agent negotiation, the ontology itself is not the object being negotiated. A similar approach is proposed by [26], where agents agree on a common ontology in a decentralized way. Rather than being the goal of each agent, the ontology mapping is a common goal for every agent in the system. [2] presents an ontology negotiation model which aims to arrive at a common ontology which the agents can use in their particular interaction. We, on the other hand, are concerned with delivering mapping pairs found by a group of agents using argumentation. [20] describes an approach for ontology mapping negotiation, where the mapping is composed by a set of semantic bridges and their inter-relations, as proposed in [15]. The agents are able to achieve a consensus about the mapping through the evaluation of a confidence value that is obtained by utility functions. According to the confidence value the mapping rule is accepted, rejected or negotiated. Differently from [20], we do not use utility functions. Our model is based on cooperation and argumentation, where the agents change their arguments and by argumentation they select the preferred mapping.

## 6. Final Remarks and Future Work

This paper presented a composite matching approach based on the argumentation formalism to map legal core ontologies. The matching process, takes two ontologies as input and determines as output correspondences between the semantically related entities of those ontologies. It can help users to reuse and compare information from different sources.
We extended a state-of-art argumentation framework, namely Value-based Argumentation Framework (VAF), in order to represent arguments with confidence degrees. The VAF allows to determine which arguments are acceptable, with respect to the different preferences represented by different agents. Our extension associates to each argument a confidence degree, representing the confidence that a specific agent has in that argument. We assumed that the confidence degrees is a criteria which is necessary to represent the ontology matching domain.

We have used different agents’ output which use distinct matching algorithms in order to verify the behavior of our model. The model was applied to legal core ontologies: LKIF and CLO. However, our approach is not restrict to legal domain. The proposed argumentation model seems to be useful for general ontology matching (see, for example [25][24], where we applied our model for other domains).

In the future, we intend to develop further tests considering also agents using constraint-based matching approaches (i.e., the similarity between two terms can be based on the equivalence of data types and domains, of key characteristics, or relationship cardinality); use the ontology’s application context in our matching approach (i.e., how the ontology entities are used in some external context, which is especially interesting, for instance, to identify WordNet senses that must be considered to specific terms); and test our approach for less high-level ontologies. Moreover, we plan to extend our model to multilingual ontology matching. Next, we will use the matching result as input to an ontology merge process in a question answering system for the law domain.

Acknowledgements

The first author is supported by the Programme Alban, the European Union Programme of High Level Scholarships for Latin America, scholarship no.E05D059374BR.

References


Copyright Licenses Reasoning using an OWL-DL Ontology

Roberto GARCÍA¹, Rosa GIL
Universitat de Lleida, Spain

Abstract. In order to extract the full potential from Internet-wide content sharing and reuse, the underlying copyright issues must be taken into account. The novel requirements are not satisfied by traditional Digital Rights Management. Open licensing initiatives seem more appropriate, but they lack the required computerised support. Our proposal facilitates interoperability while providing a rich framework that accommodates copyright law and copes with custom licensing schemes. It is based on the Description Logic variant of the Web Ontology Language (OWL-DL) and constitutes an ontology that conceptualises the copyright domain. The ontology provides the building blocks for flexible machine-understandable licenses and facilitates implementation because DL reasoners can be directly used for license checking. However, some preliminary transformations of the licenses models are required in order to overcome the Open World Assumption inherent in OWL-DL, which limits DL-based license reasoning.

Keywords. Digital Rights Management, Copyright, Licensing, Description Logic, Reasoning

Introduction

Traditionally, copyright management has been achieved through Digital Rights Management (DRM) systems. For instance, they have been used by record companies to protect music sold on the Internet and in enterprises in order to control content access.

DRM focuses on controlling content access, the last step in the copyright value chain, and pays little attention to the previous ones: creation, derivation, recording, communication, etc. This is enough in closed domains, like enterprise DRM or vertical content distribution channels.

However, traditional DRM is showing its limitations in Internet-wide scenarios or when it must accommodate new copyright schemes like open source or open access. For instance, a key scenario with these requirements is inter-organisational scientific and technological knowledge sharing and reuse among universities, research centres, etc.

On the other hand, there are open licensing initiatives, like Creative Commons, which show really promising results. However, they lack the required computerised support and flexibility to scale to Internet-wide copyright management.

Our proposal facilitates interoperability and automation, while providing a rich framework that accommodates copyright law and custom licensing schemes. It is based

¹Corresponding Author: Jaume II, 69; E-25001 Lleida, Spain; E-mail: rgarcia@diei.udl.cat
on a copyright ontology, which is implemented using the Description Logic variant of
the Web Ontology Language. This approach facilitates implementation because
existing Semantic Web tools can be easily reused.

The rest of this paper is organised as follows. First, the next subsection explores
existing initiatives and their limitations are presented, from classical and standard
DRM to open access proposals like Creative Commons. Then, the Semantic Web
approach to copyright-aware DRM is presented in Section 1, which is materialised in
the Copyright Ontology detailed in Section 2 and implemented using Semantic Web
tools as it is shown in Section 3. Finally, Section 4 presents the conclusions and the
future work.

Related Work

The DRM Watch review on DRM standards [1] shows that interoperability is a key
issue for DRM systems. It arises in the content distribution scenario, for instance when
a user wants to consume content in any of the devices he owns, or in the organisational
DRM scenario, when content flows through organisations or external content is used in
order to derive new one.

The main response to DRM interoperability requirements has been the settlement
of many standardisation efforts. One of the main ones is ISO/IEC MPEG-21 [2], whose
main interoperability facilitation component is the Rights Expression Language (REL)
[3].

The REL is a XML schema that defines the grammar of a license building
language, so it is based on a syntax formalisation approach. There is also the MPEG-21
Rights Data Dictionary (RDD) that captures the semantics of the terms employed in the
REL, but it does so without defining formal semantics [4].

This syntax-based approach is also common to other DRM interoperability efforts
and one of main causes of the lack of production implementations also observed in the
DRM Watch review. Despite the great efforts in place, the complexity of the DRM
domain makes it very difficult to produce and maintain implementations based on this
approach.

The implementers must build them from specifications that just formalise the
grammar of the language and force the interpretation and manual implementation of the
underlying semantics. This has been feasible for less complex domains, for instance
when implementing a MPEG-4 player from the corresponding specification. However,
this is hardly affordable for a more complex and open domain like copyright, which
also requires a great degree of flexibility.

Moreover, the limited expressivity of the technical solutions currently employed
makes it very difficult to accommodate copyright law into DRM systems.
Consequently, DRM standards follow the traditional access control approach. They
concentrate their efforts in the last copyright value chain step, content consumption,
and provide limited support for the other steps.

In fact, just Internet publishing risks are considered and the response is to look for
more restrictive and secure mechanism to avoid access control circumvention. This
makes DRM even less flexible because it ties implementations to proprietary and
closed hardware and software security mechanisms.

The limited support for copyright law is also a concern for users and has been
criticised, for instance, by the Electronic Frontier Foundation [5]. The consequence of
this lack is basically that DRM systems fail to accommodate rights reserved to the public under national copyright regimes.

Consequently, the DRM world remains apart from the underlying copyright legal framework. As it has been noted, this is a risk because DRM systems might then incur then into confusing legal situations. Moreover, it is also a lost opportunity because, from our point of view, ignoring copyright law is also ignoring a mechanism to achieve interoperability.

It is true that copyright law diverges depending on local regimes but, as the World Intellectual Property Organisation’s promotes, there is a common legal base and fruitful efforts towards a greater level of copyright law worldwide harmonisation.

A new approach is necessary if we want to take profit from the Internet as a content sharing medium. The existence of this opportunity is clear when we observe the success of the Creative Commons initiative [6], whose objective is to promote content sharing and reuse thorough innovative copyright and licensing schemes.

However, despite the success of Creative Commons licenses, who estimates more than 140 millions of works licensed under its terms, this initiative is not seen as an alternative to DRM. The main reason is the lack of flexibility of the available licensing terms. There are mainly six different Creative Commons licenses, all of them non-commercial, and no mechanism for easy extension and adoption of alternative licensing schemes.

Moreover, Creative Commons licenses are available in three formats: a legal version for lawyers, a more readable version for average users and as metadata for computers consumption. However, the Creative Commons metadata is not a formal representation of the licenses; it just provides a reduced set of terms for building computer-oriented licenses. There are three kinds of permissions (reproduction, distribution and derivative works), one prohibition (commercial use) and four requirements (attribution, notice, share alike and source code).

Consequently, although it is possible to provide computer support for simple services like content search, there are no mechanisms for customisation and advanced computerised support that enable an Internet-wide copyright-based alternative to DRM systems.

To conclude the related work overview, the closest initiatives in the Semantic Web field are the generic policy languages KAoS [7] and Rei [8]. KAoS is based on OWL and it is able to reason about policies by ontological subsumption. However, it requires some OWL-Full reasoning capabilities and its implementation is based on a theorem prover, which causes serious scalability problems. On the other hand, Rei is based on rules that overcome OWL expressivity limitations. However, this prevents it from exploiting the full potential of the OWL language. In fact, Rei rules knowledge is treated separately from OWL ontology knowledge due to its different syntactical form.

1. A Semantic Web Approach to DRM

Our proposal tries to solve the limitations observed in the current DRM and Creative Commons approaches. The underlying reason for all of them is the lack of technological tools that allow building a flexible and expressive representation framework.

Such framework must deal with the underlying legal framework and, simultaneously, be automated in order to benefit from computerised support. This would make possible to extract all the potential from Internet-wide knowledge sharing and reuse with the support of accurate copyright management mechanisms.

The first objective is to overcome the limitations of purely syntactic approaches, like XML, and their lack of formal semantics. The best way to formalise semantics is to use ontologies in order to build an expressive and flexible computer-supported copyright management.

Moreover, as we want a Web-wide scope, the best choice is to use an ontology language based on Web technologies. The clear choice is Semantic Web ontologies based on the OWL standard [9], which provides a set of primitives that make possible to build web-sharable conceptualisations.

The increased expressivity of web ontologies allows us to include the underlying legal framework into the formalisation and to build the rest of the system on top of it. This is a key issue because, in order to build a generic framework that facilitates interoperability, the focus must be placed on the underlying legal, commercial and technical copyright aspects.

This is the approach for the Copyright Ontology, detailed in the following section. The expressiveness and generality of the resulting conceptualisations allows coping with the shortcomings of existing approaches and, additionally, the ontology can be used as an interoperability facilitator for the main DRM standards [10].

The ontology is implemented as an OWL Web ontology based on the Description Logic (DL) variant, OWL-DL. This implementation facilitates DRM systems development as license checking is implemented using existing Semantic Web reasoners.

To the best of our knowledge, there is just one other ontological framework for DRM, OntologyX. However, it is a commercial product for which there is little publicly available information. In any case, from the available information, it is clear that OntologyX concentrates on the kind of actions that can be performed on governed content and it does not take into account the underlying legal framework. Moreover, it currently lacks formal semantics and can be seen more like a rights dictionary than as a fully-fledged ontology.

2. The Copyright Ontology

The copyright domain is quite complex so we face its conceptualisation in three phases. Each phase concentrates on a part of the whole domain. First, the objective is the more primitive part, the Creation Model. Second, there is the model for the rights part, the Rights Model, and finally a model for the available actions, the Action Model, which is built on top of the two previous ones.

2.1. Creation Model

The Creation Model conceptualises the different forms a creation can take, which are classified depending on the classical top ontological points of view [11,12,13]:

---

1 Copyright Ontology, http://rhizomik.net/ontologies/copyrightonto
— **Abstract**: something that cannot exist at a particular place and time without some physical encoding or embodiment.

- **Work**: is a distinct intellectual or artistic creation. It includes literary and artistic works, music, pictures and motion pictures, but also computer programs or compilations, like databases.

— **Object**: corresponds to the class of ordinary objects and includes digital objects.

- **Manifestation**: the materialisation of a work in a concrete medium, a tangible or digital object.
- **Fixation**: the materialisation of a performance in a concrete medium, a tangible or digital object.
- **Instance**: the reproduction, copy, of a manifestation, a fixation or another instance.

— **Process**: something that happens and has temporal parts or stages.

- **Performance**: the expression in time of a work. Performers or technical methods might be involved in the process.
- **Communication**: the transmission of a work among places at a given time. It is a process performed when the public is not present at the place and or time where the communication originates. It includes broadcasts, i.e. one to many, but also communications from a place and at a time individually chosen.

There are many relations among the different forms a creation can take during its life cycle, see Figure 1, as it evolves from an abstract idea, i.e. a Work, towards something that can be consumed by end users, e.g. and **Instance** or a **Communication** or **Performance**.

![Figure 1. Creation model showing the different views on creation](image)

This relations are named following the same pattern, e.g. for a **Manifestation** there is the relation **isManifestationOf**, which relates it to the original **Work** that it materialises, and the reverse relation **hasManifestation**, that relates a **Work** to all its manifestations.

**2.2. Rights Model**

The Rights Model follows the World Intellectual Property Organisation recommendations. It includes economic plus moral rights, as promoted by WIPO, and copyright related rights, see Figure 2.
The most relevant rights in the context of DRM systems are economic rights as they are related to the production and commercial aspects of copyright. Reproduction, Distribution, Public Performance, Fixation, Communication and Transformation Right are the economic rights. The Rights Model in the Copyright Ontology provides the following hierarchy of copyright rights:

- **Economic Rights**:
  - Reproduction Right
  - Distribution Right
  - Public Performance Right
  - Fixation Rights (Sound Record and Motion Picture Rights)
  - Communication Rights (Broadcasting and Making Available Rights)
  - Transformation Rights (Adaptation and Translation Rights)

There are also the moral rights, which are always held by the creator and cannot be commercially exploited, and the related or neighbouring rights, the rights of the other actors also involved in the exploitation of works: performers, producers and broadcasters.

- **Moral Rights**:
  - Attribution Right
  - Integrity Right
  - Disclosure Right
  - Withdrawal Right

- **Related Rights**:
  - Performers Rights
  - Phonograms Producers Rights
  - Broadcasters Rights

### 2.3. Action Model

The last model, the Action Model, corresponds to the primitive actions that can be performed on the concepts defined in the Creation Model, as it is shown in Figure 3. Actions are regulated by the rights in the Rights Model. For the economic rights, these are the governed actions:

- **Reproduction Right**: *reproduce*, commonly speaking *copy*.
- **Distribution Right**: *distribute*. More specifically *sell*, *rent* and *lend*.
- **Public Performance Right**: *perform*; it is regulated when it is a public performance and not a private one.
- **Fixation Right**: fix, or record.
- **Communication Right**: communicate when the subject is an object or retransmit when communicating a performance or previous communication, e.g. a re-broadcast. Other related actions, which depend on the intended audience, are broadcast or make available.
- **Transformation Right**: derive. Some specialisations are adapt or translate.

![Diagram of relations between Action and Creation Models](image)

One of the biggest criticisms against DRM is that they do not respect some special permissions that many copyright legal systems provide to end-users. These permissions are commonly called fair use, fair dealing or user rights. Although some of them are referred to as rights, e.g. the right to quote, they constitute exceptions to copyright and should be considered as end-user privileges and not rights.

These privileged actions, normally restricted by copyright, may be done without the authorization of the copyright owner in circumstances specified in the law. Moreover, these exceptions do not mean that the exceptional use is always free. Some require the user to pay a compensation. For instance, in some countries, there are levies on digital recording equipment and media.

These are the main copyright exceptions:
- **Quotation Right**: quote, a limited extent reproduce action of a source protected work, which is clearly mentioned.
- **Education Right**: educational act, any reproduce, communicate or perform action with educational or research purposes.
- **Information Right**: inform, any copyright governed act with informative purposes.
- **Official Act Right**: official act, any copyright governed act that is part of an official act.
- **Private Copy Right**: reproduce privately, a reproduce act that produces a reproduction solely for private consumption.
- **Parody Right**: parody, any copyright governed act with parody or caricature purposes.
- **Temporary Reproduction Right**: reproduce temporally, a reproduce act that produces a temporal reproduction.
The action concepts are complemented with a set of relations that link them to the action participants. This set is adopted from the linguistics field and it is based on case roles [14]. The case roles in the Action Model are shown in Table 1.

<table>
<thead>
<tr>
<th>Role</th>
<th>initiator</th>
<th>resource</th>
<th>goal</th>
<th>essence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Action</td>
<td>agent, effector</td>
<td>instrument</td>
<td>result, recipient</td>
<td>patient, theme</td>
</tr>
<tr>
<td>Process</td>
<td>agent, origin</td>
<td>matter</td>
<td>result, recipient</td>
<td>patient, theme</td>
</tr>
<tr>
<td>Transfer</td>
<td>agent, origin</td>
<td>instrument, medium</td>
<td>experiencer, recipient</td>
<td>theme</td>
</tr>
<tr>
<td>Spatial</td>
<td>origin</td>
<td>path</td>
<td>destination</td>
<td>location</td>
</tr>
<tr>
<td>Temporal</td>
<td>start</td>
<td>duration</td>
<td>completion</td>
<td>pointInTime</td>
</tr>
<tr>
<td>Ambient</td>
<td>reason</td>
<td>manner</td>
<td>aim, consequence</td>
<td>condition</td>
</tr>
</tbody>
</table>

The previously introduced pool of primitive actions and case roles allows building models for events and value chains in the copyright domain. For instance, Figure 4 shows how we can build a model for the value chain of serials adapted from literary works.

First, the creator adapts the original literary work, e.g. Alexandre Dumas’ “The Count of Monte Cristo”, in order to produce a serial. The resulting adaptation is realised as a script that is performed by some actors, e.g. Gerard Depardieu, and recorded into a motion picture. This motion picture is finally broadcasted to users who can tune the resulting communication.

This is just the skeleton of the value chain. In order to give a more detailed model, each step in the value chain should be modelled as an event for the corresponding action and associated participants through case roles.

However, the objective is not just to model the actual events that capture the life cycle of a given creation. Prior to these events, licenses among the involved parties are established in order to govern the value flux. Consequently, the ontology must be enriched with permissions, prohibitions and obligations [15].

---

**Figure 4.** Literary works adapted to serials value chain.
2.4. Modelling Copyright Licenses

Copyright provides a legal framework that governs creations life cycles and tries to assure a fair compensation for all the involved parties, from authors to consumers. Copyright licenses are built on top of this legal framework and establish the terms for concrete interaction among these parties.

Licenses should capture the obligations, permissions and prohibitions that make sense in the copyright domain. The semantics of the license terms are captured by the ontology described so far, but it lacks the terms that capture the semantics of obligations, permissions and prohibitions.

In order to produce a homogeneous and usable conceptualisation, we have incorporated this terms in the ontology using the concepts that capture the semantics of obligations, permissions and prohibitions as they appear in contracts from a natural language point of view, i.e. using the corresponding actions and case roles (e.g. the verb to agree on a specific action as the natural way to model a permission).

The additions are detailed next and they are related to a generic contract modelling language, the Business Contract Language (BCL) [16], in order to illustrate how these additions make the Copyright Ontology a copyright contracts modelling tool. In the following subsections, each BCL building block is introduced and then related to its Copyright Ontology counterpart.

2.4.1. Roles

The simplest building blocks in BCL are roles, e.g. Purchaser, which are captured in a generic way by the Copyright Ontology case roles. For instance, there is not a specific Purchaser case role but it is implicit in the agent case role when applied to a Purchase action.

2.4.2. Event Patterns

BCL uses event patterns as the way to state what is obliged, permitted or prohibited by a contract; they are referenced from policies that establish their modality. They are also naturally captured by the ontology terms described so far. The proposed actions and case roles are used to model event patterns in the copyright domain.

For instance, Figure 5 shows a pattern for all copy events in a Peer to Peer network performed by agent “granted” who copies “content01” from “PeerA” to two peers from the set “PeerB, PeerC, PeerD” at any time point six months after “2006-01-01”.

