Projekt LOIS: Juristische Ontologien und Thesauri

Erich Schweighofer[*]*, Doris Liebwald*

Arbeitsgruppe Rechtsinformatik, Universität Wien
Universitätsstraße 2, 1090 Wien
erich.schweighofer@univie.ac.at
doris.liebwald@univie.ac.at

Schlagworte: Ontologien, Thesauri, elektronischer Kommentar, semantisches Netz

Abstract: Dieser Beitrag beschreibt das LOIS Projekt sowie dessen Weiterentwicklung zum Werkzeug für automatisierte Anwendungen wie dynamischer elektronischer Kommentar oder Verwaltungsapplikationen. Der im Rahmen des LOIS Projekts aufgebaute multilinguale Thesaurus dient nicht nur dem Retrieval, sondern ist auch Grundlage für den Aufbau einer umfassenden Rechtsontologie. Mit dem Projektentwurf einer Vernetzung des Thesaurus mit dem hybriden wissensbasierten System wird ein Weg zum Aufbau einer umfassenden Ontologie gezeigt.

1. Einleitung

Die benutzerfreundliche, stabile und effiziente Lösung juristischer Informationssysteme als Textspeicher, Archiv und Suchhilfe bedingt – wie auch im Internet – eine stärkere Fokussierung der Forschung auf semantische Indexierung. Das Schlagwort des *Semantic Web* [Berners-Lee et al 2001] lautet – übertragen auf den Rechtssektor – *Legal Semantic Indexing*. Dieses Thema ist Kern der Rechtsinformatik seit vielen Jahren: die Frage der besten und zweckmäßigsten Formalisierung des Rechts zwecks computergestützter Verarbeitung, sei es mit logischen, begrifflichen oder anderen, insb sprachbezogenen Formalisierungen. Seit einigen Jahren sind Ontologien der Schlüssel zur expliziten Beschreibung von Konzepten in der Domäne Recht [Schweighofer/Liebwald 2004; Gruber 1993, 1999]. Neben der Wissensrepräsentation und der Metabeschreibung des Rechtssystems ist die Automatisierung von (einfachen) juristischen Entscheidungen ein wichtiges Ziel dieser Forschungen.

---

[*] Derzeit karenziert, beschäftigt bei der Europäischen Kommission in Brüssel. Die Ansichten sind jene des Autors.

Das Projekt LOIS (Lexical Ontolgies for legal Information Sharing) beschäftigt sich mit der Schaffung eines multilingualen Thesaurus für verbessertes Retrieval und stellt eine der vielen Vorarbeiten in dieser Richtung dar. In diesem Beitrag soll insb auf die vorgesehene Weiterentwicklung zur semantischen Indexierung von Rechtsinformationssystemen und damit zum elektronischen Kommentar eingegangen werden.

2. Projekt LOIS

Ziel des EU-geförderten eContent Projekts LOIS (Lexical Ontologies for legal Information Sharing)[1] ist es, für Juristen und Laien einen leichteren mono- und multilingualen Zugang zu mono- und crosslingualen Rechtsdatenbanken zur Verfügung zu stellen. Den Benutzern soll es ermöglicht werden, Anfragen an ein juristisches Information-Retrieval-System in ihrer eigenen Sprache zu stellen, aber auch Dokumente in anderen als dieser Sprache zu finden. Sind keine Ergebnisse zum Begriff oder verwandten Begriffen verfügbar (eq_synonym- oder eq_near_synonym-Beziehungen), sollen die hierarchischen Strukturen die Reformulierung der Suche durch den Benutzer unterstützen.

Die juristischen WordNets werden in sechs verschiedenen Sprachen (Italienisch, Niederländisch, Portugiesisch, Deutsch, Tschechisch und Englisch) mittels EuroWordNet[2]-Technologie [Vossen et al 1997] aufgebaut und vernetzt, wobei die LOIS-WordNets mit EuroWordNet kompatibel bleiben und über *plug-in*-Relationen als Erweiterung für die Subdomäne Recht dienen können. Während der zweijährigen Projektdauer sollen etwa 5.000 Synsets für jede der sechs Sprachen „lokalisiert", als Rechtsbegriff formal dargestellt und über einen *inter-lingual-index* (ILI) vernetzt werden.

