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

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

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

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

1. Introduction

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

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

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

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

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

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

---

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

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

1.1. Terms and Semantics

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

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

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

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

---

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

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

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

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

1.2. Legal Concepts[7]

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

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

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

---

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

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

1.3. Semantic Spaces in Law

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

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

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

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

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

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

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

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

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

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

[13] On the federal and the state level.

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

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

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

2.1. The Extendable Markup Language XML

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

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

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

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

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

2.2. Legal Ontologies

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

3. Two Practical Examples: N-Lex and EUROVOC

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

3.1. The Experimental N-Lex Project[26]

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

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

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

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

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

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

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

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

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

3.2. The EUROVOC-Thesaurus

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

4. Conclusions

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

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

---

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

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

References

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

Biagioli, C. et al. (eds.): Proceedings of the V Legislative XML Workshop (Florence 2006). European Press Academic Publishing, Florence 2007

Bing, J. (ed.): Handbook of Legal Information Retrieval. Elsevier, New York 1984

Breuker, J.A. et al.: Use and Reuse of Legal Ontologies in knowledge engineering and Information Management. In: Benjamins (eds.): Law and the Semantic Web. Lecture Notes in Computer Science Vol. 3369, Springer, Berlin et al. 2005, 36-64

Bydlinski, F.: Juristische Methodenlehre und Rechtsbegriff[2] *("Legal Methodology and Nomen Juris")*. Springer, Wien 1991

Fellbaum, C. (eds.): WordNet: An Electronic Lexical Database. MIT Press, MA 1998

Galinski, C.: Wozu Normen? Wozu semantische Interoperabilität? (*"Why Norms? Why Semantic Interoperability?"*). In: Pellegrini/Blumauer (eds.): Semantic Web: Wege zur vernetzten Wissensgesellschaft. Reihe X.media.press, Springer, Berlin et al. 2006, 47-72

Hendler, J. et al.: Integrating Applications on the Semantic Web. Journal of the Institute of Electrical Engineers of Japan Vol. 122(10), October 2002, 676-680

Hohfeld, W.N.: Fundamental Legal Conceptions as Applied in Judicial Reasoning. The Yale Law Journal Vol. 26/8 (June 1917), 710-770

Koivunen, M.-R./Miller, E.: W3C Semantic Web Activity. In: Proc. of the Semantic Web Kick-off Seminar (Helsinki 2001). HIIT Publications, Helsinki 2002/1, 27-43 (available at `http://www.w3.org/2001/12/semweb-fin/w3csw`)

Lesmo, L. et al.: The next EUR-Lex: What should be done for the needs of lawyers belonging to different national legal systems? In: Proc. of the JURIX EU-Info Workshop. Brussels 2005 (available at `http://www.di.unito.it/\ensuremath{\sim}guido/PS/jurixWorkshopPaper.pdf`)

Liebwald, D.: An Evaluation of "New EUR-Lex": All Tasks Achieved and All Problems Solved? MR-Int 3/2005 (European Media, IP & IT Law Review), Verlag Medien & Recht, Vienna, 156-160

Liebwald, D.: Einheitsschnittstellen zu Rechtssystemen am Beispiel von N-Lex (*"Unified Interfaces to Legal Systems Using the Example of N-Lex"*). In: Schweighofer et al. (eds.): Aktuelle Fragen der Rechtsinformatik 2007. Boorberg, Stuttgart et al. 2007 (in print).

Liebwald, D.: Evaluierung juristischer Datenbanken *("Evaluation of Legal Databases")*. Verlag Österreich, Vienna 2003.

Liebwald, D.: Semantische Räume als Strukturhintergrund der Rechtsetzung ("*Semantic Spaces as Structural Patterns of Legislation*"). In: Bildungsprotokolle der Kärtner Verwaltungsakademie zu den Klagenfurter Legistik-Gesprächen 2006, Klagenfurt 2007 (in print).

Matthijssen, L.: Interfacing between Lawyers and Computers: An Architecture for Knowledge-based Interfaces to Legal Databases. Kluwer Law International, The Hague et al. 1999

Sartor, G.: The ONE-LEX project and the informational unification of the laws of Europe. In: Proc. of the Klagenfurter Legistikgespräche 2005. Bildungsprotokolle Vol. 12, Kärntner Verwaltungsakademie, Klagenfurt 2006, 193-202.

Schacherreiter, J.: Legal culture und europäische Harmonisierung *("Legal Culture and European Harmonization")*. Juridikum 2006/1, Verlag Österreich, Vienna, 17-21

Schweighofer, E./Liebwald, D.: Advanced lexical ontologies and hybrid knowledge based systems: First steps to a dynamic legal electronic commentary. AI&Law Journal Vol. 15 (February 2007), Springer International, NL

Uschold, M.: Where Are the Semantics in the Semantic Web? AI Magazine 24(3): Fall 2003, AAAI Press, Menlo Park, 25-36

Uschold, M./Gruninger, M.: Ontologies: Principles, Methods and Application. The Knowledge Engineering Review 11(2)/1996, Cambridge University Press, 93-115

Wahlgren, P.: The Quest for Law. Jure AB, Stockholm 1999.

Warta, P.: Zauberworte – Verwandlungen des Gleichheitsgrundsatzes in der Judikatur des österreichischen Verfassungsgerichtshofes (*"Magic Words – Metamorphoses of the Principle of Equality in the Legal Practice of the Austrian Constitutional Court"*). In: Schweighofer et al. (eds.): Aktuelle Fragen der Rechtsinformatik 2005. Boorberg, Stuttgart et al. 2005, 576-583.