![Figure 5. Pattern for a copy action in a P2P scenario](image-url)
2.4.3. Modalities

Then, there are the terms to state the modality of these event patterns in copyright contracts. BCL defines explicitly the modalities using the Obligation, Permission and Prohibition terms. The Copyright Ontology does the same but in an implicit way, following the same “action plus case roles” approach used for event patterns. Additional classes and relations are added in order to attach modalities to event patterns. The objective is to state that the set of actions corresponding to the pattern is permitted, obliged or prohibited, depending on the particular construct that is attached to the pattern, as it is detailed in the next subsections.

2.4.3.1. Permissions

BCL Permissions are captured by a new action class, Agree, and the permitted pattern is linked using the theme case role, whose semantics are to point to the object of an action.

Following with the example in Figure 5, in order to authorise the pattern that it models, an agreement like the one shown in Figure 6 can be modelled. The agreement between “granter” and “granted” in the upper part of authorises the pattern pointed by the theme case role, the previous P2P copy pattern at the centre of the figure.

2.4.3.2. Obligations

BCL Obligations are captured in the copyright contracts as event patterns that must be satisfied at some time point after the event pattern that triggers the obligation is exercised. They are modelled using the consequence case role that links the triggering pattern to the one that is obliged.

For instance, in the bottom part of Figure 6 it is stated that, if the copy action is exercised, the consequence is that the “granted” agent must transfer 3 Euros to the “granter” agent before 24 hours from the copy action.

![Diagram](image-url)

**Figure 6.** Agreement that permits the P2P copy pattern whose consequence is an economic obligation
2.4.3.3. Prohibitions
BCL Prohibitions are captured by another action, Disagree. Like for the Agree action, the theme case role is used to link it to the object of the action, in this case to the pattern that is prohibited.

For instance, in the previous scenario, the contract might also state that it is forbidden that the “granted” agent changes “content01” using a Disagree pattern with the corresponding Transform action pattern as its theme.

2.4.4. Guards
BCL Guards are patterns that must be satisfied in order to activate the evaluation of another event pattern, thus acting as a precondition. The condition case role is used to model guards. It is applied to the pattern that is guarded and it links to the pattern that establishes the precondition. The approach is similar to the obligation case captured by the consequence case role but, in this case, the condition case role establishes an a priori condition.

For instance, in the P2P scenario the Copy pattern might by guarded by a Transfer one that requires that the “granted” agent makes a 1 Euro prepayment to the “grantor” agent before the former can excise the permitted P2P Copy action.

3. OWL Implementation

The previous conceptualisation is just an abstraction of the copyright domain. An implementation is required if we want to use it to build a computerised copyright management system. The Semantic Web approach is also productive in this respect because existing tools can be used to make the implementation quite straightforward.

The ontology has been implemented using the DL variant of the Web Ontology Language (OWL-DL), which is constrained in order to be managed by Description Logic (DL) reasoners. Such reasoners guarantee that OWL-DL ontologies can be put into practice, i.e. reasoned over, in a decidable and tractable way.

Existing DL reasoners are used to automatically check if actions on copyrighted content are authorised or not. As it has been shown, licenses are composed of Agree or Disagree actions, linked through a theme relation to patterns of actions that are correspondingly authorised or forbidden.

The pattern is implemented as an OWL class made up from the combination of classes for actions, e.g. Copy or Access, and a set of OWL Restrictions. Each restriction defines a constraint on how members of the class, the domain, are related through the specified property to other ones, the range class. The available restrictions are:

- allValuesFrom: all the values for the range of the restricted property must pertain to the given class. For instance, all values of the agent relation must pertain to the Publisher Subscribers class or, for the pointInTime relation, to the time range [2007/01/01–2007/06/30]. In order to support the later, custom datatypes reasoning is required [17].

- someValuesFrom: there is at least one value that pertains to the given range class.

- hasValue: the range is limited to a specific individual, not a class of them. For instance, the theme of a Copy action must be the individual “doi:10.1032/…”.

- cardinality: this restriction limits the number of individuals that can be connected through the restricted property. A maximum, minimum or exact
cardinality can be defined. For instance, the recipients of an action can be limited to just two individuals.

Restrictions are combined using the intersection, union and complement logical operators in order to compose the patterns of actions. For instance, Figure 7 shows the conceptual model for a license that combines commercial and open access terms.

![Diagram](image_url)

**Figure 7.** Agreement on a copy action under commercial and open access terms

The upper part shows an Agree that permits two Copy patterns, connected through the theme relation. The one of the left grants “Publisher Subscribers” to copy some content identified by a DOI at any time point six months after 2007-01-01. Any attempt to exercise this action pattern is subject to a commercial condition, a compensation of 3€. On the other hand, the Copy pattern on the right grants anyone to copy the same content, once the period of six months is surpassed, if the aim is non-commercial.

The constraints on the kinds of actions, their agents, time points, etc. are then implemented using OWL Restrictions, which are combined using the logical operators in the OWL language. Figure 8 shows the pattern build up from the combination of such kind of restrictions for the example presented in Figure 7. For the set of all copy actions on “doi:10.1032/...”, the light grey area, two subsets are selected and their union constitutes the licensed actions pattern, the dark grey areas.

As it can be seen in Figure 8, each intersected restriction reduces the set of actions. For instance, the non-commercial pattern does not include any restriction on the agent of the action. Consequently, the licensed actions set includes any non-commercial copy action performed by anyone later than 2007-07-01. Table 2 shows the DL notation for the class definition that models the commercial copy pattern.

**Table 2.** OWL-DL Class for the commercial copy action pattern

<table>
<thead>
<tr>
<th>Pattern</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Copy (\cap)</td>
<td>(1)</td>
</tr>
<tr>
<td>(\forall \text{pointInTime} \geq 2007-01-01T00:00:00, \leq 2007-06-30T23:59:59 \cap)</td>
<td>(2)</td>
</tr>
<tr>
<td>(\forall \text{agent}\cdot\text{PublisherSubscribers} \cap (\geq 1 \text{agent}) \cap)</td>
<td>(3)</td>
</tr>
<tr>
<td>(\exists \text{theme}, {\text{urn:doi.10.1032}/...} \cap (\leq 1 \text{theme}))</td>
<td>(4)</td>
</tr>
</tbody>
</table>
Each intersected restriction reduces the initial set of actions, which corresponds to all the Copy actions. First, (2) models the time range as a restriction on the pointingTime case role to a custom datatype. This case role is a functional property so no additional constraints on cardinality are required. The last constraints, (3) and (4), restrict the range of agent to one or more instances of the “PublisherSubscribers” class and theme to just the instance “urn:doi.10.1032/…”.

From this point, the implementation is quite straightforward. DL reasoners are specially suited for classifying individuals into classes when the latter are based on necessary and sufficient conditions. They can answer if an individual, considering its relations to other individuals and attribute values, satisfies all the restrictions of a class pattern and, thus, can be classified as an instance of that class.

In the context of the Copyright Ontology implementation, this functionality is used to check if a particular action, modelled as an individual, is allowed or not by a license. This corresponds to the fact that the action individual is classified into a class pattern that is the theme of an Agree. Another reading is that the license agrees on performing a set of actions that includes the requested one.

However, before the action is authorised, it is also necessary to check that any existing condition is met and that there is not any disagreement on the action. The DL reasoner is also useful for this part. It is checked if the precondition pattern is instantiated, so the precondition is satisfied, and that the checked action is not classified into a class pattern that is the theme of a Disagree.

To sum up, it is checked that there is an agreement on the action and no disagreement, and that the precondition is satisfied. This behaviour allows modelling complex licenses, revocation and avoiding the Open World Assumption inherent to OWL-DL as it is detailed in the next section.

3.1. Overcoming the Open World Assumption

The main problem of the OWL-DL implementation presented in the previous section is the Open World Assumption (OWA). This problem arises when the DL reasoners are trying to classify a given instance into the classes for license patterns. It can be said that
the reasoner is very “conservative” as, although the necessary and sufficient conditions are met, it will not classify an instance into a class if new facts can make it retract from this decision.

In some cases this is the desired behaviour but this is not the case for this license checking implementation. The intention is to make a local Close World Assumption and make a decision on the currently available facts as the outcome is to decide if the current action should be authorized or not at check time.

There are some OWL-DL restriction primitives that lead to OWA problems:
- **maxCardinality** (≤ n): the reasoner is conservative with this restriction as, although the cardinality restriction might be satisfied at a given time point, new facts can make the cardinality greater than n, i.e. (>n). The cardinality restriction is also affected as it is the conjunction of maxCardinality and minCardinality restrictions.
- **allValuesFrom** (∀R.C): the situation in this case is that, although at the current time all the values for the R property are in the C class, in the future, there might be new facts that involve R with a value not in C, i.e. R(¬C).

On the other hand, other OWL-DL restrictions, or their combination, are not affected by the OWA and thus do not affect the license checking implementation. Some of them are:
- **minCardinality** (≥ n): there is no OWA problem here as once the reasoner can check that the cardinality is equal or greater than n, i.e. (≥ n), new facts cannot make this inference false, i.e. (< n).
- **someValuesFrom** (∃R.C): once there is some R whose value is in C, new facts cannot make that there does not exist some R with a C value, i.e. ¬(∃R.C).

Therefore, there is not an OWA problem here.
- **FunctionalProperty** R: this constraint makes an allValuesFrom restriction OWA insensitive.

There are many ways to overcome OWL-DL’s OWA through epistemic operators [18] and non-monotonic OWL extensions [19]. However, in these cases it is necessary to get outside standard OWL-DL and, what is even more inconvenient, there are just preliminary implementations of these approaches.

In order to implement an OWL-DL based license checker, we have adopted a more pragmatic approach, which does not require additional language constructs neither a reasoner different from the existing OWL-DL ones. Figure 9 illustrates this approach. As it can be observed, a maxCardinality restriction defines a set of accepted cardinality values, e.g. from zero to two. However, as new facts are known under an OWA, instances previously classified into this restriction can “get out” of the corresponding set.

On the other hand, it can be observed that a maxCardinality restriction (≤ n) has an opposite set corresponding to the minCardinality restriction (≥ n+1). As it has been shown, the minCardinality construct is not affected by OWA. Therefore, the idea is to take profit from this fact and make the reasoner look for the opposite set, the OWA-insensitive one, and check that it is not satisfied. This implies that the reverse is satisfied by the current set of facts at hand and overcomes the OWA assumption that makes the reasoner not able to infer that. The same applies to the allValuesFrom restrictions and the reversed someValuesFrom restrictions.
Therefore, the approach is to negate the restriction and to undo this at the metalevel, i.e. to check outside the DL reasoner that the negated restriction is not satisfied and thus it can be inferred that the original one does. The negation is modelled at the metalevel using the Disagree class, which is the opposite of the Agree class.

Therefore, the allValuesFrom restriction on the theme class of an Agree, i.e. ∀R.C, is converted into the reversed someValueFrom restriction on the theme class of a Disagree, i.e. ∃R.¬C. On the other hand, the maxCardinality restriction on the theme class of an Agree, i.e. (≤n R), is converted into a minCardinality restriction on the theme class of a Disagree, i.e. (≥n+1 R).

The previous method is applied to class patterns and all the OWA-sensitive constructs are moved to a new class pattern which is disagreed in order to model the metalevel negation. This new class results from the disjunction of all the transformed restrictions and is intersected with the original pattern, which is now composed by just the OWA-insensitive restrictions and is what remains as the subject of the Agree.

Continuing with the commercial copy pattern described in Table 2, the following example illustrates this mechanism. Table 3 shows the class patterns that overcome the OWA and result from the previous transformation. Note that pointInTime is defined as a functional property in the Copyright Ontology so it is not affected by the OWA and remains unchanged (2), like the other OWA-insensitive constructs (3). All of them build Pattern ′ that corresponds to the OWA insensitive part of the original Pattern, i.e. the pattern that is agreed.

\[
\begin{align*}
\text{Pattern}^{'} & \equiv \text{Copy} \cap \\
& \forall \text{pointInTime} \geq 2007-01-01T00:00:00, \leq 2007-06-30T23:59:59 \cap \\
& (\geq 1 \text{ agent}) \cup \exists \text{ theme.\{urn:doi:10.1032/\ldots\}} \\
\text{Pattern}'' & \equiv \text{Pattern}^{'} \cap \\
& (\exists \text{ agent}.\neg \text{PublisherSubscribers} \cup \\
& (\geq 1 \text{ theme}) 
\end{align*}
\]

\textbf{Table 3. OWA-insensitive classes for the commercial copy action pattern in Table 2}

On the other hand, Pattern ′′ contains the transformed OWA sensitive constructs in the original Pattern, i.e. the disagreed pattern. There is the someValuesFrom restriction (5) corresponding to the allValuesFrom restriction (3) in Table 2 and the minCardinality restriction (6) corresponding to the maxCardinality one (4) in Table 2.

The combination of both patterns, the first the theme of an Agree and the second of a Disagree, ends up building a pattern like the one shown in Figure 10. The set of actions that is authorised corresponds to the darker part, i.e. Pattern ′ minus Pattern ′′.

The figure also shows three example instances, in N3 notation, situated into the class patterns under which they are classified. Consequently, just instances of Pattern ′
that are not instance of Pattern” are authorised. The values that make instances to not be classified into Pattern’ are highlighted using bold letters. It is assumed that “Roger” is an instance of PublisherSubscribers and that “Matt” does not.

Figure 10. Interpretation of the class patterns in Table 3

3.2. Creative Commons Scenario

As it has been pointed out in the introduction, one of the more successful copyright licensing approaches is Creative Commons (CC). However, the machine-readable version of CC licenses models a very limited part of the licenses semantics, i.e. many details in the human-readable version are not captured in the computer one.

Consequently, current CC licenses are just appropriate for license search computerised support. In order to enable more sophisticated services for CC licenses, we have modelled them using the Copyright Ontology. Thus, we can make their semantics explicit and propagate the previously described license reasoning capabilities to the enormous pool of CC licenses.

As the semantics are implicit in the human-readable version, it has been necessary to interpret and manually map the licenses to Copyright Ontology concepts. Fortunately, there is a limited set of predefined licenses’ so this is an easily affordable process.

For instance, Table 4 presents the model for the “Attribution Share-Alike” (by-sa) CC license. This license allows licensees to remix, tweak, and build upon the copyrighted work even for commercial reasons, as long as they credit the author and license their new creations under identical terms.

Consequently, in order to license under these terms any manifestation of “myWork”, the three “By-Sa” class patterns are granted. The first one, “By-Sa_1”, lets licensees to copy, distribute, communicate and make available any manifestation of “myWork” if they give credit to the author. The second one, “By-Sa_2” authorises

5 Creative Commons Licenses 3.0, http://creativecommons.org/licenses
derivations of the work and the third one, “By-Sa_3” establishes the same terms than for the original work for those derivations.

<table>
<thead>
<tr>
<th>Table 4. Copyright Ontology model for the CC by-sa license</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>By-Sa_1</strong> ≡ (Copy ∪ Distribute ∪ Communicate ∪ MakeAvailable) ∩</td>
</tr>
<tr>
<td>( \forall ) theme. MyManifestation ( \sqcap (\geq 1 ) theme) ( \sqcap</td>
</tr>
<tr>
<td>( \exists ) condition. AttributeMe</td>
</tr>
<tr>
<td>MyManifestation ≡ Manifestation ( \sqcap \forall ) isManifestationOf. {myWork}</td>
</tr>
<tr>
<td>AttributeMe ≡ Attribute ( \sqcap \exists ) theme. {myWork} ( \sqcap \exists ) recipient. {me}</td>
</tr>
<tr>
<td><strong>By-Sa_2</strong> ≡ Derive ( \sqcap</td>
</tr>
<tr>
<td>( \exists ) theme. {myWork} ( \sqcap</td>
</tr>
<tr>
<td>( \forall ) result. NewManifestation ( \sqcap (\geq 1 ) result)</td>
</tr>
<tr>
<td>NewManifestation ≡ Manifestation ( \sqcap</td>
</tr>
<tr>
<td>( \forall ) isManifestationOf. NewWork ( \sqcap (\geq 1 ) isManifestationOf)</td>
</tr>
<tr>
<td>NewWork ≡ Work ( \sqcap \exists ) isDerivationOf. {myWork}</td>
</tr>
<tr>
<td><strong>By-Sa_3</strong> ≡ (Copy ∪ Distribute ∪ Communicate ∪ MakeAvailable) ∩</td>
</tr>
<tr>
<td>( \forall ) theme. NewManifestation ( \sqcap (\geq 1 ) theme) ( \sqcap</td>
</tr>
<tr>
<td>( \exists ) condition. AttributeMe</td>
</tr>
</tbody>
</table>

4. Conclusions and Future Work

We are not profiting from the full potential of Internet-wide content sharing and reuse because the underlying copyright issues are not made explicit and dealt with. Instead, the reaction is to protect content using security mechanisms that limit the possibilities.

A good example of the potential of a less restrictive approach is Creative Commons licensing schemes for open access and reuse of content. However, Creative Commons does not constitute and alternative to DRM. It lacks the required flexibility to incorporate additional license terms, like commercial ones, and advanced computerised support.

Our semantic web approach to copyright management constitutes an alternative. It provides an expressive conceptual framework, the Copyright Ontology, which provides the building blocks for flexible machine-understandable licenses.

Altogether, it constitutes a tool that helps people state the copyright conditions for the content they share and how it might be reused. A way to build an Internet-wide licensing network adapted to particular needs: commercial or non-commercial, open or closed access, reusable share-alike content, etc.

Moreover, thanks to its OWL-DL implementation and mechanisms to overcome the Open World Assumption, it can be put into practice quite easily by using existing DL reasoners. License reasoning allows checking if a particular action is granted by a pool of licenses. This capability can be propagated to existing license modelling languages, like DRM standards [10] or Creative Commons licenses.

From these results, future work focuses on combining the DL layer, which deals with patterns and events, with a rule-based metalevel. Currently, the metalevel is implemented procedurally, which is possible due to the limited range of interpretations of the DL classifications. However a metalevel implementation based on Semantic
Web rules would make available a greater level of flexibility and new functionalities. For instance, rules can facilitate incorporating penalties into the system, i.e. obligations that take place when obligations are violated [20]. Currently, obligations are just monitored in order to detect violations.

References

Integrated Applications
This page intentionally left blank
An Ontology-Based Decision Support System for Judges

Pompeu CASANOVAS a,1, Núria CASELLAS a and Joan-Josep VALLBÉ a

a Institute of Law and Technology, UAB, Barcelona (Spain)

Abstract
This article briefly outlines how the Iuriservice application was envisioned, the modeling process of the Ontology of Professional Judicial Knowledge, and the contribution of the users to the global development of this system. Iuriservice offers a semantically enabled FAQ search system for the Spanish judges in their first appointment. The system is now at the first stage of implementation in the Spanish Judicial School. Users may input questions to the system in natural language to obtain access to a database of experience-based answers to practical day-to-day questions. In order to offer the question-answer pair from the application database that best matches the input question, the search system is enhanced using ontologies. Some lessons learned are also outlined.

Keywords. Decision support system, knowledge management, legal ontologies, professional knowledge, user validation

Introduction

Legal and judicial practices are nowadays intensive knowledge management domains. Lawyers, judges and other legal professionals face an exponential growth of rules, judgments and legal documents. In addition, legal professionals perform their tasks in distributed, rather than isolated, network environments [2]. To address these two issues of current legal practice we have designed Iuriservice as a specialized semantic frequently asked questions (FAQ) for the judicial domain.

Iuriservice is a decision-support application that allows easier, faster searching of professional judicial knowledge through the use of ontologies; searches are conducted within a repository of previously stored questions regarding judicial practice. This FAQ search system enables the user to search a database of stored question-answer pairs using a natural language interface.

In this article we will first describe how the application was envisioned and designed, then we will briefly describe the construction process of the Ontology of Professional Judicial Knowledge (OPJK), and finally we will describe the involvement of the users in the validation process towards the implementation of the application.