Ähnliche Begriffe in verschiedenen Sprachen sind zu Synsets verbunden, um die lexikalischen Einträge der jeweiligen Sprachen abzubilden. Jedes Synset enthält eine Glosse (kurze Definition). Taxonomische und lexikalische Beziehungen werden benutzt, um die Synsets zu vernetzen und semantische Bedeutungen darzustellen. Rechtssystemspezifische einmalige Begriffe können in Literalen als Teil eines Synsets dargestellt und durch Hinzufügung entsprechender Metain-

---

[1] Siehe auch http://www.loisproject.org.

[2] Siehe auch http://www.illc.uva.nl/EuroWordNet/. EuroWordNet beruht seinerseits wiederum auf dem Princeton WordNet, http://wordnet.princeton.edu/.

formationen explizit gemacht werden. Dadurch wird eine Verknüpfung von europäischen und typisch nationalen Begriffen erlaubt.

Die LOIS-Datenbank besteht aus fünf Modulen:

- Lexikalische Datenbank
- Rechtsbegriffsdatenbank
- Rechtsdokumentenindex (eindeutige Bezeichnungen)
- Rechtsklassifikationsindex (Fundstellennachweis des geltenden Gemeinschaftsrechts)
- Rechtsontologie auf oberster Ebene

Die Grundlage der lexikalischen Datenbank ist das bereits existierende italienische Juridical WordNet JWN [Gamgemi et al 2003, Dini et al 2005], eine Erweiterung der italienischen EWN-Initiative. Das italienische JWN wird als ILI benutzt, um die sprachspezifischen LOIS-WordNets abzubilden. Die Strukturen der sprachspezifischen WordNets wurden identisch zum italienischen JWN aufgebaut; manuelle Überarbeitung, Anpassung und Integration sind derzeit noch in Arbeit. Die lexikalische Datenbank basiert auf einer geringeren Spezifikationsstufe der Rechtssprache, um einen sehr hohen Grad der Konkordanz zwischen den Begriffen verschiedener Sprachen zu sichern.

Die Rechtsbegriffsdatenbank ist aus Legaldefinitionen aufgebaut, die mittels eines speziell entwickelten (semi-)automatischen Definitionsextraktionstools aus EU-Richtlinien gewonnen wurden. Für die Subgruppe Verbraucherschutzrecht werden auch nationale Umsetzungen und andere dazugehörige nationale Bestimmungen berücksichtigt. Die resultierenden Definitionen der verschiedenen Sprachversionen werden automatisch verbunden; nur nationale Umsetzungsmaßnahmen müssen von Hand hinzugefügt werden. Für die Rechtsbegriffsdatenbank werden die englischen gesetzlichen Definitionen als ILI benutzt. Die Datenbankarchitektur unterscheidet nicht zwischen der lexikalischen Datenbank und der Rechtsbegriffsdatenbank, hinzugefügte Metadaten markieren jedoch die Ableitung der verschiedenen Begriffe und Begriffstypen.

Hinsichtlich der semantischen Beziehungen zwischen Synsets innerhalb einer Sprache sind derzeit vornehmlich Synonymie/Antonymie in Gebrauch, bezüglich der taxonomischen Beziehungen Hyperonymie/Hypernymie. Alle Synsets jeder Sprache sind über Äquivalenzrelationen mit dem ILI verlinkt, wobei für jedes Synset in einem einsprachigen WordNet mindestens eine Äquivalenzbeziehung zum ILI besteht – entweder direkt oder indirekt durch verwandte Synsets. Für die Rechtsbegriffsdatenbank wurden die besonderen Beziehungen umgesetzt_in/umgesetzt_als geschaffen.

Weiters wurde eine Ontologie auf oberster Ebene mit ungefähr 50 Begriffen vorgeschlagen, um ein gemeinsames semantisches Netz zu schaffen. Der ILI stellt lediglich eine unstrukturierte Liste von Bedeutungen dar. Daher werden durch die Ontologie auf oberster Ebene ergänzende Informationen bereitgestellt, indem Metainformationen zur Datenbank hinzugefügt werden.

Die Ergebnisse des LOIS-Projekts sind für die Entwicklung der lexikalischen Ontologie von hoher Bedeutung. Bei Vorliegen guter Begriffsdefinitionen und Relationen zwischen den Begriffen sollte es einfach sein, das Lexikon zu einer Ontologie umzuformen.