1Corresponding Author: Pompeu Casanovas, Institute of Law and Technology, Faculty of Law, Building B, Campus UAB, Bellaterra 08193 (Barcelona), Spain; E-mail: pompeu.casanovas@uab.cat.
1. Problem Description and Application Design

The need for the system and its initial design were established as a result of two thorough ethnographic surveys carried out with the collaboration of the Spanish General Council of the Judiciary (CGPJ). In 2002-2003 the Institute of Law and Technology conducted a national survey on newly recruited judges [1]. The main aim of the survey was to identify the most frequent problems that young judges usually faced in their first appointment. In particular, in depth interviews were made to 130 judges of the 52nd class, out of 378 young judges who had completed their studies at the School between 1997-1999.

From that survey, the most important features were

- the fact that new entrants to the Spanish judiciary were faced with a great variety of cases, procedures, hearings, decisions, and rulings, and
- that in a number of situations (especially those in which legal protocols regarding procedures were unclear or missing) these judges solved the cases asking to a peer or a more experienced judge for professional advice.

Although peer consultation is the traditional way of sharing and distributing knowledge within the judicial profession, it might prove to be an uncertain knowledge management process (i.e. different contradictory solutions to the same situation may coexist), which also could slow down the performance of both the judge asking for advice and the judge responding, especially during the late hours of the on-call period. Thus, a system that stored and managed this professional knowledge could offer solutions to the judicial profession in order to distribute, maintain, and avoid inconsistencies of this practical knowledge shared by the professionals (professional judicial knowledge, PJK).

Yet, if a system was to provide efficient support in a reliable manner, the construction of the repository of knowledge and its accuracy and validity (answers to questions regarding practical problems of the judicial profession) were critical. Therefore a second, extended ethnographic survey targeting the data for the application (context and content of problems for the newly recruited judge) was carried out. This ethnography was designed as a three-step process. The process consisted of: 1) oral interviews, 2) literal transcriptions of these interviews, and 3) the elaboration a list of the practical problems that the judges had encountered during their first years of practice. With this final step, we obtained a list of nearly 800 questions regarding problems faced during their first appointment that required consultation and peer expertise. The survey covered 14 Spanish Autonomous Communities (out of 17) and was carried out by a team of fourteen researchers (trained lawyers and law professors from the UAB Law School). Preliminary findings were consistent with the core hypothesis of research and prototype development [7]: the usual practice within the group of newly recruited judges consisted of requiring and gathering information from their seniors regarding practical matters either for case resolution or court management.

Therefore, we considered it would be useful to produce a repository of professional judicial knowledge to support judicial activity, integrated in a web-based system with fast
and reliable access for all the entrant judges. This repository could be fed from the practical questions obtained from the field work and it could function as a Frequently Asked Questions database. Then, the system should offer access to this database, containing the judicial expert knowledge, through a simple natural language interface. Through this interface, the judge might describe the problem at hand and the application ought to respond with a list of relevant question-answer pairs that offer solutions to that particular problem.3

This system might be a solution for what we may call “generation gap”. Professional knowledge (PK) is individual, often personal, but largely shared with other people working in the same domain. It is tacit and implicit in professional behavior. Therefore, the bulk of PK is socially, but unequally distributed. We refer as professional judicial knowledge (PJK) to the knowledge belonging to magistrates and judges showing these properties. It is different from general legal knowledge, the one acquired at the Law School. But it is still legal-centred, because precedent, rules and statutes constitute the core of judges’ experience.

Unfortunately, when judges retire, some professional judicial knowledge disappears, and the new generation of judges has to discover it again. Gathering part of this knowledge in an available repository could help them to overcome this “generation gap”.

This Iuriservice repository consists of the practical questions gathered in the surveys and their corresponding answers, which have been provided by a team of eight experienced magistrates from the Spanish Judicial School. Figure 1 shows an example of a question/answer pair that Iuriservice could provide.

**Question.** While on duty, a judge receives a call from a hospital reporting a sexual assault. The victim has not made yet an official report of the incident. Procedures to be followed? Which rules apply?

**Answer.** As for the procedures to be followed, a forensic scientist should be sent to the hospital in order to examine the victim and to take samples. If the crime has not yet been officially reported, the judge except in very exceptional circumstances may begin no procedures. Provided that it is clear from the telephone call alone that this is a case of sexual assault and that no other crime has been committed, then criminal proceedings must be initiated by the victim...

Figure 1. Example of a question-answer pair contained in the Iuriservice FAQ repository.

Do notice that the practical nature of questions reflects also some problems stemming from the diverse functions fulfilled by different professionals involved in the legal practice. For instance, coordination problems between judges and prosecutors in court are typical (see Figure 2), and judges at their first appointment must develop certain organizational skills to solve them.

---

2 In 2000, 47% of members of the 52nd class of judges declared to have no computer skills. Those who declared to have some were mostly used to text processors, and only 35% of the training judges declared to be Internet users [14]. Yet in 2002 the use of Internet had increased among judges: 60.6% of the interviewed inexperienced judges declared to use Internet [1]. Further surveys show that the use of Internet is gradually increasing with the training at the Judicial School and during their first appointment [15].

3 See [7] and [8] for further information about the system.
Question. I have a prosecutor assigned to my court, this is the one who signs the documents and gives the formal OK on behalf of the Prosecution, but he does not show up at the hearings. Another prosecutor is coming. Is this normal? May I require his presence?

Answer. No. This is a prosecutors’ office internal matter. The principle of unity ruling the Public Prosecution implies that any of them can carry out her function independently to her concrete assignment in a particular moment.

Figure 2. Example of a practical problem within the judicial office.

2. A Legal Professional Ontology: OPJK

Effectiveness is an important design issue for the system: it should be able to locate within the repository the best possible matching stored question to the user’s input question that tackles the problem. Due to this requirement, the application could benefit from the use of semantic technologies, through the use of ontologies, in order to provide more accurate searches than the basic keyword search. Several ontologies were modeled to this end. In this section we will describe the Ontology of Professional Judicial Knowledge (OPJK) that models the knowledge contained in the questions gathered from the field work (the domain knowledge).

As stated above, one of the main corpuses for knowledge modeling obtained through the field work were the nearly 800 practical questions formulated by the newly recruited judges. These questions contain the professional judicial knowledge gathered during daily practice at courts and constitute the repository of the application. The Ontology of Professional Judicial Knowledge, then, models and represents formally this professional (not purely theoretical) judicial knowledge in order to make the questions accessible through a natural language interface. The construction process dealt with several tasks, first, the establishment of the ontological requirements, then, the analysis of the corpus and extraction of relevant terms (concepts), and finally, the ontology construction and formalization processes.

2.1. Requirements

At first, the domain and goal of the ontology, the design guidelines and knowledge sources, and user and application requirements were specified. Some of those requirements were specified as follows. As stated above, the main purpose of the ontology was to aid (providing more accuracy) the retrieval of relevant stored question-answer pairs regarding judicial practice (semantically similar to the input natural language question). Thus, the ontology domain ought to cover the concepts contained in the questions included in the database: concepts from the legal domain from the point of view of the judicial practice. At the same time, those questions dealt with several legal domains, from family law or commercial law to criminal law. Therefore, the legal ontology to be modeled was to include concepts from different legal subdomains. In this sense, the main knowledge source was to be the corpus of the questions obtained in the field work and stored in the database of the system. This corpus would be used as the text for automatic term extraction and knowledge acquisition. Other design requirements were the tight collaboration between domain experts and ontology engineers in a distributed environment for the construction of the ontology, the information retrieval nature of the tool (Juriservice), the judge as the user of the tool, and Spanish as the language to be used as input and output natural language.
2.2. Corpus Analysis and Term Extraction

The corpus of questions was used as the input text for the construction of the ontology. Several tools were used in order to analyze the text to offer more information regarding the content of the text. For example, ALCESTE\(^4\) (see Figure 3) as a textual statistics software, allowed a fruitful clustering analysis and a first glance of a possible subdomain classification. Later, ONTOGEN\(^5\) (see Figure 4) allowed legal experts to refine former subdomain classifications for the corpus of questions [18,13]. Other tools were also used to extract relevant terms that could be included in the ontology [12].

\(^4\)http://www.image.cict.fr/index_alceste.htm
\(^5\)http://ontogen.ijs.si/
TextToOnto and later Text2Onto provided lists of extracted terms (see Figure 5). With the improved version of Text2Onto which supports Spanish, verbs do not appear in the concept extraction; it extracts only nouns: "caso" [case], "juicio" [trial], "parte" [party], "juez" [judge], "persona" [person], "orden" [injunction], "vista" [trial], "policia" [police/policeman], "guardia" [on-duty], "prueba" [evidence], "procedimiento" [process], "funcionario" [civil servant], "juicio rápido" [fast trial], etc. However there are still some features which need improvement such as the appearance of "¿Cómo" [how], "¿Cuál" [which], "¿es" [is]. Approximately, only the 8% of the concepts extracted either made no sense (for example, "alcalde de imputado" [mayor of impeached] or were not concepts (such as the above mentioned). At this stage, approximately 1500 concepts were extracted from the text.

Figure 5. Text2Onto concept extraction.

However, the main difficulty encountered with the use of term extraction and ontology extraction tools apart from the use of Spanish language as input language, was the lack of contextual references that this list of extracted terms provided. Due to all these difficulties and the need to treat the extracted terms within their context (the specific question), the ontology was developed manually from the selection of relevant terms for each of the questions. A middle-out strategy was followed and the extracted lists of terms were used as guidelines towards conceptual relevance.

2.3. Ontology Construction and Formalization

The methodology used was influenced by the particular needs of the ontology domain and language and also by the fact that the responsibility for ontology building was shared by domain experts and ontology engineers alike during all the construction process. Domain experts were not only involved in the knowledge elicitation process but also dur-

\[\text{http://ontoware.org/projects/text2onto/}\]
ing the requirement specification phase, the knowledge acquisition and elicitation, the formalization process, and later during the evaluation process. The construction process followed a middle-out strategy; selected terms contained in the questions were included in the ontology, formalized at first in RDF and currently in OWL, and were generalized or specified when necessary. For that reason, the reuse of top or upper ontologies was taken into account when a certain level of generality was achieved and the reuse of a top ontology could be beneficial towards organizational purposes. The top level (module) of the PROTON ontology was used for this purpose. Moreover, the distributed modeling environment (ontology engineers and domain experts) forced distance discussions and cooperative modeling that required to take into account methodological support. The DILIGENT methodology (DIstributed, Loosely-controlled and evolvInG Engineering of ONTologies) was taken into account to support such relations.

2.4. Ontology Refinement and Evaluation

At the moment, the Ontology of Professional Judicial Knowledge used in Iuriservice has a total of 119 concepts and 561 instances. This ontology was evaluated regarding effectiveness. The measurement compared the precision of the user’s query response between: (1) a search based on keywords, morphological analysis, and synonyms, and (2) a search based on keywords, morphological analysis, synonyms, and ontologies (semantic distance). See the results of the effectiveness tests in Table 1. An improvement on effectiveness can be noticed when semantic distance and ontologies are added to keyword searches.

<table>
<thead>
<tr>
<th></th>
<th>Keywords</th>
<th>Keywords + Ontologies (semantic distance)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Success</td>
<td>28.57%</td>
<td>45.71%</td>
</tr>
<tr>
<td>Failure</td>
<td>71.43%</td>
<td>54.29%</td>
</tr>
</tbody>
</table>

Taking these results into account, OPJK is currently being refined: concepts and concept definitions are being revised and added, and the instance/class distinction is also under verification. When this refinement process is complete, the ontology will be evaluated and further measurements will be performed.

3. User-centered Approach

Regarding validation and testing activities, the development of this application has followed the spiral model of software development process. Users were involved in the whole development cycle of the Iuriservice application. In fact, the different usability and users tests were made possible by the helpful intervention of the Magistrates from

---

7 PROTON (<http://proton.semanticweb.org/>) was also being developed within the SEKT consortium. See [11] for more details.

8 A corpus of 68 FAQs from the gender violence domain was used. Two types of tests were performed taking into account the (1) user’s input questions with the same meaning as the target FAQ question, and (2) the user’s input questions with different meaning from the target FAQ question. See [13] and [3] for further information regarding effectiveness and efficiency measurements.
the Spanish Judicial School, a group of trainee judges, and a group of legal experts from the academia. In the early phases users needs were investigated, and the results were fed into the development process. During the actual development of the application different approaches were used. The methods included the inspection of the prototypes (including its interface) by experts (Heuristic Evaluation) and Cognitive Walkthroughs with early prototypes. In this way it was assured that the prototype delivered to users did not show major defects and met their needs and requirements [10]. We founded our Heuristic Evaluation test upon the 10 standard usability heuristics proposed in [17], and they were presented to the usability experts translated in Spanish in order to make them useful and helpful. Besides, we also followed the basic instructions and methods appeared in [4].

Regarding the Heuristic Evaluation, the main improvements were related, among others, to visibility and system status (directed to inform the users about the status of the system, through appropriate feedback within reasonable time), error prevention (a carefully designed system should prevent the user from experiencing system error messages), and help and documentation (although the system provided some help and documentation, it was found not to be sufficient).

As for the Cognitive Walkthroughs, the tasks selected were fully representative of the tasks the system would be able to perform when it were completed, and the results proved to be very useful for the improvement of the system.

Finally, some field tests of the Iuriservice application were designed to provide a test as realistic as possible of the application by judges and legal experts [5]. Clearly the user group of main interest for the study was young judges in their first position. These judges have a high workload, and they often require urgent support outside normal hours of work. The tests were prepared for a group of judges, who participated in a test lasting for two days, and a group of legal experts.

![Figure 6. Field tests at the Judicial School.](image)

On the first day a number of typical cases were presented to the judges, and they were asked to solve the cases in a randomized order, using their usual tools, and their normal working procedures. Examples of these cases may be seen below.

These data served as a baseline for the tests on the second day. During this second day of tests, first, an introduction to Iuriservice was given, and then all participants carried out some easy exercises to familiarize themselves with Iuriservice. After that, the
Case 1. How to proceed when the judge is called on-duty and there is no coroner to perform the removal of the corpse?

Case 2. When a person with a serious mental disorder becomes violent, may the judge decide the confinement in a mental health centre without the prosecutor’s or the coroner’s attendance?

Figure 7. Examples of cases presented to the judges.

participants had to solve a number of cases from the same sample as those used on the first day, again in randomized order. Both, their usual working tools and Iuriservice were available on their PC. In sum, during the two days each person solved all cases, half with and half without the use of Iuriservice.

The subjects recorded the time taken to solve the task and the number of queries executed. The decisions made were noted (and were assessed by independent experts). After the completion of all test cases, the subjects were asked to rate Iuriservice, and to complete the SUMI (Software Usability Measurement Inventory) questionnaire [16].

The same procedure was carried out with a group of 10 legal experts, but with a smaller number of cases.

The comparison of the time taken by the subjects, as may be seen in Table 2, for the same tasks either with or without the use of Iuriservice shows a significant difference. All tasks were solved in a much shorter time when Iuriservice was also available for the task. The results strongly suggest that Iuriservice provided an extra information service, complementing the existing databases, which allowed the users to find the required information quickly and easily in many situations.

Table 2. Mean time to solve one case [6]

<table>
<thead>
<tr>
<th></th>
<th>Without Iuriservice</th>
<th>With Iuriservice</th>
</tr>
</thead>
<tbody>
<tr>
<td>Judges</td>
<td>23 minutes per case</td>
<td>9 minutes per case</td>
</tr>
<tr>
<td>Legal experts</td>
<td>20 minutes per case</td>
<td>5 minutes per case</td>
</tr>
</tbody>
</table>

In conclusion the subjects were able to use Iuriservice without any apparent problems. The tool was easy to learn, and users rated it as highly positive. In fact, judges expected the introduction of Iuriservice into their working environment to be beneficial, both in terms of efficiency of their activity and reduction of their workload.

4. Lessons Learned and Conclusions

During 2008, the system will be implemented at the Judicial School. Up to now, we have learned several lessons from the whole process. Usually, when describing method-

9The results of the SUMI questionnaire (http://sumi.ucc.ie/) show that subjects assessed the Iuriservice application as highly positive [6]. The five measures are defined as: efficiency (refers to the user’s feeling that the software enables him to perform the task(s) in a quick, effective and economical manner; affect (it means emotions, and it refers to the positive user feeling of the user being mentally stimulated and pleased as a result of interacting with the software); helpfulness (refers to the user’s perceptions that the software communicates in a helpful way and assists in the resolution of operational problems); control (refers to the feeling that the software is responding in an expected and consistent way to input and commands); and learnability (refers to the feeling that the user has that it is relatively straightforward to become familiar with the software).

10The data are means for 9 judges and 9 legal experts under the condition without Iuriservice (data missing from one subject in each group), and 10 judges and 10 legal experts with Iuriservice [6].
ology and ontology building, scholars divide the processes into several ideal consecutive cycles: knowledge acquisition, ontology construction, user validation planning, etc. Yet our work within institutional settings, with real users and legal and organizational constraints, brought about the need to adapt the ontological engineering cycle to the time and other constrictions of the institution. The time and capabilities of the institution follow a social and political dynamic that does not match necessarily the needs that developers may have. For example, the staff of the Spanish Judicial School was completely renewed during our collaboration, including its director. As this staff was answering the questions to be included in the system database, we could not use the complete set of question-answer pairs for the user tests. The new team is again fully involved in the development process.

For the actual development of Iuriservice, to respect and adapt to these institutional constrictions has been critical. As a consequence, knowledge acquisition, ontology construction or user testing activities have been carried out at the same time, as an ongoing process.

The process towards the construction of the OPJK ontology and the design of the Iuriservice application has been time consuming and complex. The full process has required: surveys, the use of different textual analysis techniques, the establishment of a distributed and cooperative ontology construction process, the need for refinement and application tests, and, also, different user validation activities. Currently, the ontology is being refined to improve efficiency and effectiveness measures, and the application is being modified to comply with some final user requirements expressed in the field tests.

Acknowledgements

We thank our colleagues at the Institute of Law and Technology (UAB), iSOCO (Intelligent Software Components, S.A.) and the SEKT Project, including V.R. Benjamins, M. Blázquez, Th. Bösser, J. Contreras, S. Fernández-Langa, B. Fortuna, A. Jakulin, A. Kiryakov, J.L. López-Cobo, E.-M. Melchior, M. Poblet, Ch. Tempich, J. Völker and D. Vrandečić without whom the research would not have been possible. We also should like to thank the staff from the Spanish Judicial School for their dedication and help. This research has been partially funded by EU in the IST-2003-506826 Project SEKT (Semantically Enabled Knowledge Technologies) and the Spanish national projects SEC-2001-2581-C02-01/02, Observatorio de Cultura Judicial [Observatory of Judicial Culture], and SEJ 2006-10695, Implementación de Iuriservice [Iuriservice Implementation].

References


This page intentionally left blank
A two-level knowledge approach to support multilingual legislative drafting

T. AGNOLONI\textsuperscript{a}, L. BACCI\textsuperscript{a}, E. FRANCESCONI\textsuperscript{a,1}, W. PETERS\textsuperscript{b}, S. MONTEMAGNI\textsuperscript{c}, G. VENTURI\textsuperscript{c}

\textsuperscript{a} ITTIG-CNR, Florence, Italy
\textsuperscript{b} Dept. Computer Science, Univ. of Sheffield, UK
\textsuperscript{c} ILC-CNR, Pisa, Italy

Abstract. The quality of legislative drafting process at European and national levels is highly influenced by the legal drafters control over the multilingual complexity of European legislation and over the linguistic and conceptual issues involved in its transposition into national laws. The DALOS project aims at ensuring coherence and alignment in the legislative language, providing law-makers with a knowledge (ontological-linguistic) resource and knowledge management tools to support the multilingual legislative drafting process. This paper outlines the activities within DALOS, aiming at the definition of the characteristics of the knowledge resource, at its implementation, at its integration in a legislative drafting environment for the project prototype.

Keywords. Multilingual legal drafting, Legal ontologies, NLP techniques

1. Introduction

Quality in European and national legislation is one of the main purposes of the current initiatives of the European Commission. In the Mandelkern report on Better Regulation \cite{1} the need for a coordinated action by Member States was solicited to simplify the EU regulatory environment, to enhance the quality of EU legislation as well as to rationalise the transposition of Community legislation into national law. The Mandelkern Group on Better Regulation \cite{1} in particular stressed on “quality of regulation” as an essential precondition to enhance the “credibility of the governance process” and to contribute to “welfare of citizens, business and other stakeholders”. The analysis of \cite{1} identifies some significant problems in the “lack of simplicity, clarity and accessibility of European provisions – such as unclear, confusing terminology, incomplete or inconsistent regulations or use of vague terms”. Such problems, together with different legislative cultures within the Commission, are causes of accessibility difficulties of EU regulation. Problems of conceptual misalignment arise in particular when transposing Euro-

\textsuperscript{1}Corresponding Author: E. Francesconi, via de’ Barucci 20, Florence, Italy; E-mail: francesconi@ittig.cnr.it
ITTIG-CNR group involved in DALOS includes also Daniela Tiscornia and Pierluigi Spinosa.
pean directives into national laws: they are not only caused by the content and the quality of European legislation, but they are also the result of national practices, national legal differences and different cultures.

Coherence, interoperability and harmonization in the legislative knowledge of, and control over, the legal lexicon is therefore considered a precondition for improving the quality of legislative language and for facilitating access to legislation by legal experts and citizens. In a multilingual environment and, in particular, in EU regulations, only the awareness of the subtleties of legal lexicon in the different languages can enable drafters to maintain coherence among different linguistic versions of the same text, as well as their transposition in national laws.

Better regulation guidelines therefore have been given [1], which are targeted to guarantee more coherence in legal texts to enhance understandability and uniform interpretation of EU law in each national legal system [2]. In particular they aim to encourage:

- Coherence within EU law (internal coherence [3]), avoiding inconsistent definitions of EU legal terms as well as avoiding contradictory use of legal terms within different sectorial legislative interventions [2] (see, in EC Directives on consumer law, terms like Timeshare, Distant Contract, Unfair [3]);
- Coherence outside the EU law (external coherence [3]), avoiding that a same legal concept can be expressed in different ways in a Directive and in its transposition into a national law [2].

To face these problems the DALOS project has been launched within the “eParticipation” framework, the EU Commission initiative aimed at promoting the development and use of Information and Communication Technologies in the legislative decision-making processes. The aim of such initiative is to foster the quality of the legislative production, to enhance accessibility and alignment of legislation at European level, as well as to promote awareness and democratic participation of citizens to the legislative process.

In particular, DALOS aims to ensure that legal drafters and decision-makers have control over the legal language at national and European level, by providing law-makers with linguistic and knowledge management tools to be used in the legislative processes, in particular within the phase of legislative drafting.

Nowadays the key approach for dealing with lexical complexity is the ontological one [4], by which it is possible to characterize the conceptual meaning of lexical units, as well as to provide a detailed description of the semantic properties of the related concepts as well as their relationships.

In this paper the development of a two level (ontological and linguistic) knowledge resource for DALOS is described. In particular, in Section 2 the complexity of the multilingual legal scenario is addressed; in Sections 3 the design principle of the DALOS knowledge resource and the characteristics of its Knowledge Organization System (KOS) are presented; in Section 4 the phases for implementing the DALOS resource are introduced; in Section 5 and 6 the implementation of the DALOS knowledge resource organized in two levels (lexical and ontological)

---

2Drafting Legislation with Ontology-based Support
is illustrated; in Section 7 a description of the prototypical integration of the DA-
LOS resource in a legislative drafting environment is given; finally in Section 8
some conclusions are reported.

2. Approaching the multilingual EU legal scenario for legal drafting

In legal language any vocabulary originated by a law system is an autonomous
linguistic resource. To approach the complexity of the multilingual scenario as well
as to align concepts of the EU legal domain one cannot transfer the conceptual
structure of a legal system to another, because of different national legal contexts
and legislative cultures within EU Member States, as discussed in [5]. One of
the widely accepted approach consists in developing parallel alignment between
different terms in different languages representing the same concept.

A similar problem arises as regards the obligation of EU Member States
to implement European Directives into national laws. Far from being a straight
transposition, this process usually includes a further step in which European
Directives are subject to interpretation which can lead to questionable results (see

In the literature different methods exist for approaching the multilingual com-
plexity of European law, for example controlled vocabularies, implemented in ter-
iminology database (such as IATE run by all the main EU Institutions), thesauri
(as EUROVOC, maintained by OPOCE\(^3\)), semantic lexicons or lightweight on-
tonologies (as WordNet [6], EuroWordNet [7] and, in the legal domain, JurWordNet
[8]). Alignment of multilingual terminologies can be effectively obtained by using
a pivot language. More expressive descriptions of concepts associated with lexi-
cal units can be contained in domain ontologies (or statute specific ontologies),
representing concepts used in a specific statute (as IPROnto [9]). More general
organizations of domain concepts are addressed in literature as core ontologies
(as LRI-Core [10], LKIF [11] and CLO [12] for the legal domain), while founda-
tional concepts categories applicable to all domains are usually addressed in top
or foundational ontologies (as SUMO [13], DOLCE [14]).

The integration of different lexical resources (heterogeneous because belong-
ing to different law systems, or expressed in different languages, or pertaining to
different domains) can be carried out in different fashions:

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

The methodological approach chosen in DALOS is the third one: it requires
the definition of mapping procedures between semantic lexicons, driven by the
reference to an ontological layer where the basic entities which populate a legal
domain are described.

This approach has been chosen because:

\(^3\)Office for Official Publications of the European Communities
• it provides a higher degree of modularity for the knowledge architecture: lexical units of a generic lexicon can be associated to concepts pertaining to various domains, thus allowing to reuse the same lexical resource in different domains;
• while lexical units are characterized by lexical properties (hyponymy, hyperonymy, fuzzynymy, etc.), an ontological layer provides more detailed description of domain-specific concepts (i.e. Buyer, Seller, etc.) and semantic relations between concepts (i.e. has_object_role, has_agent_role, etc.).

A previous experience to approach the multilingual EU scenario for information retrieval tasks is the LOIS project whose aim was the creation of a lexical resource for the legal field: it is composed by about 35,000 concepts in six European languages (Dutch, English, German, Portuguese, Czech, and Italian, linked by English) [15].

The LOIS database contains terms extracted from EU Directives on the consumer protection domain; it uses the semantics of WordNet and EuroWordNet, which is centered around the notion of “synset”. A Synset is a set of one or more uninflected word forms (lemmas) with a synonymous meaning: for example action, trial, proceedings, law suit form a noun synset because they can be used to refer to the same concept. Because each synset denotes a particular meaning, polysemous words (with more than one meaning) occur in more than one synset. A synset is often further described by a gloss, explaining the meaning of the concept.

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

Cross-lingual linking is based on equivalence relations between synsets from the individual language wordnets and English synsets, which function as the pivot. These relations denote a.o. complete equivalence, near equivalence, or equivalence as a hyponym or hypernym. The network of equivalence relations in LOIS through the pivot Inter-Lingual-Index [7], determines the interconnectivity of the indigenous wordnets.

The structure of the LOIS database can be sufficient for cross-lingual retrieval tasks. When using this resource for different purposes, as for example for legislative drafting, more detailed views of concepts and their domain-specific relations might be needed at language or jurisdiction-independent levels.

Legal drafting of European law in particular can be effectively supported by a jurisdiction-independent representation of legal concepts and their relationships in order to obtain a view of the actors involved in a specific domain to be regulated, the roles they play in the same domain, the relationships with other actors, which can support the activity of regulating a typical situation, for example a transaction on the Internet between a consumer which wants to buy a good or a service from a specific type of seller. This information is not included in a semantic

---

4 created within the European project LOIS (Legal Ontologies for Knowledge Sharing, EDC 22161, 2003-2006)
multi-lingual lexicon, which, on the other hand, contains language dependent lexical manifestations of such concepts as well as lexical relations. What is needed therefore is a “distinction between conceptual modeling at a language-independent level and a language and culture specific analysis and description of discourse-related units of understanding” [16].

3. Design of the DALOS resource

As discussed, DALOS is targeted to providing a knowledge resource for legislative drafting. The identification of the scenario of use is particularly important because it contributes to identify the characteristics of the knowledge to be described and to avoid the common tendency to indiscriminately mix domain knowledge and knowledge on the process for which it is used (drafting, reasoning, searching, etc.). Such a mixing prevents knowledge representations from being automatically reusable outside the specific context for which the knowledge representation was originally developed [17].

The peculiarity of the “legislative drafting” activity can be identified in its function of norms creation on specific domains. In this context the use of an ontology is of primary importance. Laws in fact usually contain provisions [18], which deal with entities, expressed by lexical units, but they do not provide any general information on them: for example, the Italian privacy law regulates the behaviour of the entity “Data controller” who is the owner of a set of personal data, but such law does not give any additional information on this role and the relationships with other entities within the considered domain [19]. A formalized description of entities using an ontology of the domain to be regulated allows to obtain such additional information.

For the DALOS resource we want to avoid that the knowledge to be used as support for the activity of norms creation on a specific matter is mixed with the knowledge on the general process of drafting which, obviously, is matter independent (see also [19]). Therefore what is needed as a support for legislative drafting is a knowledge and linguistic resource giving a description of a domain to be regulated (domain knowledge), namely concepts of such domain, as well as their lexical manifestations in different languages.

For the aim of developing a project pilot, the “consumer protection” domain has been chosen, for its particular interest as a regulatory environment, which influences the competitiveness of EU Member States businesses and their ability to grow and create jobs.

In the first phase of the project we addressed the provision of the specification for the DALOS resource, in particular the definition of its Knowledge Organization System (KOS).

As discussed in Section 2, the DALOS resource is organized in two layers of abstraction (see Fig. 1):

- the Ontological layer containing the conceptual modeling at a language-independent level;
- the Lexical layer containing lexical manifestations in different languages of concepts at the Ontological layer.
Figure 1. Knowledge Organization System (KOS) of the DALOS resource.

Basically, the Ontological layer acts as a layer that aligns concepts at the European level, independently from the language and the legal order, where possible. Moreover, the Ontological layer allows to reduce the computational complexity of the problem of multilingual term mapping (N-to-N mapping). Concepts at the Ontological layer act as a “pivot” meta-language in a N-language environment, allowing the reduction of the number of bilingual mapping relationships from a factor $N^2$ to a factor $2N$. Concepts at the Ontological layer are linked by taxonomical as well as object property relationships (has_object_role, has_agent_role, has_value, etc.).

On the other hand, the Lexical layer aims at describing language-dependent lexical manifestations of the concepts of the Ontological layer. At this level lexical units are linked by pure linguistic relationships (hypernymy, hyponymy, meronymy, etc.).

The connection between these two layers is represented by the relationship between concepts and their lexical manifestations:

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

In the DALOS KOS this link is represented by the hasLexicalization relationship.

The use of a two-level knowledge architecture provides a higher degree of modularity to the knowledge resource:

1. the Lexical layer is a lexical database which can be upgraded and reused in different domains with respect to the one considered in DALOS;
2. the Ontological layer provides a more detailed semantic description of the lexical units at the Lexical layer, as well as relationships between concepts.
Entries and relationships at both levels are described by exploiting the expressiveness of RDF/OWL semantic Web standards. The combination of these two levels of knowledge allows synsets at the Lexical layer to obtain two different kinds of properties:

- **linguistic properties**, namely EuroWordNet lexical relations that come directly from a linguistic and statistical analysis of texts;
- **semantic properties**, coming from the classification of synsets into classes of the Ontological layer, describing the consumer protection domain for the pilot case; such semantic properties are valid only within this particular domain.

For example, a statistical analysis of the texts on consumer protection law revealed that *consumer* and *supplier* are related lexical units, therefore, in our model, at the Lexical layer they have been linked by a *fuzzynym* WordNet property. At the Ontological layer such relation can be more semantically characterized as regards the consumer protection domain: such lexical units are considered as lexicalizations of the classes *Consumer* and *Supplier* respectively, while the relations between them pass through the concept *Commercial_transaction* which links *Consumer* and *Supplier* by a *has_agent_role* property (see Fig. 1).

As discussed in Section 2, for European law drafting purposes the use of the Ontological layer can provide law-makers with a detailed view of a specific scenario, providing support in identifying actors, roles and relationships involved in a particular situation to be regulated. On the other hand, for national law drafting implementing EU directives, a view on language-dependent lexical manifestations of concepts at the Ontological layer is necessary to guide legal drafters in choosing the most appropriate term to express a specific concept (see for example the term *right of withdraw* which in Italian legislation has been translated into two ways: *recesso* and *risoluzione*; in these cases the DALOS service may support the drafter in choosing the right translation, providing information on the contexts in which the different translations are used as well as on the related frequency of use).

4. Phases of the DALOS resource implementation

The DALOS ontological-linguistic resource is implemented by three main activities:

1. Semi-automatic term extraction on the domain of consumer protection law from a set of selected texts using NLP tools (Lexical layer implementation);
2. Manual ontology construction on the “consumer protection” domain (Ontological layer implementation);
3. Connection between Ontological layer and Lexical layer by the *hasLexicalization* property implementation.

The first activity (implementation of the Lexical layer) is carried out using different NLP tools: in particular GATE\(^5\), owned/provided and maintained by

\(^5\)General Architecture for Text Engineering
the Department of Computer Science of the University of Sheffield, specifically used in this task to process English and Dutch texts, as well as T2K\(^6\), jointly developed by CNR-ILC and University of Pisa, tailored to process Italian texts.

The second activity (implementation of the Ontological layer) is an intellectual one, which aims at describing a scenario to be regulated.

The third activity (connection between the two knowledge levels) is implemented by an intellectual activity targeted to provide links between concepts at the Ontological layer and synsets at the Lexical layer by the \texttt{hasLexicalization} property.

5. Lexical layer implementation

Hand–crafted lexical and ontological resources usually need to be continuously extended and refined in order to incorporate up–to–date knowledge. Term acquisition from legal texts, based on NLP techniques, can play an important role in this process of implementing lexical resource and suggesting concepts for the Ontological layer on the basis of significant terminological entries. Terms acquisition from legal texts, based on NLP techniques, can assist this process (see, for instance, [20], [21], [22] [23]) by suggesting lexical relations. In the DALOS project we concentrate on the semi–automatic implementation of a lexical resource by means of the acquisition of terminological knowledge from texts belonging to the “consumer protection” in Italian and English for the pilot case (extensions to other languages, in particular Dutch and Spanish, are currently under development).

DALOS Lexical layer is built up by using NLP tools on a data set composed by 16 directives (9 original versions; 7 consolidated texts) and 42 judgements (33 from the Court of Justice, 9 from the Court of First Instance)\(^7\) on the domain of interest.

5.1. Term Extraction from Italian Texts

To implement the Italian version of the DALOS Lexical layer we used T2K (Text–to–Knowledge), a hybrid ontology learning system combining linguistic technologies and statistical techniques [24].

T2K does its job in two basic steps:

1. extraction of domain terminology, both single and multi–word terms, from a document base;
2. organization and structuring of the set of acquired terms into proto–conceptual structures, namely
   - fragments of taxonomical chains, and
   - clusters of semantically related terms.

As far as term acquisition is concerned, a pipeline of computational tools for linguistic analysis (AnIta [25]) is used to extract candidate terms from texts. These linguistic processing modules are in charge of:

---

\(^6\)Text-to-Knowledge

\(^7\)see the DALOS corpora at \url{http://www.dalosproject.eu}, “Documentation” section.
1. tokenisation of the input text;
2. morphological analysis (including lemmatisation);
3. syntactic parsing, articulated into two different steps:
   (a) chunking (including morpho-syntactic disambiguation);
   (b) dependency analysis.

Candidate terms may be one word terms (“single terms”) or multi-word terms (“complex terms”). The acquisition strategy differs in the two cases.

Single terms are identified on the basis of frequency counts in the shallow parsed texts, after discounting stop-words. The acquisition of multi-word terms, on the other hand, follows a two-stage strategy. First, the syntactically chunked text is searched for on the basis of a set of chunk patterns. Chunk patterns encode syntactic templates of candidate complex terms: for instance, adjectival modification (e.g. *organizzazione internazionale* ‘international organisation’), prepositional modification (e.g. *commercializzazione di autovetture* ‘marketing of cars’), including more complex cases where different modification types are compounded (e.g. *commercio di prodotti fitosanitari* ‘trade of fitosanitary products’). Secondly, the list of acquired potential complex terms is ranked according to their log-likelihood ratio [26].

In T2K recognition of longer terms is carried out by iterating the extraction process: acquired complex terms are projected back onto the original text and the acquisition procedure is iterated on the newly annotated text. The method proves helpful in reducing the number of false positives consisting of more than two chunks [27]. The iterative process of term acquisition yields a list of candidate single terms ranked by decreasing frequencies, and a list of candidate complex terms ranked by decreasing scores of association strength. The selection of a final set of terms to be included in the TermBank requires some threshold tuning, depending on the size of the document collection and the typology and reliability of expected results.

In what follows we exemplify what we discussed so far with preliminary results obtained from the DALOS consumer law corpus, including Directives, Regulations and case law on ‘protection of consumers’ economic and legal interests, for a total of 284,795 word tokens. With the best configuration we obtained a TermBank of 2,074 terms, namely 174 single and 1,900 multi-word terms, which is currently being evaluated by domain experts.

Table 1 contains a fragment of the automatically acquired TermBank. For each selected term, the TermBank reports its prototypical form (in the column headed “Term”), its frequency of occurrence in the whole document collection, and the lemma of the lexical head of the chunk covering the term (see column “Lemmatised headwords”). The choice of representing a domain term through its prototypical form rather than the lemma (as typically done in ordinary dictionaries) follows from the assumption that a bootstrapped glossary should reflect the actual usage of terms in texts. In fact, domain-specific meanings are often associated with a particular morphological form of a given term (e.g. the plural form). This is well exemplified in Table 1 where the acquired terms headed by *diritto* ‘right’ can be parted into two groups according to their prototypical form: either singular (e.g. *diritto di revoca* ‘termination right’) or plural (e.g. *diritti*
Table 1. An excerpt of the automatically acquired TermBank

d’autore ‘copyrights’). It should be noted, however, that reported frequencies are not limited to the prototypical form, but refer to all occurrences of the abstract term.

5.2. Italian term organization and structuring

In the second extraction step, proto–conceptual structures involving acquired terms are identified. We envisage two levels of conceptual organization. Terms in the TermBank are first organized into fragments of head–sharing taxonomical chains, whereby commercio dei medicinali ‘trade of medicines’ and commercio elettronico ‘electronic trade’ are classified as co–hyponyms of the general single term commercio ‘trade’. In this way, single and multi–word terms are structured in vertical relationships providing fragments of taxonomical chains such as the one reported below:

applicazione
dell’articolo
della direttiva
della legge
tariffa
delle disposizioni
delle sanzioni
applicazione
delle sanzioni amministrative
applicazione
delle sanzioni previste
del presente decreto
del regolamento

where the acquired direct and indirect hyponyms of the term applicazione ‘enforcement’ are reported. In this example, it can be noticed that terms sharing the head only are the direct hyponyms of the root term. Further hyponymy levels
can be detected when two or more multi-word terms share not only the head but also modifiers, as in the case of the *applicazione delle sanzioni amministrative* ‘enforcement of administrative sanctions’ with respect to the more general term *applicazione delle sanzioni* ‘enforcement of sanctions’. The number of extracted hyponimic relations from the DALOS corpus is 911 referring to 172 hyponym terms.

The second structuring step performed by T2K consists in the identification of clusters of semantically related terms which is carried out on the basis of distributionally-based similarity measures [28]. For each term (both single and complex) in the TermBank, we extracted a set of 1.071 semantically related terms referring to 238 terminological headwords.