3. Thesaurus und multilinguales Information Retrieval

Ein Thesaurus ist als Dokumentationssprache eine geordnete Zusammenstellung von Begriffen und ihren (vorwiegend natürlichsprachlichen) Bezeichnungen zum Indexieren, Speichern und Wiederauffinden von Inhalten [Schweighofer 1999, 63 f]. Der Dokumentinhalt soll präzise komprimiert und exakt dargestellt werden. Thesauri zielen daher – wie Ontologien – auf die Beschreibung der Welt, wobei der Unterschied in der Tiefe der Beschreibung und im Element der Automatisierung liegt. Multilinguale Thesauri – wie zB EUROVOC – unterstützen wesentlich die Wanderung zwischen verschiedensprachigen Begriffssystemen.

Im Projekt LOIS umfasst ein Thesauruseintrag die Bezeichnung (Header), eine Definition (Glosse) sowie Beziehungen. Dieser Thesauruseintrag wird mit sechs europäischen Sprachen gleichgesetzt.

Für das multilinguale Information Retrieval steht neben der Suche mit allen Bezeichnungen desselben Begriffs auch die Einbeziehung der Glossen, dh des Kontextes, zur Verfügung. Das Problem liegt einerseits in der Repräsentation des Kontextes, andererseits in der Berechnung der Übereinstimmung. Tests mit dem multilingualen Tool[3] sind für Mitte 2005 geplant, und zwar vorerst nur mit den Bezeichnungen. Aufgrund der Testresultate wird zu entscheiden sein, ob und wie diese Erweiterung stattfinden soll.

---

[3] Der Prototyp steht unter http://195.110.142.254:8091/docd/clir/ zur Verfügung.

4. Konzeption einer umfassenden Ontologie

In der Informatik und in der Wissensrepräsentation wird unter einer Ontologie eine explizite formale Spezifikation einer gemeinsamen Konzeptualisierung verstanden [Gruber 1992, 1993; Schweighofer/Liebwald 2004]. Kern jeder juristischen Ontologie [Breuker et al 2002, Breuker/Winkels 2003] ist die Beschreibung der realen Welt (des Weltwissens) sowie des Rechtssystems. Die Vernetzung der Begriffswelten des Weltwissens und des juristischen Wissens ist eine Zentralaufgabe der Ontologie.

Die umfassende Ontologie [Schweighofer/Liebwald 2004] ist ein neuer Lösungsweg zum dynamischen elektronischen Kommentar. Damit soll die Beschränkung des Rechtsinformationssystems auf umfassende und präzise Dokumentation des Rechts überwunden und durch Hinzufügung einer weiteren Abstraktionsebene ein wesentlicher qualitativer Mehrwert geschaffen werden. Neben dem dynamischen elektronischen Kommentar sollen auch automatisierte Rechtsanwendungen durch die umfassende Ontologie ermöglicht werden.

„Umfassend“ bedeutet, dass eine vollständige Beschreibung des Rechtssystems (das wäre das Rechtsinformationssystem) sowie des Weltwissens (das wäre zB das Semantic Web [Koivunen/Miller 2001, Berners-Lee et al 2001] oder die englischsprachige lexikalische Datenbank WordNet [Miller et al 1990, Fellbaum 1998]) Ausgangspunkt der Ontologie ist. Dieses Wissens soll in Frames umgeformt und umfassend verlinkt werden.

Die Ontologie soll zwei Frames umfassen: Rechtframes und Sachframes. Jedes Frame hat insb folgende Attribute bzw Relationen: Name, Beschreibung, hierarchische Einordnung (Ober- bzw Unterordnung, Klassifikation), begriffliche Einordnung (wie Homonym, Synonym und Polysem) und juristische Einordnung (wie Tatbestand bei Sachbegriffen oder Sachframes und Normen bei juristischen Begriffen). Diese Verlinkung ist neben der Beschreibung der wichtigste Teil der umfassenden Ontologie. Das Rechtframe hat drei Varianten: materielle Norm, formelle Norm und Rechtsbegriff. Hinsichtlich des Sachgebietsframes wird unterschieden zwischen Agenten-, Objekt- und Prozessframes.