In what follows, clusters of semantically related terms are exemplified:

- **disposizioni** ‘provision’
  - norme, disposizioni relative, decisione, atto, prescrizioni
- **legge** ‘law’
  - regolamento, protocollo, accordo, statuto, amministrazioni comunali
- **pubblicità ingannevole** ‘misleading advertisement’
  - pratiche commerciali, procedimento, pubblicità comparativa, clausole abusive, pubblicità
- **cmv** (comitato per i medicinali veterinari) ‘committee for veterinary medicines’
  - comitato, cpmp (Committee for Proprietary Medicinal Products), commissione, membri, consiglio

It should be appreciated that in these clusters of semantically related words different classificatory dimensions are inevitably collapsed; they include not only quasi-synonyms (as in the case of *disposizioni* ‘provision’ and *norme* ‘regulations’), hypernyms and hyponyms (e.g. *comitato* ‘committee’ and *cmv* (comitato per i medicinali veterinari) ‘committee for veterinary medicines’), but also looser word associations. As an example of the latter we mention the relation holding between *legge* ‘law’ and *amministrazione comunale* ‘municipal administration’, or between *comitato* ‘committee’ and *membri* ‘members’.

### 5.3. Term Extraction from English Texts

To implement the English version of the DALOS Lexical layer two term extraction applications are used. The overall focus of the extraction effort was on nominal term candidates. The set will be further expanded in the future with verbal term candidates.

The first application, TermExtractor [29] offers a comprehensive package of algorithms for the selection of relevant terms from any text corpus. It extracts a list of “syntactically plausible” term candidates (e.g. compounds, adjective-nouns, etc.), and determines the termhood on the basis of two entropy-based measures: Domain Relevance and Domain Consensus, which are used to select only the terms which are relevant to the domain of interest and consensually referred throughout the corpus documents. Domain Relevance is computed with reference to a set of

---

contrastive terminologies from different domains. It filters term candidates using Lexical Cohesion, which measures the degree of association of all the words in a terminological string. Domain Consensus computes the overall significance of the term candidates across documents. Furthermore, it allows various other settings, e.g. lemmatization and exclusion of nested terms. Tab. 2 below illustrates the result from the TermExtractor tool. The weight is the overall score (between 0 and 1) computed on the basis of a weighted mean of the scores for domain relevance, domain consensus and lexical cohesion.

<table>
<thead>
<tr>
<th>term</th>
<th>weight</th>
<th>domain relevance</th>
<th>domain consensus</th>
<th>lexical cohesion</th>
</tr>
</thead>
<tbody>
<tr>
<td>member state</td>
<td>0.945</td>
<td>1.000</td>
<td>0.863</td>
<td>1.000</td>
</tr>
<tr>
<td>free movement</td>
<td>0.754</td>
<td>1.000</td>
<td>0.797</td>
<td>0.330</td>
</tr>
<tr>
<td>official journal</td>
<td>0.747</td>
<td>1.000</td>
<td>0.722</td>
<td>0.545</td>
</tr>
<tr>
<td>protection of consumers</td>
<td>0.731</td>
<td>1.000</td>
<td>0.783</td>
<td>0.139</td>
</tr>
<tr>
<td>financial service</td>
<td>0.729</td>
<td>1.000</td>
<td>0.729</td>
<td>0.288</td>
</tr>
<tr>
<td>internal market</td>
<td>0.728</td>
<td>1.000</td>
<td>0.673</td>
<td>0.477</td>
</tr>
<tr>
<td>council directive</td>
<td>0.722</td>
<td>1.000</td>
<td>0.705</td>
<td>0.346</td>
</tr>
<tr>
<td>national legislation</td>
<td>0.713</td>
<td>1.000</td>
<td>0.762</td>
<td>0.060</td>
</tr>
<tr>
<td>natural person</td>
<td>0.712</td>
<td>1.000</td>
<td>0.673</td>
<td>0.412</td>
</tr>
<tr>
<td>community level</td>
<td>0.711</td>
<td>1.000</td>
<td>0.740</td>
<td>0.116</td>
</tr>
<tr>
<td>national law</td>
<td>0.707</td>
<td>1.000</td>
<td>0.672</td>
<td>0.282</td>
</tr>
<tr>
<td>personal data</td>
<td>0.704</td>
<td>1.000</td>
<td>0.677</td>
<td>0.314</td>
</tr>
<tr>
<td>interest of consumers</td>
<td>0.698</td>
<td>1.000</td>
<td>0.712</td>
<td>0.111</td>
</tr>
</tbody>
</table>

Table 2. TermExtractor Results

The second application, TermRaider, has been developed by the University of Sheffield in GATE. GATE is an architecture, a framework and a development environment for Language Engineering (LE) applications, which support efficient and robust text processing.

GATE uses NLP based techniques to assist the knowledge acquisition process for ontological domain modelling, applying automated linguistic analyses to create ontological knowledge from textual resources or to assist ontology engineers and domain experts by means of semi-automatic techniques.

For term extraction from the DALOS corpus, the following actions are performed within GATE:

- Tokenization and sentence splitting divide up the text into manageable units
- Part of speech tagging
- Lemmatization yields the citation form for each word form
- A multi word unit grammar defines the sequences of part of speech tags that constitute noun phrases
- The computation of term frequency/inverted document frequency (TF/IDF) [31], a technique widely used in information retrieval and text mining taking into account term frequency and the number of documents in the collection, computes the saliency of term candidates for each document

http://www.gate.ac.uk
• All term candidates with a TF/IDF score higher than an empirically determined threshold are selected.

The part of speech sequences defined by the grammar are exemplified in Tab. 3. The threshold for the tf-idf score was experimentally set to 5:

<table>
<thead>
<tr>
<th>Phrase</th>
<th>Term</th>
<th>tf/idf score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single nouns</td>
<td>consumer</td>
<td>10.3</td>
</tr>
<tr>
<td>Multiple noun</td>
<td>credit agreement</td>
<td>7.5</td>
</tr>
<tr>
<td>combinations:</td>
<td>distance sales</td>
<td>10.2</td>
</tr>
<tr>
<td></td>
<td>package travel contract</td>
<td>11.2</td>
</tr>
<tr>
<td>Noun-preposition-</td>
<td>code of conduct</td>
<td>8.8</td>
</tr>
<tr>
<td>-noun combinations</td>
<td>protection against victimization</td>
<td>11.2</td>
</tr>
<tr>
<td></td>
<td>seller of goods</td>
<td>10.1</td>
</tr>
</tbody>
</table>

Table 3. Part of speech sequences

The combination of TermExtractor and GATE enables us to obtain more reliable term candidates (where both agree, termhood is very likely), and GATE allows us to extend our techniques to other languages such as Dutch, which forms part of DALOS. For further processing such as determining relations between terms, GATE gives us much more in terms of NLP potential for later analysis of the terms in their textual environment.

The overall result of the English term extraction is a set combining the TermExtractor and tf/idf results. This set contains 2664 terms, of which 833 are multi word units, and 1831 are single nouns.

5.4. English term organization and structuring

The conceptual structuring of newly acquired terms exploits various linguistic properties of the candidate terminology. First, the analysis of the internal structure of term candidates enables the creation of links between term candidates themselves, or between term candidates and existing terminology from the LOIS database. For nominal phrases, modifiers are separated from their heads, and a hyponymic relation is automatically established if both the full phrase and the head are part of the term candidate set or the LOIS terminological vocabulary. For example, “circulation of information” has “circulation” as hypernym, and “arrangement for payment” is a subconcept of “arrangement”. Further, identity relations between modifiers of multi-word units and term candidates indicate thematic relations between the two terms, as in the case of “reference” and “reference period” on the one hand, and “mutual recognition” and “mutual recognition procedure” on the other. In addition, lexicosyntactic patterns (so-called Hearst patterns, e.g. “such as”, “including”, “and other” [32]) are indicative of semantic relations. When applied to the legal texts, they define hyponymic relations between candidates. For instance, the definition: “‘total cost of the credit to the consumer’ means all the costs, including interest and other charges, which the consumer has to pay for the credit;” supports the creation of a hyponymic link between the term candidates “interest” and “charge”. Our intention is to extend these techniques in the near future with the extraction of verbal relations between term candidates and statistical collocational information.
6. Ontological layer implementation

The Ontological layer of the DALOS resource is aimed at providing an alignment of concepts at language-independent level. It acts not only as a pivot structure for language-dependent lexical manifestations, but it provides a more semantically characterized description of the chosen domain in terms of concepts and their relations, exploiting the expressiveness and reusability of the RDF/OWL semantic Web standards for knowledge representation. This allows also to validate the developed knowledge resource with respect to existing foundational or core ontologies.

As discussed in Section 4 the Ontological layer is the result of an intellectual activity aimed at describing the domain of the consumer protection, chosen for the pilot case. An intellectual approach has been chosen to strictly reach the project objectives with a predictable degree of reliability in describing semantically qualified relations between concepts.

Classes and properties have been implemented on the basis of the terminological knowledge extracted from the chosen Directives on the consumer protection law (see Section 5), in particular from the “definitions” contained, maintaining coherence to the design patterns of the Core Legal Ontology (CLO) developed on top of DOLCE foundational ontology [33] and on the “Descriptions and Situations” (DnS) ontology [34] [35] within the DOLCE+ library. The DALOS ontology copes with the entities of the chosen domain and their legal specificities. In this knowledge architecture the role of a core legal ontology is to provide entities/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 such as, in our case, DOLCE.

As regards domain-specific concepts, the DALOS Ontological layer is designed to stress the distinction identified by the “Descriptions and Situations” ontology, extended by CLO within the legal domain, between intensional specifications like norms, contracts, roles, and their extensional realizations in the same domain as cases, contract executions, agents. This distinction is correlated by the so called Norm ↔ Case CODeP design pattern [36]. According to the Norm ↔ Case CODeP, intensional specifications like norms use tasks, roles, and parameters, while extensional realizations like legal cases conform to norms when actions, objects and values are classified by tasks, roles, and parameters respectively. The matching is typically performed when checking if each entity in a legal fact is compliant to a concept in a legal description [36].

The distinction stressed by DALOS is strictly linked to the activity of legislative drafting addressed by the project. Apart from more technical provisions like ‘amendments’ on existing norms, legislative drafting can in fact be considered as an activity that creates norms on generic situation descriptions, qualifying them by, for example, deontic terms [11]. Speaking according to CLO, this activity deals with description (intensional specifications) of generic situations (also called “situational frameworks” in [11]), giving them a normative perspective. For

---

10 http://www.loa-cnr.it/ontologies/CLO/CoreLegal.owl
11 DOLCE+ library, http://dolce.semanticweb.org
12 Conceptual Ontology Design Pattern
example the Directive 97/7/EC of 20 May 1997, at Art. 7 paragraph 1 states that “Unless the parties have agreed otherwise, the supplier must execute the order within a maximum of 30 days from the day following that on which the consumer forwarded his order to the supplier”: it states that, unless differently agreed, the generic situation in which the supplier executes an order to the consumer, following a consumer request, is obliged, and this obligation has to be satisfied within a maximum of 30 days from the consumer request.

A normative perspective of generic situations is the result of the legislative drafting activity; it results in legislative text paragraphs grouped in articles, which can be semantically qualified as provisions [18], namely fragments of a regulation (for example an obligation for a role towards a task).

A support to legislative drafting can therefore include: 1) a taxonomy of provision types able to give a normative perspective to generic situations; 2) a knowledge resource supporting the description of generic situations in a specific domain, as well as giving an ontological perspective to entities involved in such situations [37]. The DALOS Ontological layer aims at representing this second kind of knowledge resource, tailored for the consumer protection domain pilot case.

![Figure 2. Excerpt of the DALOS Ontological Layer.](image)

The Ontological layer is therefore populated by the conceptual entities which characterize the consumer protection domain. The first assumption is that all concepts defined within consumer law are representative of the domain and, as a consequence, that several concepts used in the definitional contexts pertain to the ontology as well, representing the basic properties or, in other words, the ‘intensional meaning’ of the relevant concepts. Similarly, the Ontological layer contains generic situations having a legal relevance in the chosen domain.

Such domain-specific concepts are classified according to more general notions, imported from CLO, such as Legal role and Legal situation. Examples
of some concepts obtained by the definitions from the consumer law domain are Commercial_transaction, Consumer, Supplier, Good, Price. The specific roles they play ([35]) are illustrated in Fig. 2.

On the other hand, the main entities derived from CLO are axiomatized, disjoint classes, characterized by meta properties, such as Identity, Unity and Rigidity. The most relevant distinction is between Roles (anti-rigid) and Types (which are rigid). Roles, according to [35], are anti-rigid since they are “properties that are contingent (non-essential) for all their instances”. Types on the other hand can play more roles at the same time. For instance, a legal subject (either a natural or artificial person) can be a seller and a buyer. Domain-specific requirements are expressed by restrictions over ontological classes, for instance by defining Consumer as a role that can be played by Natural_person only.

The first version of the DALOS Ontological layer contains 121 named classes with necessary & sufficient definitions, resulting in the OWL-DL language.

7. The application prototype

An application prototype has been developed within the project in order to show how the DALOS resource can be accessed and exploited to provide multilingual lexical and semantic support in legislative documents drafting concerning the “consumer protection” domain.

xmLegesEditor is an open source legislative drafting environment developed at ITTIG-CNR [38] for supporting the adoption of legal national standards (XML and URN NIR standards). Briefly, xmLegesEditor is a visual XML editor able to support legislative drafters in the production of standard compliant normative documents, providing advanced features for structural and semantic markup as well as user-friendly tools for constructing persistent hyperlinked normative references [39] [40]. xmLegesEditor is one of the three regulation-drafting environments being evaluated in the SEAL project (Smart Environment for Assisting the drafting and debating of Legislation) [14]. See [41] [42] for a comparison with the other environments.

The DALOS extension of xmLegesEditor provides integrated access from the drafting environment to the knowledge resource produced in DALOS.

As discussed in Section 6, the DALOS resource is provided in the W3C standard format for semantic resources RDF/OWL. This allows easy and flexible integration in an application through specifically suited software libraries. Jena [43] is the most popular framework for the development of semantic Web applications written in Java. It provides, among many other features, high level methods for accessing and manipulating RDF/OWL resources. Its use inside xmLegesEditor, also written in Java, allows flexible access to the DALOS resource. It is possible for example to upgrade the Lexical layer with lexical units in different languages, in a dynamic and transparent way.

---

13 NormeInRete, http://www.normeinrete.it
14 SEAL is a project in the e-Participation initiative of the European Commission. See http://www.eu-participation.eu/seal
7.1. Accessing the DALOS resource within the application prototype

The DALOS version of xmLegesEditor provides additional panels for accessing the DALOS knowledge resource. Different kinds of access have been considered following its multiple layer architecture. They are implemented in several application panels (Fig. 3).

The **Ontology Browser** panel presents to the user the Ontology hierarchy in a tree view allowing the browsing through the ontological classes down to the classified terms. This provides the drafter with a smart overview of the domain ontology and the related term classification.

Lexical units grouped in synsets according to the strategy described in previous sections, is also accessible from a plain list **Terms** panel, from which the user can obtain all the domain terms in the chosen language, or express a textual query over the lexical forms of the extracted synsets specifying simple query parameters. Defined terms (in the domain corpus) are highlighted in the list for their particular relevance in preventing ambiguity and favoring harmonization (see [44]).

**Details** and **Sources** panels present to the user general lexical information on the selected synset. In particular term definitions (if present), as its textual fragment from the domain corpus, and a list of different variants (in a WordNet sense) are shown in the Details panel. The Sources panel shows the list of hyperlinks to external document fragments in the domain corpus where the selected term has been defined and where it has been used, as well as its variants (Fig. 4). This provides the drafter with easy access to the contexts of use and definitions of each
term in the existing legislation, obtaining a valuable feedback on its pertinence in a current context.

Multilingual support is managed by offering to the user the possibility to choose the main language, i.e. the language in which the lexical units are shown within the editor. It is possible to change the main language by reloading the whole knowledge resource inside the editor. In the Details panels it is then possible to look at the corresponding definition, relations and sources of the selected synset in one of the other supported languages provided by the term alignments.

In the Lexical Properties panel (Fig. 4), the linguistic relations established in the Lexical layer, based on the EuroWordNet semantic model, are dynamically accessed from the OWL resources and presented to the user. In this panel WordNet relations like “hyponymy” and “fuzzynymy” (as “Related Term”) for the selected synset, with respect to other synsets in the selected language, are shown.

Similarly, the Semantic Relations panel shows semantic relations of the selected synset with other synsets inferred from their classification in the domain ontology as described in Section 3. These are dynamically inferred by accessing the OWL resources through the reasoning engine provided by the Jena library and presented to the user in a hierarchical view.

Figure 4. Additional DALOS panels

7.2. Using the DALOS resource in the application prototype

The integration of xmLegesEditor with the DALOS knowledge resource is completed by providing functions for the interaction with the editing panels. Once a synset is selected, it is possible to choose one of its lexical variants as a normalized term to be inserted in the text being drafted at the cursor position, or to markup the selected word with a reference to the chosen synset.

The inserted term will be highlighted in the text as being taken from the DALOS resource as well as marked up in the underlying XML document with a reference to the DALOS resource it belongs to. This means that the text in-
cludes information on term origin. For example, by opening a document within xmLegesEditor and clicking on the highlighted terms, all the previously described panels are populated with several types of information about the selected term or concept. This provides a self explaining view of the document. It is to be noticed that the DALOS term markup, though implemented in the prototype application for Italian standard (NormeInRete standards [39]) compliant documents, is not constrained to any particular format and can be applied to different XML standards, using either xmLegesEditor for DTD or XMLSchema compliant XML documents, or different XML editors after having integrated the DALOS linguistic-ontological resource.

7.3. Using the DALOS resource to improve the quality of legislative texts

The integration of the DALOS resource into xmLegesEditor provides users with facilities which aims at enhancing the quality of legislative documents in different scenarios, which can be grouped into three main working situations:

1. new legislative documents drafting;
2. existing documents checking;
3. transposition of European directives into national laws.

The first situation deals with drafting new provisions; in this case the user has the possibility to browse the Ontological layer, identifying concepts which are relevant for a situation to be regulated, as well as clusters of related concepts and semantic relationships which give a view of the actors and their relationships on the situation to be regulated. Going down to the Lexical layer the user may choose, in a specific language, the more appropriate lexical manifestations of such concepts, checking also the contexts in which they are used within the European legislation (Source panel), and insert the chosen lexical unit, with a proper XML annotation, within the text. Similarly, the user may search if a term is included in the Lexical layer and which is his conceptual description at the Ontological layer.

The second situation deals with checking the terminological accuracy of an existing text. The system may check all the lexical units of the text also contained in the DALOS Lexical layer, as well as verifying their pertinence to the context, and, in case, substitute them with more appropriate terms.

The third situation deals with facilities able to guarantee terminological and conceptual coherence in the transposition of European legislation into national laws. Usually European directives use general concepts and terminology to better cope with different Member States legal cultures; the use of the same term in a transposition law and in the transposition language version of the related directive might not be appropriate for a national legal culture and legislation (an example is the case of the term ‘electronic signature’, used in the Directive 1999/93/EC, and terms used in the Italian legislation (D. Lgs. 7 March 2005, n. 82) where three variants are distinguished: a more general concept of ‘firma elettronica’ (electronic signature), and more specific ones as ‘firma elettronica qualificata’ (qualified electronic signature) and ‘firma digitale’ (digital signature)). By using the DALOS resource the legal drafter can be supported in choosing the more appropriate concept and terminology to be used within the transposition law.
8. Conclusions

The main purpose of the DALOS project is to provide law-makers with linguistic and knowledge management tools to be used in the legislative processes, in particular within the phase of legislative drafting. The aim is to keep control over the legal language, especially in the EU legislation multilingual environment, enhancing the quality of the legislative production, as well as the accessibility and alignment of legislation at European level.