Als Zwischenschritte zur umfassenden Ontologie werden das hybride wissensbasierte System sowie der ontologische Thesaurus angesehen [Schweighofer/Liebwald 2004]. Das hybride wissensbasierte System [Schweighofer 1999] ist eine wissensbasierte Umformung des Informationssystems einer Rechtsordnung durch (semi-)automatische Methoden, wie Darstellung der materiellen Regeln als logische Sätze,

Einsatz von korpusbezogenen Begriffsdeutungen oder Zusammenfassungen und automatisch generierten Verweisungen. Der ontologische Thesaurus unterscheidet sich vom üblichen Thesaurus durch die umfassendere Beschreibung, die Einbeziehung des Weltwissens sowie auch die Verlinkung mit diesem.

5. Vom Thesaurus zur Ontologie

Der Thesaurus liefert das Grundgerüst für die Ontologie: ein System der Rechtsbegriffe als Beschreibung des Rechtssystems. Die bereits bestehenden Begriffshierarchien und Relationen müssen weiter ausgebaut und ergänzend mit Attributen versehen werden. Sachbegriffe sind im Rechtsthesaurus nicht enthalten; ein erster Schritt wird jedoch durch die jeweiligen Begriffsbeschreibungen gesetzt. Sachbegriffe sind als Begriffe aufzunehmen, zu beschreiben und sowohl untereinander als auch mit den Rechtsbegriffen zu vernetzen.

6. Projekt: Vernetzung Thesaurus – wissensbasiertes System

Um die Zweckmäßigkeit des Ansatzes zu prüfen, soll als Zwischenschritt in einem kleineren Rechtsgebiet (zB Staatsbeihilfenrecht oder Datenschutz) ein erster Prototyp der umfassenden Ontologie geschaffen werden. Als Projektstart ist Herbst 2005 vorgesehen.

Ausgangspunkt sind folgende Grundlagen: Textkorpus, Rechtsthesaurus und Sachthesaurus. Der Textkorpus aller Normen, Gerichtsentscheidungen sowie sonstiger relevanter Texte wird mit Hilfe der Rechtsinformationssysteme aufgebaut. Der Rechtsthesaurus stammt aus den Begleitarbeiten zum Projekt LOIS. Der Sachthesaurus basiert auf den Arbeiten des WordNet.

Im ersten Schritt wird dieses Gemenge von Wissenselementen (semi-)automatisch in die Framestrukturen der umfassenden Ontologie umgeformt. Als wesentlichste Vorarbeit ist es im zweiten Schritt notwendig, dass der Textkorpus in ein hybrides wissensbasiertes System umgeformt wird. Die Sätze werden als quasi-logische Regeln dargestellt (zB AustLII[4] oder SoftLaw[5]). Die Verweisungen sind automatisch zu extrahieren (zB AustLII). Die Dokumente bzw Normen werden automatisch klassifiziert sowie auch die Cluster inhaltlich beschrieben

---

[4] Siehe http://www.austlii.edu.au/ oder die AustLII Papers im Journal of Information, Law & Technology (JILT) unter http://elj.warwick.ac.uk/jilt/.

[5] Siehe http://www.softlaw.com.au/.

(zB GHSOM, LabelSOM [Schweighofer et al 2001]). Weiters sind korpusbasierte Begriffsanalysen zur Verbesserung des Thesaurus vorgesehen (zB KONTERM [Schweighofer 1999]).

Der dritte Schritt der Verlinkung zwischen den jeweiligen Frames stellt den schwierigsten Teil des Aufbaus der Ontologie dar. Hierzu sind hermeneutische Regeln vorgesehen, wie diese auch bei automatischen Verlinkungen oder Textanalysen verwendet werden. Sprachmuster mit semantischer Bedeutung werden zueinander in eine juristisch relevante Beziehung gebracht, wie: Tatbestandsbegriffe und Rechtsnorm bzw Rechtsbegriff oder Begriffsausprägungen durch Verweis auf Gerichtsentscheidungen. Durch diese strukturierte Vernetzung wird ein „Rohkommentar" geschaffen, der eine wesentliche und unentbehrliche Hilfe einer Kommentierung oder einer (einfachen) automatischen Applikation sein soll.

Für all diese Schritte sind bereits Prototypen vorhanden, sodass die Verwirklichung möglich scheint. Schwierig ist es jedoch, die jeweilige Performance und Qualität sowie die Integration der verschiedenen Programmtools zu erzielen. Der Aufbau einer Wissensbasis hermeneutischer Regeln wird vieler Tests zur Qualitätsverbesserung bedürfen. Bei Erfolg wird dieses Zwischenprojekt wichtige Grundlagen für die Konzeption und Implementierung der umfassenden Ontologie schaffen.