In this paper we presented the DALOS resource, which is organized into two knowledge layers (the Ontological and Lexical layers). The motivations of this kind of architecture and the methodologies for its implementation have been presented. In particular, we have discussed the principle on the basis of which the Ontological layer has been developed, as well as NLP techniques used to implement the Lexical layer. Finally we have illustrated the use of the DALOS resource in the xmLegesEditor legislative drafting environment, along with facilities aiming at enhancing the quality of legislative texts.

As future developments the extension of the Lexical layer to Dutch and Spanish is being carried on, as well as methodologies to build automatically or semi-automatically the DALOS Ontological layer can be investigated. Similarly xmLegesEditor can be adapted to serve as the interface for modification of the DALOS knowledge resource, for example to locally enrich the lexicon or to classify terms, to correct or align terms in different languages and, in general, to modify the knowledge resource, which can then later be submitted to a community of experts for its dynamic upgrade according to users’ contributions.

The availability of a document archive marked up with a vocabulary of normalized terms derived by DALOS modules can also be useful in documents indexing to provide Semantic Web-oriented retrieval services. Moreover, as terms in XML texts will be linked to the Ontological layer through the Lexical layer, it will be possible to provide more advanced query features exploiting semantics for extracting norms or document fragments using more complex retrieval inferences.

The application prototype (the integrated environment composed by the drafting tool and the knowledge resource) is under test and evaluation by legislative offices of the Italian Parliament and CNIPA\textsuperscript{15}. Following the extension to Dutch and Spanish, the system will be also tested by other public administration users in the Netherlands and Spain.

References


\textsuperscript{15}Italian Minister of Reforms and Innovations in Public Administration / National Center for Information Technology in Public Administration
ies based on ontologies,” in *Proceedings of the V Legislative XML Workshop*, pp. 161–172, Euro

legislation,” in *In Proceedings of the International Conference on Artificial Intelligence


1998.


Modeling for Advanced Application Domains*, vol. 3289, Springer Berlin / Heidelberg,
2004. ISSN 0302-9743 (Print) 1611-3349 (Online).


legal concepts,” in *Proceedings of the Workshop on Legal Ontologies and Artificial In
elligence Techniques* (P. Casanovas, M. Biasiotti, E. Francesconi, and M. Sagri, eds.),
pp. 43–63, 2007. ISSN 1613-0073, CEUR Workshop Proceedings, ISSN 1613-0073, online

*Law and the Semantic Web (Benjamins, Casanovas, Breuker, and Gangemi, eds.),* Springer
Verlag, 2005.


tologies with dolce,” in *Proceedings of the 13th International Conference on Knowledge
Engineering and Knowledge Management (EKAW02)* (C. M. A. O. L. S. A. Gangemi, N.


[16] K. Kerremans and R. Temmerman, “Towards multilingual, termontological support in

ricore, two core ontologies for law.,” in *In Proceedings of EKAW Workshop on Core

LEGONT* ’97, 1997.


based web information retrieval system,” *Lecture Notes in Computer Science*, vol. 3369,
pp. 185–200, 2005.

[22] S. Walter and M. Pinkal, “Automatic extraction of definitions from german court deci
sions,” in *Proceedings of the COLING-2006 Workshop on Information Extraction Beyond

[23] P. Cimiano, “Ontology learning and population from text,” in *Algorithms, Evaluation and

una piattaforma linguistico-computazionale per l’estrazione di conoscenza da testi,” in


The e-Sentencias Prototype: A Procedural Ontology for Legal Multimedia Applications in the Spanish Civil Courts

Pompeu CASANOVAS a, Xavier BINEFA b, Ciro GRACIA b, Emma TEODORO a, Núria GALERA a, Mercedes BLÁZQUEZ d, Marta POBLET e, Jordi CARRABINA c, Marius MONTON c, Carlos MONTERO c, Javier SERRANO c and José Manuel LÓPEZ-COBO d

a Institute of Law and Tecnology
UAB, Spain

b Digital Video Understanding Group
ETSE, UAB, Spain
{Ciro.Gracia, Xavier.Binefa}@uab.cat

c Laboratory for HW/SW Prototypes and Solutions (CEPHIS)
ETSE, UAB, Spain

d Intelligent Software Components (iSOCO)
Madrid, Spain
{ozelin, mercedes}@isoco.com

e ICREA Researcher at the Institute of Law and Tecnology
UAB, Spain
marta.poblet@uab.cat

Abstract. Search, retrieval, and management of multimedia contents are challenging tasks for users and researchers alike. We introduce a software-hardware system for the global management of the multimedia contents produced by Spanish Civil Courts. The ultimate goal is to obtain an automatic classification of images and segments of the audiovisual records that, coupled with textual semantics, allows an efficient navigation and retrieval of judicial documents and additional legal sources. This paper describes our knowledge acquisition process, sets a typology of Spanish Civil hearings as performed in practice, and a preliminary procedural ontology at its actual stage of development (e-Sentencias ontology). A discussion on procedural, contextual and multimedia ontologies is also provided.

Keywords. Semantic Search, Civil Courts, Legal Multimedia, Legal Procedural Ontology, Multimedia Ontology, Contextual Ontology, HW/SW Acceleration Platforms, Reconfigurable Devices, Speaker Diarization, Video Segmentation

Introduction

Legal professionals are used to consume an important part of their time searching, retrieving, and managing legal information. However, the recent explosion of multimedia
legal contents has resulted in rising costs and requires more management capacities than ever before. Improving the functionalities for search, retrieval, and management of multimedia legal documents is paramount to fully unlock the potential of those contents for legal practice and to develop specific management solutions for different profiles of legal users: lawyers, judges, prosecutors, court clerks, etc.

In Spain, the Civil Procedure Act of January 7th, 2000 (1/2000) introduces the video recording of oral hearings. As a result, Spanish civil courts are currently producing a massive number of multimedia files that have substituted the written transcripts and have become part of the judicial file, together with suits, indictments, injunctions, judgments and pieces of evidence. Lawyers, prosecutors and judges need to access these contents when preparing similar cases or, if necessary, when appealing to superior courts. To date, nevertheless, there is no available system within the judicial domain to automatically index, tag, or annotate those audiovisual contents.

Furthermore, the provisions made by the 1/2000 Civil Procedure Act for the video recording of civil proceedings in Spain do not include a protocol establishing how to obtain audiovisual records. Rather, and since an ever growing number of Autonomous Communities in Spain hold competences on the organization of the judicial system, there is a plurality of standards, formats, and methods to produce audiovisual records. As a result, analogical and digital standards coexist with different recording formats. The support in which copies are provided to legal professionals (i.e. to prepare an appeal) may also consist of VHS videotapes, CDs or DVDs. In addition, the procedures to store, classify, and retrieve audiovisual records may vary even from court to court, with no common database available to store the audiovisual records.

The use of ontologies applied to legal multimedia was already faced in the e-Court Project [12]. The aim of the e-Sentencias Project is to develop a software-hardware system for legal professionals to access and manage the multimedia files connected to their cases across different platforms, formats, and languages. The project includes the development of both an ontology-based meta-search engine and a specific web-based, reconfigurable hardware platform capable to accelerate and improve the quality of the document retrieval process. The objectives of this approach are: (i) to facilitate search, discovering and management of multimedia legal contents based on their meaning; (ii) to optimize search results by automatically linking video-to-text; (iii) to improve the organization of the search memory, and (iv) to save time to users. The e-Sentencias Project, in sum, combines legal and contextual ontologies with algorithmic techniques. In this regard, e-Sentencias involves technologies such as the Semantic Web, ontologies, NLP techniques, audio-video segmentation, and IR. The ultimate goal is to obtain an automatic classification of images and segments of the multimedia records that, coupled with textual semantics, allows the efficient navigation and retrieval of judicial documents and additional legal sources.

Section 1 of the paper offers a brief typology of civil procedures covered by e-Sentencias. In Section 2 we present an overview of the knowledge acquisition process towards the construction of a conceptual structure to classify video segments and the development of legal ontology applications. Section 3 describes the development of the e-Sentencias ontology. Section 4 sets the discussion on our research goals focusing on the state of the art of procedural, contextual and multimedia ontologies. Section 5 shows the structure and architecture of the video system prototype at the present stage of research and, finally, we conclude in section 6 by offering some conclusions and expected results.
1. Civil Procedures in Spain

The main inputs of e-Sentencias are multimedia files that contain hearings of civil cases [vista oral]. The 1/2000 Act establishes two basic procedures or “declarative processes”: the ordinary proceedings and the verbal proceedings (see Figs 1 & 2). The main differences between the two procedures lie both in the value of the case—more or less than €3000, respectively—and the legal object at dispute.

The steps of civil processes also vary depending on the specific procedure. On the one hand, ordinary proceedings [juicio ordinario] start with a separate, independent motion called preliminary hearing [audiencia previa] to resolve pre-judiciary issues (documents, evidences to be accepted, etc.), while in the verbal proceedings [juicio verbal] all steps take place at the same judicial event. On the other hand, the claim of the plaintiff is contested in written terms in the ordinary proceeding, while in the verbal proceeding is replied orally.

2. Knowledge Acquisition Methodology

The figures below represent the steps of civil processes as formally set by the 1/2000 Act. However, the routines and constrains of daily practice in court result in a far more complex typology, revealing interesting differences between the formal provisions of the law and the actual development of the hearings. Thus, we identified 14 different types of proceedings populating the ordinary/verbal division. Therefore, a new typology of

![Figure 1. Steps of the process in ordinary proceedings [juicio ordinario].](image-url)
hearings has to be superimposed over the original division. This fact adds some complexity to the conceptualization of the hearings, because each particular judgement has to be classified as (i) preliminary hearing, ordinary proceeding, verbal proceeding, or precautionary/provisional measure and (ii) within one type of the proposed typology.

To proceed with the knowledge acquisition process, we represented the new typology as a set of workflows able to contain (and interact with) other structures of practical knowledge (i.e. sub-typologies, concepts, keywords, content groups, etc). The basic units of the workflows are ‘procedural stages’. Stages are the unities in which 1/2000 Act segments civil procedures and they are also the basic video sequences of the e-Sentencias prototype. In workflows, procedural stages include: (i) actions (rhombus) which imply to move forward or backward to another stage; (ii) procedural legal concepts, (iii) and content of legal concepts (boxes). An example of such a workflow is shown in Fig. 4.

At this point, we followed two different knowledge acquisition strategies. First, we set up a focus group composed of professional attorneys, judges and solicitors. We used standard eliciting techniques (e.g. interview, discussion, and brainstorming) to extract from their professional knowledge the conditions of use and estimated frequencies of expressions and legal concepts as they appear in recorded hearings. As a result of that, we elicited a list of roughly 900 terms and legal practical expressions, and we linked them to the general typology of 14 oral hearings and to concrete subjects contained in each type. Figure 5 shows this kind of work for one of the legal processes (subtypes).
Figure 3. Typology of civil hearings as emerged in daily practice.

Figure 4. General structure of the hearing in ordinary proceedings [juicio ordinario].
Figure 5. Estimated frequency of terms associated with hearings.

The second strategy focuses on the transcripts of the hearings. We transcribed a small set of fifteen hearings marking their different steps and coding manually legal concepts (i.e. 'injunction', 'cause of necessity', 'deed', etc.), legal expressions (i.e. 'with the permission of your Honor'), and practical procedural rules that are implicit in the video sequences, such as the following piece shows:

```xml
<actor name="judge" tc="00.01.30">    
   Let us see mr. *** DEFENDANT STANDS UP AND
   APPROACHES TO THE MICROPHONE come to the
   microphone [PROCEDURAL FORMULA, EXCLUSIVE USE
   BY THE JUDGE]
</actor>
<actor name="defendant" tc="00.01.31">    
   yes
</actor>
<actor name="judge" tc="00.01.38">    
   and answer the questions that both attorneys are going to
   formulate, starting by the attorney of the plaintiff [GENERAL
   RULE: IF BOTH PARTIES HAVE REQUESTED
   EXAMINATION, THE PLAINTIFF'S ATTORNEY ALWAYS
   COMES FIRST IN EXAMINING THE DEFENDANT, AND
   THEN CONTINUES THE DEFENDANT'S ATTORNEY].
</actor>
```

Figure 6. Transcription of a Civil Hearing.

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

To facilitate both the transcription and annotation processes we set up an annotation tool
interface (Fig. 7). A second list of terms was then extracted, and a third list of concepts was built by convention according to the following parameters: (i) speech markers, (ii) pragmatic markers, (iii) cognitive markers.

Speech markers are of sociolinguistic nature (pitch on word, pitch on phrases, short pause, emphatic stress, overlapping speech, rising shift in intonation ...). Pragmatic markers are of discursive and rhetoric nature (opening speech, ending expression, ending pause, competitive talk, spacing out ...). Cognitive markers are of communicative (link of subjects to reference and co-reference phenomena: metaphor, anaphora, catafora) and emotional nature (stress, plea, anxiety, surprise ...).

This methodology comes from the fields of linguistic, discourse and cognitive analysis, reconstructing the main specific frameworks in which situated meaning is produced [17,31]. We may notice here that there is no need to use all the results obtained through this kind of annotation work to build up the final ontology. Using discourse, pragmatic and cognitive markers is only a way to clarify and organize the experts’ knowledge on these particular legal frameworks. However, this is crucial in order to identify the most common patterns of meaning used in each type of civil proceeding, avoiding misinterpretations.

This expert linguistic knowledge allows grasping the main features needed to build the procedural cognitive contexts in which ontologies may add semantics to the further diarization of the multimedia content. Diarization refers to the process of audio stream segmentation. We will be come back to it later on.

It may be noticed by now that MPEG-7 descriptors which are designed primarily to describe low-level audio or visual features such as color, texture, notion, audio energy, and so forth, or attributes of AV content such as location, time, quality and so forth, have been loosely followed to interpret the main features of the speech [29]. On the contrary, higher-level AV features such regions, segments, objects and events have been substituted for the cognitive and emotional features of the actors link to the content of the hearing.
3. The e-Sentencias Ontology

We have proceeded to model the e-Sentencias ontology on the basis of the procedural knowledge practically used in the Court hearings. This procedural knowledge is what we have been labeling as “professional legal knowledge” (PLK) [13,14,15,16]. PLK is the kind of knowledge which is produced through the daily practice and behavior of professional people. In this case, as a result of the knowledge acquisition process, we obtained several lists of concepts, expressions and terms ordinarily used and shared by judges, lawyers, court secretaries and clerks in the civil procedures. We organized them following the discourse and cognitive guidelines described so far, but bearing in mind the needs and requirements of the final users of the system.

To date, the e-Sentencias ontology consists of 16 classes, 22 attributes and 169 instances. The ontology provides a model of the following aspects:

- The hearings as found in the civil jurisdiction practice and their different phases, stages and actions.
- The agents who take part in each one of the phases of the hearings (lawyer, plaintiff, judge, prosecutor, etc.).
- The contents dealt with in each one of the modelled hearings.

We may identify the different phases the hearings in different proceedings consist of, bearing in mind that within each phase we have identified an order of precedence among the phases. This order of precedence does not take place within a hearing, but rather, and depending on the different actions, the hearing jumps into another phase. For this purpose, we have created the classes Phase and Hearing in the ontology, where a hearing is characterized by a set of phases and agents involved and by the fact that it is related to a certain type of civil-jurisdiction proceedings. An example of an instance within the hearing is the one shown in Fig. 9.

Furthermore, the type of claim is also modeled (or else the issues dealt with) within the class hearing, which make sense within the course of a hearing, as can be seen in the right part of Fig. 9. As mentioned above, there is a relationship of precedence between the different phases a hearing consists of. For instance, Fig. 4 shows that the oral hearing of the ordinary proceedings has five phases: (i) opening of the hearing, (ii) ver-
...ification that the action is pending, (iii) evidence, (iv) final conclusions, (v) and end of the hearing. The hearing consists of a set of phases. We define the order of precedence within the transition stages. Figure 10 offers an example of instance of phase. In such class we determine to which hearings a certain phase relates to (in the example, the phase of evidence takes place in all hearings) and we determine also what sub phases the phase is divided into (the phase of evidence consists of three sub phases: carrying out the evidence, admitting the evidence and putting forward the evidence).

Another aspect which has been taken into account when modeling the ontology are the specificities identified during the hearings, as, although there is an order of precedence of the phases within the hearing, there are certain points where we change from one phase to the other, that is to say, there is a diagram of stages. A clear example can be seen in the prior graphic (Fig. 4), where, for example, if there is coincidence between the phases we pass directly from phase 2 to checking the subsistence of proceedings in phase 5 or end phase of the hearing. These transitions between phases have been modeled in the class StageTransition. Another example can be found in Fig. 11.

In the stages of transition we may identify the initial stage, the final stage, the hearing where the transition takes place, as well as the related action. Furthermore, we have also identified a series of sub actions which may or not occur during the transition. Thus, in the period for submitting the evidences [proposición de prueba], different sub actions may occur, such as the oral and contradictory reports of the experts, the interrogatories made to the witnesses, etc., which may or not be carried out within this period.

Finally, this ontology is used in combination with a set of thesaurus of synonyms where different expressions related to one instance are specified, for example, for the
instance of the phase ‘Statements of the Claim’, the following synonyms have been defined:

- Statement of the claim of the party
- This party ratifies the statement of claim
• This party ratifies the written statement of the claim
• I reaffirm again
• I ratify
• I confirm the written statement of defence
• I ratify the written defence
• I require these proceedings
• I petition these proceedings to be admitted to evidence
• Pleas provided as grounds for the petition
• Challenging of the document
• Statement regarding a plea of the other party
• Put forward an exception

These synonyms have a common trait: they are all expressions which are formulae used by the agents involved in the proceedings and will thus allow us to carry out the semantic searches in the audiovisual context by processing the audio.

4. Discussion: Procedural, Contextual and Multimedia Ontologies

In practice the use of ontologies for different tasks and purposes requires to consider the particular task as context for the ontology. The reason is that ontologies are often not really designed independent of the task at hand [21]. In general, the context of use has an impact on the way concepts are interpreted to support certain functionalities. As some aspects of a domain are important in one context but do not matter in another one, an uncontextualized ontology does not necessarily represent the features needed for a particular use. In order to solve this problem, we have to find ways to enable the representation of different viewpoints that better reflect the actual needs of the application at hand.

When talking about viewpoints, we can distinguish two basic use cases. In the first case, the aim is to provide means for maintaining and integrating different existing viewpoints. In the second use case, one may want to extract a certain viewpoint from an existing model that best fits the requirements of an application.

In many application domains it is acknowledged that the creation of a single universal ontology is neither possible nor beneficial, because different tasks and viewpoints require different, often incompatible conceptual choices. As a result, it may be required to support situations where different parties commit to different viewpoints that cannot be integrated by imposing a global ontology. This situation demands for a weak notion of integration, in order to be able to exchange information among the viewpoints [40]. Stuckenschmidt describes one of such examples from oncology. Oncology is a complex domain where several specialties, e.g. chemotherapy, surgery, and radiotherapy are involved in a sequence of treatment phases, each representing a particular viewpoint.

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

Researchers on contextual ontologies use to define ‘context’ as local (not shared with other ontologies) and opposed to content ontologies themselves (shared models of a domain). ‘It can be argued that the strengths of ontologies are the weakness of contexts and vice-versa’ [11]; also, [21]. In this specific sense, contextual ontologies have to deal
at least with three problems: the directionality of information flow, local domains and context mapping.

Directionality of information flow means keeping track of the source and the target ontology as specific piece of information; local domains means giving up the hypothesis that all ontologies are interpreted in a single global domain; context mapping means that two elements (concepts, roles, individuals) of two ontologies, though extensionally different, are contextually related, e.g., because they both refer to the same object in the world [11]. The e-Sentencias ontology faces these three problems, and it is being modeled within the workflow of each type of hearing. Thus, it is context-sensitive. We modeled the directionality of the informational flow as transitional states of the different steps of the oral hearings. We treated the specific domain of modeling an ordinary proceeding (juicio ordinario) as a local domain.