7. Konklusionen

Durch den Aufbau eines multilingualen Thesaurus schafft das LOIS Projekt die Basis für effizientes und exaktes multilinguales Retrieval. Dem Thesaurus kommt aber auch eine wichtige vorbereitende Rolle für den Aufbau einer Rechtsontologie zu. Mit dieser Wissensbasis wird der Kern für einen dynamischen elektronischen Kommentar sowie für automatische Rechtsanwendungen gelegt. Mit dem Projektentwurf einer Vernetzung des Thesaurus mit dem hybriden wissensbasierten System wird ein Weg zum Aufbau einer umfassenden Ontologie gezeigt.

Literatur

[Berners-Lee et al 2001] *Berners-Lee, T. et al*, The Semantic Web (2001), Scientific American 05/2001, NY, http://www.scientificamerican.com/

[Breuker et al 2002] *Breuker, J. et al*, Ontologies for legal information serving and knowledge management, in: Proceedings of the 15th Jurix (London, UK, 2002), IOS Press, Amsterdam et al, 2002, 73-82

[Breuker/Winkels 2003] *Breuker, J./Winkels, R.*, Use and reuse of Legal ontologies in knowledge engineering and information management, Workshop on Legal Ontologies (in conjunction with the 9th ICAIL, Edinburgh, UK, 2003), http://www.lri.jur.uva.nl/~winkels/legontICAIL2003.html

[Dini et al 2005] *Dini, L. et al*, Cross-lingual legal information retrieval using a WordNet architecture, LOAIT (in conjunction with the 10th ICAIL, Bologna, IT, 2005), in print

[Fellbaum 1998] *Fellbaum, C.* (Hrsg), WordNet: An Electronic Lexical Database (1998), MIT Press, Cambridge, MA

[Gangemi 2003] *Gangemi, A./Sagri, M.-T./Tiscornia, D.*, Jur-Wordnet, a Source of Metadata for Content Description in Legal Information, Workshop on Legal Ontologies (in conjunction with the 9th ICAIL, Edinburgh, UK, 2003), http://www.lri.jur.uva.nl/~winkels/legontICAIL2003.html

[Gruber 1992] *Gruber, T.R.*, ONTOLINGUA: A Mechanism to Support Portable Ontologies, Knowledge System Laboratory (1992), Stanford University, CA

[Gruber 1993] *Gruber, T.R.*, A Translation Approach to Portable Ontology Specifications. Knowledge Acquisition Vol 5/2 (1993), Academic Press, London et al, UK, 1993, 199-220 (199).

[Koivunen/Miller 2001] *Koivunen, M.-R./Miller, E.*, W3C Semantic Web Activity, Semantic Web Kick-off Seminar/Finland 2001, http://www.w3.org/2001/12/semweb-fin/w3csw

[Miller et al 1990] *Miller, G.A. et al*, Five Papers on WordNet, CSL Report 43 (1990), Cognitive Science Laboratory, Princeton University (ftp://ftp.cogsci.princeton.edu/pub/wordnet/5papers.ps)

[Schweighofer 1999] *Schweighofer, E.*, Rechtsinformatik und Wissensrepäsentation (1999), Forschungen aus Staat und Recht 124, Springer, Wien/NY

[Schweighofer et al 2001] *Schweighofer, E., Rauber, A., Dittenbach, M.*, Automatic text representation, classification and labeling in European law, in: Proceedings of the 8th ICAIL (St. Louis, Missouri, US, 2001), ACM Press, New York, NY, 2001, 78-87

[Schweighofer/Liebwald 2004] *Schweighofer, E./Liebwald, D.*, Konzeption einer Ontologie der österreichischen Rechtsordnung, in: Schweighofer, E./Liebwald, D./Kreuzbauer, G./Menzel, T (Hrsg), Informationstechnik in der juristischen Realität, Aktuelle Fragen der Rechtsinformatik 2004, Verlag Österreich, Wien 2004, 39-48

[Vossen et al 1997] *Vossen, P./Díez-Orzas, P./Peters, W.*, The Multilingual design of the EuroWordNet Database, in: Vossen et al (Hrsg), Proceedings of the ACL/EACL-97 workshop Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications (Madrid, July 12th, 1997), Madrid, ES, 1997, 1-8