Regarding the context mapping, we are dealing with two separate ontologies with a different scope. The aim of the first one is to model judicial professional knowledge – the OPIK ontology we have been developing since the European Project SEKT [18]. The second one models the procedural steps and development of the Spanish verbal and ordinary hearings, including all their phases and some common elements contextually related (e.g., the agents involved in the hearings). Figure 12 shows the contextual relations of transitional stages for the opening of the hearings [apertura].

Strictly speaking, multimedia ontologies take a different point of view onto the context. Context is defined currently in a more cognitive way as 'the set of interrelated conditions in which visual entities (e.g. objects, scenes) exist' [23,25]. This grounds the strategy of the direct vs. indirect exploitation of the knowledge base to annotate the content of the videos, using visual and content descriptors alike. ‘The main idea of our approach lies in a way to associate concepts with instances that are deemed to be prototypical by their annotators with regard to their visual characteristics’ [9].

Figure 12. Transitional stage graph (apertura).
This definition of context entails a theoretical approach in which ‘actions and events in time and space convey stories, so, a video program (raw video data) must be viewed as a document, not a non-structured sequence of frames’ [36,37]. In such an approach, visual low level features, object recognition and audio speaker diarization – the process of partitioning the audio stream in homogenous segments and clustered according to speaker identity – are crucial to analyze e.g. a sport or movies sequences. The distance between low-level features and their possible interpretation is known as ‘the semantic gap’.

Several approaches to this perspective problem and the so-called ‘semantic gap’ may be found in the recent literature: (i) multi-context ontologies vs. mono-context ontologies [1,2,20]; (ii) low-level descriptors [pixel color, motion vectors, spatio-temporal relationships] vs. semantic descriptors [person, vehicle ...] [33,3,9]; (iii) modal keywords of perceptual concepts [aural, visual, olfactory tactile, taste] vs. content topics [24,25]; (iv) cross-media annotation [19].

In the e-Sentencias ontology we have not properly taken a pure multimedia perspective because the audiovisual documents that are recorded in Spanish courtrooms do not convey actions, but legal narratives. Motion and color are generally uniform, since they are not considered the relevant aspect of those documents. Thus, court records are technically very poor (see Fig. 13), filmed using a one-shot perspective (the camera is situated above and behind the judge, who never appears on the screen). Rather than telling a story, the video structures a single framework in which a story is referred, conveyed and constructed by the procedural actors (judge, counsels, testimonies, secretary, and court clerks).

Here lies the layered exophoricty of the legal discourse. Exophoricty means that what is said in court refers to other scenarios and ambiences situated outside the court. However, actions, events and stories are referred as well to a contextually embedded discourse, procedurally-driven, and hierarchically conducted by the judge (judge-centered). Therefore, a strong delay is produced between audio and video as sources of information.

Figure 13. Image quality.
A legal court video record would be completely useless without the audio, because we may only infer procedural (but not substantial) items from the motion. What is important is what is said in court, not what is done. Visual images are only ancillary related to the audio stream. This is an important feature of the records, which has to be taken into account in the task of building an ontology, because what the different users require (judges, lawyers, citizens) is the combination of different functionalities focused on the legal information content (legislation quoted, previous cases and judgments -precedent-, personal professional records, and so on). This is the reason for a hybrid user-centered approach that is the kernel of our theoretical approach.

The e-Sentencias ontology is focused on civil proceedings as they are really performed into the Spanish Courts, it takes into account the different contexts embedded into the workflows of each type of hearing, and it is addressed to multimedia content. But it is not a properly contextual or multimedia ontology, but a procedural one.

5. Structure and Architecture of the e-Sentencias Prototype

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

The structure of the application is based on two intuitive and semantically powerful metaphors: the oral hearing line and the oral hearing axis. The oral hearing line presents a timeline divided into segments. Each segment represents a different speech, produced by one of participants in the process: judge, secretary, attorneys, witnesses, etc. Each participant is represented by a different color to obtain an identification at first glance of their interventions. Therefore, it is possible to visualize specific contents of the video by merely clicking on a particular colored sequence. Moreover, it is possible to add textual information to any instant of the intervention.

The oral hearing axis consists of a column representing the different phases of the event as defined by procedural legislation. Different phases (as opening statements, presentation of evidences, concluding statements, etc.) are represented by different colors, allowing a quick access. It is also possible to access to legal documents related to each phase (i.e. pieces of evidence such as contracts, invoices, etc.) as well as to jurisprudence quoted in the oral hearing and detected through phonetic analysis. This legal information is also structured in directories and folders.

As Fig. 14 shows, the user interface is divided into two main parts: the upper part contains the video player, the oral hearing axis and the oral hearing line. The lower part is devoted to external information layers (i.e. references to articles, documents annexed, manual annotations, links to jurisprudence, etc.). This part is divided into two tabs. The first one contains important information of the selected phase, allowing the addition of the different documents presented during the phase. The second tab contains historical information of the process and all the related information available in advance.
The main functionalities offered in the upper part of the user interface are:

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

2. The oral hearing line: the timeline of sequences and interventions assigned to the different actors of the process. One single sequence of the video may contain interventions of different actors. Therefore, sequences may be either mono-colored (intervention of one single part) or multi-colored (more than one part intervening in the same sequence). The horizontal length of each segment of the timeline is proportional to its length in seconds. The application includes two modes of playing video, apart of the usual one. It is possible to select either the visualization of all the interventions by a single participant or, in turn, all the interventions on a given phase.

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

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

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

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

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

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

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

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

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

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

6. Conclusions and Expected Results

The bottom-up methods employed so far deserve a further discussion, since no clustering or semi-automatically ontology extraction tools have been used. And, yet, it is our contention that the e-Sentencias application can reach its maximum efficiency by automatically segmenting and annotating the video contents. However, this is a research goal that requires, first, a clear understanding of the conceptual content of the hearings. Therefore, we developed a research strategy in which keywords, ontologies and the user interface prototype would be worked out at the same time.

A second reason of this bottom-up approach lies on the very nature of the institutional conditions we are dealing with. Video records of Spanish courtrooms are neither officially transcribed nor digitally stored in large databases. Yet, most multimedia ontology extraction methods are applied to vast collections of digital images or video sets stored in libraries or large databases (Yahoo news, broadcasting TV, sports such as soccer). The knowledge acquisition and ontology building processes are therefore addressed to fill up the semantic gap between metadata, low-level (physical features) and high level descriptors (interpretation, human interaction) in order to perform a more effective classification, information retrieval or search and browsing within databases or websites.

Several techniques have been proposed: (i) Automatic Image Annotation (AIA) and Confidence Clustering (CC) [39]; (ii) fusing MPEG-7 visual descriptors [38]; (iii) pictorially and multimedia enriched ontologies [7,8]; (iv) salience and visualness calculation of multimedia entities [19]; (v) use of probabilistic grammar, Bayesian networks combined with semantic constraints [41]; (vi) inter and intra-contextual segments for video segment retrieval [32]; (vii) interoperability of OWL with the MPEG-7 MDS (MOREL ontology) [42].

In our case, however, lawyers and judges store the multimedia files obtained from the court office on individual basis (typically archiving them together with the case records). No written transcription of the hearing is provided, so that professionals have to exclusively rely on multimedia contents. Our main goal, then, is not focused on image, shot, event, scene, frame detection or recognition patterns, but offering a semi-automatic way to easily search each step or phase of the hearing in every single record.

These are the reasons why we adopted a bottom-up strategy, carefully transcribing a sample of each type of hearing, building from scratch the conceptual structure, and hand making a preliminary segmentation of the hearings. Some empirical studies show that "retrieval based on transcription of the speech in video data adds more to the average precision of the result than content-based retrieval" [22]. However, this leads to a kind of paradox, since the amount of videos to manage grows so large that it becomes impossi-
ble to process them manually. This is also the main management problem that lawyers, judges and court offices are facing with their multimedia files.

Therefore, once done the manual annotation process and built a preliminary procedural ontology, we propose to handle the problem by using speech recognition, which is an already mature technology in other domains (i.e. information services). Speech recognition technologies are being used in two different manners: (1) keyword spotting for the oral hearing axis and (2) speaker recognition for the oral hearing line [23].

The required metadata needed to draw the oral hearing axis are temporal marks at which each procedural phase of the judicial process starts or switches. Each process has its own steps, and each phase has its own actors. During these phases, all the actors start and end up their interventions with the same utterances, i.e. in Spanish, con la venia [approx. ‘thank you your Honor’], ‘no hay más preguntas Señoría’ [‘no more questions, your Honor’]. We state that by spotting these and similar utterances, we will be able to settle the starting and ending point at any step of the hearing.

This approach is simple, but relies on the effectiveness of the recognition of the keywords. In order to improve this effectiveness, we are training the acoustics models of the speech recognition system, nowadays based in a commercial motor (Loquendo), with data stemming from the videotaped hearings. We expect to recognize close to a 100% system will be trained with the whole data set.

On the other side, required metadata for the oral hearing line are the moments when actors start and finish. These moments are found by means of speaker identification techniques. This part of the research is still at an early stage and we do not have significant results yet. We should have in mind that speech recognition system cannot be completely trusted to obtain a perfect speech to text transcription yet [27].

In the e-Sentencias and other related projects we expect to obtain two different types of results. On the one hand, a fully annotated legal corpus of multimedia oral hearings classified in 14 procedural classes, as regulated by the 1/2000 Act. On the other hand, an operational system with a humancomputer interface, adapted to different user profiles, as described in this paper. Using the system prototype, the automatic capabilities of speaker interventions and phases detection will be tested against manually annotated corpus. It will also be evaluated in cross-oral hearings retrieval based on hardware accelerated and specifically implemented multimedia ontologies.

Acknowledgements

We carry out this work within the following projects: (i) e-Sentencias (e-Sentencias. Plataforma hardware-software de aceleración del proceso de generación y gestión de conocimiento e imágenes para la justicia) is a Project funded by the Ministerio de Industria, Turismo y Comercio (FIT-350101-2006-26). A consortium of: Intelligent Software Components (iSOCO), Wolters Kluwer España, IUAB Institute of Law and Technology (IDT-UAB), Centro de Prototipos y Soluciones Hardware - Software (CHEPIS-UAB) and Digital Video Semantics (Dpt. Computer Science UAB) and (ii) Metabuscador (Metabuscador semántico legal, clustering de resultados y autoclasiﬁcación de fuentes jurídicas), developed by the same team, is a Project funded by the Ministerio de Industria, Turismo y Comercio (FIT-350100-2007-161). The original legal ontology is being developed in the project OCI. Implementación de Iuriservice: FAQ inteligente de soporte para la Escuela Judicial española. SEJ2006-10695.
References


[33] Pertridis, K., Precioso, F., Athanasiadis, T., Avrithis, Y., Kompatatsiari, Y.: Combined Domain Specific and Multimedia Ontologies for Image Undertaking. 28th German Conference on Artificial Intelligence, Koblenz, Germany, September 11–14, 2005.


Authors’ Biographies

Rinke Hoekstra, Joost Breuker, Marcello Di Bello and Alexander Boer, “LKIF Core: Principled Ontology Development for the Legal Domain”

Rinke Hoekstra is researcher at the Leibniz Center for Law of the University of Amsterdam. His research interests include knowledge representation, ontology engineering, knowledge based reasoning and in general the application of these approaches to the Semantic Web and the legal domain. He is one of the founders of MetaLex, the main developer of the LKIF ontology, and is currently a member of the OWL Working Group of the W3C that develops OWL 2. He obtained his MSc in Artificial Intelligence from the University of Amsterdam in 2001.

http://www.leibnizcenter.org/information/people/rinke-hoekstra

Joost Breuker is professor in Computer Science & Law at the Leibniz Center for Law and also works at the HCSLab of the University of Amsterdam. His main areas of research are experimental cognitive psychology (text understanding: problem solving) and Artificial Intelligence (intelligent teaching systems (ITS), knowledge acquisition and engineering, artificial legal reasoning). His current work is primarily in the fields of methodologies and design patterns for building knowledge systems, artificial problem solving, in particular legal reasoning and ontologies, knowledge representation and the Semantic Web. He is one of the founders of the CommonKADS methodology and studied cognitive science at the University of Amsterdam.

http://www.leibnizcenter.org/information/people/ joost-breuker

Marcello di Bello is a graduate student at the Department of Philosophy of Stanford University. He received a MA in Logic at the Institute for Logic, Language & Computation of the University of Amsterdam in 2007.

http://www-philosophy.stanford.edu/fss/marcello.html
Alexander Boer is a senior researcher of the Leibniz Center for Law at the University of Amsterdam. His research interests include computational legal theory, legal ontology, e-Government, and generally the development of innovative concepts for the application of IT in the field of law. Drs. Boer is one of the founders of MetaLex, and chaired the technical committee of the MetaLex CEN/ISSS workshop for several years. He received his MA degree in social science informatics from the University of Amsterdam in 1998.

http://www.leibnizcenter.org/information/people/alexander-boer
Aldo Gangemi, “Introducing pattern-based design for legal ontologies”

Aldo Gangemi graduated in philosophy, and then worked in knowledge engineering and semantic technologies, including the Semantic Web as well as properly medical, legal, agricultural, and business applications. He’s senior researcher at the CNR Institute of Cognitive Sciences and Technology in Rome, and head of the Semantic Technology Lab. He also co-founded the Laboratory for Applied Ontology. His current research focuses on pattern-based ontology design, collaborative modeling, hybridation of NLP, lexical and semantic resources, and metamodels for heterogeneous knowledge integration. He has published more than 100 refereed articles in proceedings of international conferences, journals, and books, and has been involved in several projects funded by the European Union, governmental organizations, and industrial companies (either as research partner or consultant). He is responsible for CNR in the EU projects NeOn, IKS, and BONy, and has worked in the seminal ontology-related projects Galen, WonderWeb, OntoWeb, Metokis, and the Eureka project IKF. He has organized several conferences and workshops, and regularly tutors in PhD courses and international schools. In legally-related fields, he has organized workshops on regulatory and legal ontologies, and has made active research on reusable components for legal ontology design, including the Core Legal Ontology.

http://www.loa-cnr.it/gangemi.html
Alessandro Lenci, Simonetta Montemagni, Vito Pirrelli and Giulia Venturi, “Ontology learning from Italian legal texts”

**Alessandro Lenci** is a tenured researcher at the Department of Linguistics of the University of Pisa, where he teaches computational linguistics. He received a PhD in Linguistics in 1999 from the Scuola Normale Superiore in Pisa. He has extensively published in lexical semantics, computational lexicons and natural language processing. His current research themes focus on computational lexical semantics, cognitive modelling of semantic lexicon, statistical methods for natural language processing, distributional models of meaning.

**Simonetta Montemagni** is senior researcher at the Institute of Computational Linguistics of CNR in Pisa. She received a Ph.D. in Computational Linguistics from the University of Manchester Institute of Science and Technology (UMIST, UK) in 1995. Over the last several years, she contributed to more than 40 international and national projects and extensively published in the areas of linguistics and natural language processing. Her research interests include robust parsing techniques, acquisition of lexical information from electronic textual corpora and dictionaries, models of human language parsing and of language variation, Information Extraction from texts and Text Mining.

**Vito Pirrelli** received a laurea degree from the Linguistics Department of Pisa University (Italy) and a PhD in Computational Linguistics from Salford University (UK). Currently he is Research Director at the CNR Institute for Computational Linguistics in Pisa and teaches “Computer for Humanities” in Pavia. His research interests include Machine language learning, Word Processing and Computer Models of the Mental Lexicon, Information extraction and Theoretical Morphology.

**Giulia Venturi** received a laurea degree in 2006 from the Italian Linguistics Department of Pisa University (Italy). Currently she has a scholarship at the Institute of Computational Linguistics of CNR in Pisa (ILC-CNR). She is a
PhD student in Computational Linguistics at the University of Torino (Italy). Her main research interests include extraction of linguistic and extra-linguistic knowledge from domain-specific corpora and its organisation into Knowledge Organisation Systems such as domain ontologies, NLP-based analysis of legal language with a view to legal texts semantic processing.
Raquel Mochales-Palau and Marie-Francine Moens, “Automatic Argument Detection and Its Role in the Semantic Web”

Raquel Mochales Palau received the Master’s degree in computer engineering from the Universitat Politècnica de Catalunya (UPC), Barcelona, and the European Masters in Language and Speech from the TALP Research Center at the UPC, in 2006. She is currently a researcher in the Department of Computer Science at Katholieke Universiteit Leuven, Belgium. Her research interests include natural language processing, information extraction and argumentation analysis.

Marie-Francine Moens is an associate professor at the Department of Computer Science of the Katholieke Universiteit Leuven, Belgium. She holds a M.Sc. and a Ph.D. degree in Computer Science from this university. She leads a research team on text based information retrieval. Her main interests are in the domain of automated content retrieval from texts using a combination of statistical, machine learning and symbolic techniques, and exploiting insights from linguistic and cognitive theories. She is the author of many international publications and books with regard to topics such as information extraction from text, automatic indexing and summarization of text, and information retrieval and search models.
Stephan Walter, Manfred Pinkal, “Automatic Extraction of Definitions from German Court Decisions”

Stephan Walter is a researcher in Computational Linguistics at Saarland University. His research focuses on the semantics of legal language, and he is currently working on a Ph.D. thesis on the extraction of definitions from German court decisions.

Manfred Pinkal is full professor for Computational Linguistics at Saarland University. He has done extensive research in the areas of semantic processing, linguistically guided information access techniques, and spoken-language dialogue systems. He is co-founder of the Collaborative Research Center “Resource-adaptive Cognitive Processing” (funded by German Science Foundation), founding speaker of the International Research and Training Group “Language Technology and Cognitive Systems”, and co-founder and Principal investigator of the Excellence Cluster “Multimodal Computing and Interaction”. For his research in computational linguistics and language technology, he was awarded the Gottfried Wilhelm Leibniz Prize of the German Science Foundation in the year 2000.
Cassia Trojahn, Paulo Quaresma and Renata Vieira, “Mapping Core Legal Ontologies using an Extended Argumentation Framework based on Confidence Degrees”

**Cássia Trojahn** is a PhD student at Department of Informatics, University of Évora, Évora, Portugal. She works with question answering systems, ontology mapping approaches, and argumentation models for ontology mapping.

**Paulo Quaresma** is an associated professor at the Department of Informatics, University of Évora, Portugal, and a member of the CENTRIA – Artificial Intelligence Center of the New University of Lisbon. His main research interests are natural language processing and information retrieval, specially question answering systems, using symbolic and machine learning techniques.

**Renata Vieira** is an assistant professor at Faculdade de Informática, Pontifícia Universidade Católica do Rio Grande do Sul, Porto Alegre, Brazil. Her main research interests are related to natural language processing, agent communication and ontologies.
Roberto Garcia and Rosa Gil, “Copyright Licenses Reasoning using an OWL-DL Ontology”

**Roberto García** got his MSc. in Computer Science from the Universitat Politècnica de Catalunya (UPC) in 2000. Then, he completed a Master in E-Commerce and got a research assistant position at the Universitat Pompeu Fabra (UPF) where he concluded his PhD thesis in 2006 about a Semantic Web approach to DRM. He is now an assistant professor at the Universitat de Lleida (UdL), where he is combining his previous research lines with trying to get the Semantic Web in touch with “real-world” end-users. More details from http://rhizomik.net/~roberto

**Rosa Gil** got her degree in Physics from the Universitat de Barcelona in 1996. Then she completed a Master in Environment Management in 1997. She got a scholarship at the Universitat Pompeu Fabra (UPF) in 2000 and carried out her research in Semantic Web based multimedia and digital rights management. She completed her PhD thesis in 2005 about agents negotiating in the Semantic Web. She is now an associate professor at the Universitat de Lleida (UdL), where she is combining his previous research lines with trying to get the Semantic Web in touch with “real-world” end-users. More details from http://rhizomik.net/~rosa
Pompeu Casanovas, Nuria Casellas and Joan-Josep Vallbé, “An Ontology-Based Decision Support System for Judges”

**Dr. Pompeu Casanovas** is Professor of Philosophy of Law at the Autonomous University of Barcelona (UAB), Law School; Director of Advanced Research (ACQU), Consultant of Artificial Intelligence and Law at the Universitat Oberta de Catalunya (UOC); and Director of the UAB Institute of Law and Technology. He has conducted research in several institutes and universities (UCSD, OIISL, Stanford, Bologna). He has been principal investigator of over 35 national and international projects, and author of above 10 books and 100 scientific articles in the areas of legal philosophy, legal sociology, judicial studies, and AI and Law (the Semantic Web). He is the General Editor of the Research Series La Razón ñurea (Editorial Comares, Spain).

Contact him at: pompeu.casanovas@uab.cat

**Núria Casellas** is an assistant lecturer at the Law School (Department of Political Science and Public Law) of the Universitat Autònoma de Barcelona (UAB), and a researcher at the Institute of Law and Technology (IDT-UAB). Her research interests are focused on the application of artificial intelligence and knowledge management techniques to the legal domain, in particular legal knowledge acquisition and representation methodologies, and legal ontologies. She has participated in various national and European funded projects, such as SEKT (Semantically Enabled Knowledge Technologies) EU-IST 2003-506826, DALOS (Drafting Legislation with Ontology-Based Support) EU eParticipation 2006-01/024, and IURISERVICE (PROFIT-150500-2003-198 and SEJ2006-10695).

Contact her at: nuria.casellas@uab.cat

**Joan-Josep Vallbé** is an associate lecturer at the Constitutional Law and Political Science Department of the University of Barcelona, and works also as a researcher at the Institute of Law and Technology (IDT-UAB). His main research interests cover organizational decision analysis and its relation to artificial
intelligence, including expert knowledge acquisition techniques and user evaluation processes. He has been enrolled in various research projects, such as the Observatory of Judicial Culture (SEC2001-2581-C02-01), SEKT–Semantically Enabled Knowledge Technologies (EU-IST 2003-506826), DALOS–Drafting Legislation with Ontology-Based Support (VI Framework Programme, E-participation, contract 2006/01/024), and IURISERVICE (FIT-150500-2003-198 and SEJ2006-10695).

Contact him at: vallbe@ub.edu
Tommaso Agnoloni, Lorenzo Bacci, Enrico Francesconi, Wim Peters, Simonetta Montamegni and Giulia Venturi, “A two-level knowledge approach to support multilingual legislative drafting”

Tommaso Agnoloni received the Master Degree in Computer Science Engineering from Università di Firenze in 1998 and the Ph.D. in Systems Engineering and Computer Science in 2002 at the Computer Science Department (DSI), Università di Firenze. In January 2004 he joined the Italian National Research Council at the Institute of Legal Information Theory and Techniques (ITTIG-CNR), where currently he is a research assistant. His research interests include Legislative XML standards and formats, Legal Ontologies, Semantic Markup. He has been involved in various projects within the eParticipation framework (DG Information Society & Media) of the European Commission. He is coordinator of the development and one of the authors of xmLegesEditor an opensource visual XML Editor for the adoption of legal national standards.

Lorenzo Bacci received the Degree in Computer Science Engineering from University of Florence in 2005. In September 2005 he joined the Italian National Research Council at the Institute of Legal Information Theory and Techniques (ITTIG-CNR), where he currently works. His main research interests are: Knowledge Representation, Information Extraction, NLP, ontologies and Semantic Web.

Enrico Francesconi received the Ph.D. in Computer Science at the Computer Science Department of the University of Florence. In 2002 he joined the Institute of Legal Information Theory and Techniques of the Italian National Research Council (ITTIG-CNR), where currently he is a researcher. His main research interests include knowledge representation, legal ontologies, legal standards and semantic markup, legal drafting, artificial intelligence techniques for legal document classification and knowledge extraction. He is a member of the Italian and European working groups establishing XML and URI standards for legislation. He has been involved in various projects within the eParticipation
framework (DG Information Society & Media) of the European Commission and for the Office for Official Publications of the European Communities.

**Wim Peters** is a senior researcher working in the Natural Language Processing Group at the University of Sheffield (U.K.). He is interested in lexical and ontological structure, and has worked in various domains. In the legal domain he has been involved in a number of projects dealing with knowledge acquisition of EU and national legislation, and the ontological organization of this knowledge.

**Simonetta Montemagni** is senior researcher at the Institute of Computational Linguistics of CNR in Pisa. She received a Ph.D. in Computational Linguistics from the University of Manchester Institute of Science and Technology (UMIST, UK) in 1995. Over the last several years, she contributed to more than 40 international and national projects and extensively published in the areas of linguistics and natural language processing. Her research interests include robust parsing techniques, acquisition of lexical information from electronic textual corpora and dictionaries, models of human language parsing and of language variation, Information Extraction from texts and Text Mining.

**Giulia Venturi** received a laurea degree in 2006 from the Italian Linguistics Department of Pisa University (Italy). Currently she has a scholarship at the Institute of Computational Linguistics of CNR in Pisa (ILC-CNR). She is a PhD student in Computational Linguistics at the University of Torino (Italy). Her main research interests include extraction of linguistic and extra-linguistic knowledge from domain-specific corpora and its organisation into Knowledge Organisation Systems such as domain ontologies, NLP-based analysis of legal language with a view to legal texts semantic processing.
Authors’ Biographies

Pompeu Casanovas, Xavier Binefa, Ciro Gracia, Emma Teodoro, Nria Galera, Mercedes Blázquez, Marta Poblet, Jordi Carrabina, Marius Monton, Carlos Montero, Javier Serrano, and José Manuel López-Cobo, “The e-Sentencias prototype: A procedural ontology for legal multimedia applications in the Spanish Civil Courts”

Dr. Pompeu Casanovas is Professor of Philosophy of Law at the Autonomous University of Catalonia (UAB), Law School; Director of Advanced Research (ACQU), Consultant of Artificial Intelligence and Law at the Universitat Oberta de Catalunya (UOC); and Director of the UAB Institute of Law and Technology (http://idt.uab.cat). He has conducted research in several institutes and universities (UCSD, OIISL, Stanford, Bologna...). He has been principal investigator of over 35 national and international projects, and author of above 10 books and 100 scientific articles in the areas of legal philosophy, legal sociology, judicial studies, and AI and Law (the Semantic Web). He is the General Editor of the Research Series La Razón Urea (Editorial Comares, Spain).

Contact him at pompeu.casanovas@uab.es

Dr. Marta Poblet holds a Juris Doctorate from the Stanford University (2002) and a Masters degree in International Legal Studies (Stanford Law School, 2000). She graduated in Political Sciences and Sociology at the UAB (1994) and in Law at the UB (1992). She is a ICREA researcher at the Institute of Law and Technology (IDT) and a lecturer of sociology of law at the Autonomous University of Barcelona - Department of Sociology. Her research areas cover legal organizations, judicial systems, conflict resolution, ADR-ODR, and immigration phenomena related to those aspects. She has published over 20 scientific articles in Journals and books.

Contact her at marta.poblet@uab.es

Emma Teodoro graduated in Law at the UAB (1997) and holds a Master degree in Sociology of Law at the International Institute for the Sociology of Law (IISJ-Oñati, Spain, 1998-1999). She is currently enrolled in a PhD Program in Law at the University of Barcelona. She is a researcher at the IDT-UAB. She is the technical responsible of the Multimedia Laboratory of the IDT and organizes the filming, production and editing of digital video (ethnomorphic doc-
Emma Teodoro graduated in Law at the UAB (2007). She is a researcher at the IDT-UAB. Her research interests include ethnography of legal institutions and legal organizations, law and image and visual anthropology, and legal ontologies.
Contact her at emma.teodoro@uab.es

Nuria Galera graduated in Law at the UAB (2007). She is currently enrolled in a PhD Program in Law at the University of Barcelona. She is a researcher at the IDT-UAB. She had been participating in different projects regarding to developing legal ontologies as E-sentencias Project. She is currently doing project manager tasks on Online Dispute Resolution (ODR) and collaborating on the Llibre Blanc de la Mediaci a Catalunya (White Paper on Mediation in Catalonia).
Contact her at nuria.galera@uab.cat

Xavier Binefa received BSc degrees in mathematics from the Barcelona University and in computer engineering from the Universitat Autonoma de Barcelona. He received his PhD degree in computer vision from the Universitat Autonoma de Barcelona in 1996, where he is currently an associate professor in its Computer Science Department and heading its group of Digital Video Understanding. He has graduated three PhD Students and is the advisor of five more PhD candidates. His research interests are focused on Video Understanding and Statistical Pattern Recognition, oriented to the analysis, indexing, presentation and interaction with digital video. He has authored and co-authored 13 contributions on international journals and more than 50 international conferences. He has also been the research leader of 16 competitive projects, taking into account the competitive national and international projects with both public and private funding.
Contact him at xavier.binefa@uab.cat

Ciro Gracia received BSc degrees in Computer Science from the University Autònoma de Barcelona and a Msc degree in Computer Vision and Artificial Intelligence in 2008 from the same University. He worked during three
years in Catalonia Justice Department in the field of education and rehabilitation initiatives. His research interests include Pattern recognition related to Multimedia video Analysis and Indexing.

Contact him at ciro.gracia@uab.cat

Jordi Carrabina was born in Manresa, Catalonia in 1963. He graduated in Physics in 1986, and received M.Sc. (1988) and Ph.D. (1991) degrees in Microelectronics (Computer Science Program), at the University Autonoma of Barcelona. In 1986 he joined CNM, with a PhD grant where he was collaborator researcher until 1996. Since 1990, he is Associate Professor at the Microelectronics and Electronic Systems Department of the University Autonoma of Barcelona, heading the research group Embedded Computation in HW/SW Platforms and Systems Laboratory, and the CEPHIS technology transfer node from the regional IT network. Main interests are Embedded Platform-based design using Hardware Software Co-design and SoC/NoC Architectures facing integrated and reconfigurable devices, mainly oriented to multimedia applications. He has given courses in several universities in Spain, Europe and South America. He has been consultant for companies as Hewlet-Packard Barcelona, Mitsubishi Electric, Epson Electronics, Sony Europe and many SME in Spain.

Contact him at jordi.carrabina@uab.cat

Marius Monton was born in Barcelona, Catalonia in 1976. He graduated in Informatics Engineering in 2004, and received M.Sc. (2006) in Microelectronics (Computer Science Program), at the University Autonoma of Barcelona. Since 2003 is working at the research group Embedded Computation in HW/SW Platforms and Systems Laboratory, and the CEPHIS technology transfer node from the catalan IT network, and as Associate Professor at the Microelectronics and Electronic Systems Department of the University Autonoma of Barcelona. He is also working part time with Greensocs. Main interests are SystemC TLM modeling oriented to multimedia applications.

Contact him at marius.monton@uab.cat
Javier Serrano was born in Barcelona, Catalonia in 1964. He graduated in Computer Science in 1988, and received Ph.D. (1994) degree in Automatic Control (Computer Science Program), at the University Autonoma of Barcelona. In 1988 he started as research and teaching assistant at the same university. Since 1996, he is Associate Professor at the Telecommunications and System Engineering Department of the University Autonoma of Barcelona. He joined the CEPHIS technology transfer node from the regional IT network as project manager in 2005. Main interests are speech technologies and audio in embedded systems.

Contact him at javier.serrano@uab.cat

José-Manuel López-Cobo was born in 1972 in Granada (Spain). He holds a degree in Computer Engineering from the University of Málaga. He has been working since 1999 in Research & Commercial projects, in three companies. He worked at Atos Origin and iSOCO, being responsible for the management of the Semantic Web and Semantic Web Services related EC funded projects. Currently he works at XimetriX Network Thoughts as Research Manager where he is managing new projects regarding the Semantic Content Management product of the company. José-Manuel has over 30 publications in several books, journals and conferences.

Contact him at ozelin.lopez@ximetrix.com

Mercedes Blázquez Cívico has worked at iSOCO, Madrid, Spain, as a researcher since September 2000. She was graduated in Computer Science from the Polytechnic University of Madrid (UPM) in 1997. She obtained her MSC in Artificial Intelligence from UPM in November 2000. She is currently working on her PhD thesis in artificial intelligence at UPM on Knowledge Management and their applications. Her research activities include knowledge management and its applications, web 3.0 and its application on knowledge management. She is currently participating in different projects related with the application of semantics in bank innovation department and the developing of web 2.0 portals. Also, she has worked in different national research projects PROFIT: eSentencias related with the application of the Semantic Web in different legal domain data sources,
such as text, emails, video or audio, eLogistica, related with the application of semantics over a domain of logistic. Also she collaborated in the INES platform (the Spanish indicative about software and services). Also she has participated in the FP6 IP project SEKT (IST-2003-506826) as a contractor team leader during the whole duration of the project. Other national R&D projects are: PLATA, a platform to provide access to the different Web technologies Semantic Web and Web 2.0; Elcano, a Semantic Search Engine in the International Affairs sector; ONTO-H: annotation tool of cultural sources. She has published several conference/workshops papers on ontology engineering, semantic web, and knowledge management.

Contact her at Mercedes@isoco.com
# Subject Index

<table>
<thead>
<tr>
<th>Term</th>
<th>Page(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>argumentation detection</td>
<td>115</td>
</tr>
<tr>
<td>argumentation framework</td>
<td>133</td>
</tr>
<tr>
<td>argumentation structure</td>
<td>115</td>
</tr>
<tr>
<td>civil courts</td>
<td>199</td>
</tr>
<tr>
<td>common sense</td>
<td>21</td>
</tr>
<tr>
<td>computational linguistics</td>
<td>95</td>
</tr>
<tr>
<td>contextual ontology</td>
<td>199</td>
</tr>
<tr>
<td>copyright</td>
<td>145</td>
</tr>
<tr>
<td>decision support system</td>
<td>165</td>
</tr>
<tr>
<td>definition extraction</td>
<td>95</td>
</tr>
<tr>
<td>description logic</td>
<td>145</td>
</tr>
<tr>
<td>design pattern</td>
<td>53</td>
</tr>
<tr>
<td>digital rights management</td>
<td>145</td>
</tr>
<tr>
<td>document management</td>
<td>75</td>
</tr>
<tr>
<td>e-administration</td>
<td>3</td>
</tr>
<tr>
<td>e-court</td>
<td>3</td>
</tr>
<tr>
<td>e-discovery</td>
<td>3</td>
</tr>
<tr>
<td>e-governance</td>
<td>3</td>
</tr>
<tr>
<td>framework</td>
<td>21</td>
</tr>
<tr>
<td>globalization</td>
<td>3</td>
</tr>
<tr>
<td>HW/SW acceleration platforms</td>
<td>199</td>
</tr>
<tr>
<td>information overload</td>
<td>3</td>
</tr>
<tr>
<td>information retrieval</td>
<td>3</td>
</tr>
<tr>
<td>knowledge extraction from texts</td>
<td>75</td>
</tr>
<tr>
<td>knowledge management</td>
<td>165</td>
</tr>
<tr>
<td>knowledge representation</td>
<td>21</td>
</tr>
<tr>
<td>law ontologies</td>
<td>133</td>
</tr>
<tr>
<td>lawyers</td>
<td>3</td>
</tr>
<tr>
<td>legal concept</td>
<td>21</td>
</tr>
<tr>
<td>legal information systems</td>
<td>95</td>
</tr>
<tr>
<td>legal knowledge engineering</td>
<td>53</td>
</tr>
<tr>
<td>legal multimedia</td>
<td>199</td>
</tr>
<tr>
<td>legal ontology(gies)</td>
<td>3, 21, 165, 177</td>
</tr>
<tr>
<td>legal procedural ontology</td>
<td>199</td>
</tr>
<tr>
<td>licensing</td>
<td>145</td>
</tr>
<tr>
<td>LKIF</td>
<td>21</td>
</tr>
<tr>
<td>methodology</td>
<td>21</td>
</tr>
<tr>
<td>multilingual legal drafting</td>
<td>177</td>
</tr>
<tr>
<td>multimedia ontology</td>
<td>199</td>
</tr>
<tr>
<td>natural language processing</td>
<td>75</td>
</tr>
<tr>
<td>NLP techniques</td>
<td>177</td>
</tr>
<tr>
<td>ontological engineering</td>
<td>21</td>
</tr>
<tr>
<td>ontology</td>
<td>21</td>
</tr>
<tr>
<td>ontology design</td>
<td>53</td>
</tr>
<tr>
<td>ontology learning</td>
<td>75, 95</td>
</tr>
<tr>
<td>ontology matching</td>
<td>133</td>
</tr>
<tr>
<td>professional knowledge</td>
<td>165</td>
</tr>
<tr>
<td>reasoning</td>
<td>145</td>
</tr>
<tr>
<td>reconfigurable devices</td>
<td>199</td>
</tr>
<tr>
<td>semantic search</td>
<td>199</td>
</tr>
<tr>
<td>semantic web</td>
<td>3, 115</td>
</tr>
<tr>
<td>speaker diarization</td>
<td>199</td>
</tr>
<tr>
<td>user validation</td>
<td>165</td>
</tr>
<tr>
<td>video segmentation</td>
<td>199</td>
</tr>
</tbody>
</table>
This page intentionally left blank
# Author Index

<table>
<thead>
<tr>
<th>Author</th>
<th>Page(s)</th>
<th>Author</th>
<th>Page(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Agnoloni, T.</td>
<td>177</td>
<td>Lópe-Cobo, J.M.</td>
<td>199</td>
</tr>
<tr>
<td>Bacci, L.</td>
<td>177</td>
<td>Mochales Palau, R.</td>
<td>115</td>
</tr>
<tr>
<td>Binefa, X.</td>
<td>199</td>
<td>Moens, M.-F.</td>
<td>115</td>
</tr>
<tr>
<td>Blázquez, M.</td>
<td>199</td>
<td>Montemagni, S.</td>
<td>75, 177</td>
</tr>
<tr>
<td>Boer, A.</td>
<td>21</td>
<td>Montero, C.</td>
<td>199</td>
</tr>
<tr>
<td>Breuker, J.</td>
<td>3, 21</td>
<td>Monton, M.</td>
<td>199</td>
</tr>
<tr>
<td>Carrabina, J.</td>
<td>199</td>
<td>Peters, W.</td>
<td>177</td>
</tr>
<tr>
<td>Casanovas, P.</td>
<td>3, 165,</td>
<td>Pinkal, M.</td>
<td>95</td>
</tr>
<tr>
<td>Casellas, N.</td>
<td>165</td>
<td>Pirrelli, V.</td>
<td>75</td>
</tr>
<tr>
<td>Di Bello, M.</td>
<td>21</td>
<td>Poblet, M.</td>
<td>199</td>
</tr>
<tr>
<td>Francesconi, E.</td>
<td>3, 177</td>
<td>Quaresma, P.</td>
<td>133</td>
</tr>
<tr>
<td>Galera, N.</td>
<td>199</td>
<td>Serrano, J.</td>
<td>199</td>
</tr>
<tr>
<td>Gangemi, A.</td>
<td>53</td>
<td>Teodoro, E.</td>
<td>199</td>
</tr>
<tr>
<td>Garcia, R.</td>
<td>145</td>
<td>Trojan, C.</td>
<td>133</td>
</tr>
<tr>
<td>Gil, R.</td>
<td>145</td>
<td>Vallbé, J.-J.</td>
<td>165</td>
</tr>
<tr>
<td>Gracia, C.</td>
<td>199</td>
<td>Venturi, G.</td>
<td>75, 177</td>
</tr>
<tr>
<td>Hoekstra, R.</td>
<td>21</td>
<td>Vieira, R.</td>
<td>133</td>
</tr>
<tr>
<td>Klein, M.C.A.</td>
<td>3</td>
<td>Walter, S.</td>
<td>95</td>
</tr>
<tr>
<td>Lenci, A.</td>
<td>75</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
This page intentionally left blank
This page intentionally left blank
This page intentionally left blank
This page intentionally left blank
This page intentionally left blank
This page intentionally left blank