Undefined 0 (0) 1
IOS Press

Linked Legal Data: A SKOS Vocabulary for the Code of Federal Regulations

Núria Casellas *
Legal Information Institute, Cornell Law School, Cornell University,
Myron Taylor Hall, Ithaca 14850, NY, United States
E-mail: nuria.casellas@cornell.edu
Institut de Dret i Technologia, Universitat Autònoma de Barcelona
School of Law (Ed. B), Bellaterra (Cerdanyola del Vallès), 08193, Spain.
E-mail: nuria.casellas@uab.cat

Abstract. This paper describes the application of Semantic Web and Linked Data techniques and principles to regulatory information for the development of a SKOS vocabulary for the Code of Federal Regulations (in particular of Title 21, Food and Drugs). The Code of Federal Regulations is the codification of the general and permanent enacted rules generated by executive departments and agencies of the Federal Government of the United States, a regulatory corpus of large size, varied subject-matter and structural complexity. The CFR SKOS vocabulary is developed using a bottom-up approach for the extraction of terminology from text based on a combination of syntactic analysis and lexico-syntactic pattern matching.

Although the preliminary results are promising, several issues (a method for hierarchy cycle control, expert evaluation and control support, named entity reduction, and adjective and prepositional modifier trimming) require improvement and revision before it can be implemented for search and retrieval enhacement of regulatory materials published by the Legal Information Institute. The vocabulary is part of a larger Linked Legal Data project, that aims at using Semantic Web technologies for the representation and management of legal data.

Keywords: SKOS, natural language processing, information extraction, legal ontologies, term extraction, ontology learning

1. Introduction

The regulatory system represents the largest contact surface between governmental activity and the governed. It is large, complex, and enormously varied in its subject matter, and, more importantly, it currently generates large amounts of siloed data: from rulemaking materials, to implementation or guidance materials, and finding aids. Moreover, the content of regulations cannot be regarded in isolation, as legislative activities, judicial decision-making, and the daily work of the issuing agencies (e.g. datasets generated from audits or compliance evaluations performed according to their delegated authority) shape its evolution and substance.

In the United States, the Code of Federal Regulations is the codification of the general and permanent enacted rules generated by executive departments and agencies of the Federal Government "[t]he air we breathe, the water we drink, the jobs we hold, and the general welfare of our families and friends are increasingly protected and defined by rules issued by federal agencies of various sorts" [34].

*Visiting Post-doctoral researcher at Cornell University, this research is supported by a MEC/Fulbright 2010-2012 grant from the Spanish Ministry of Education. Assistant Professor at the Universitat Autònoma de Barcelona, School of Law, Spain. The author would like to acknowledge the work of Cornell's MEng students Dallas Dias, Xu Luo, Sharvari Marathe, and Ankit Singh, the feedback and assistance of Thomas R. Bruce Sara Frug, Daniel Nagy, David Shetland, and Wayne Weibel at the Legal Information Institute, and of Prof. Claire Cardie at Cornell's Department of Computer Science. Related partial work has been presented at AAAI Fall Symposium Series 2011 and at the dg.o 2012 conference.

The complexity of the regulatory system, together with the variety of the subject matter and the size of the corpus of regulatory text poses difficulties to the search, retrieval and understanding of legal information. For example, a consumer is concerned with the usage of a product and the service provided by the manufacturer. On this scenario, the retrieval of all the relevant and applicable safety and consumer related information, together with their related procedures is no trivial task due to the particularities of the legal terminology, the structure of the Code, the organization of the materials, etc.

The reuse, conversion of existing content-related thesauri, controlled vocabularies, or taxonomies in a machine-readable form or the development of a SKOS vocabulary for the Code of Federal Regulations could allow semantic search and retrieval enhancement of regulatory materials.

Moreover, the formalization of these regulatory materials, compiled in the Code of Federal Regulations, in Semantic Web machine-readable formats, together with its interlinking using Linked Data principles to other relevant datasets could facilitate the development of regulatory compliance applications in a variety of domains: pharmaceutical product development, data protection analysis, risk assessment, safety compliance, patent assessment in biotechnology ([40]), product management, record-keeping compliance, etc.

In particular, as envisioned in the Linked Legal Data project [11], a SKOS vocabulary of the Code of Federal Regulations terminology could be extended to incorporate knowledge regarding defined terms, regulated objects, obligations, etc. and support the integration of machine-readable regulatory knowledge with other relevant vocabularies or datasets. For example, Linked Data approaches applied on such a vocabulary could allow cross-jurisdictional search based on thesaurus matching and other term-based extensions (e.g. EuroVoc Thesaurus) or support the aggregation of pharmaceutical regualatory materials (e.g. DrugBank database).

After an overview of Semantic Web and Linked Data approaches in the legal domain and of the Code of Federal Regulations in section 2, this paper describes the development process of a SKOS vocabulary for the Code of Federal Regulations, taking into account the possibility to reuse existing materials, the conversion of regulatory-related thesauri, and, finally, or the application of techniques to extract the terminology from the CFR text. Section 4 contains an example of the application of the SKOS vocabulary for Linked Data purposes in the pharmaceutical domain with the DrugBank dataset. Finally, some conclusions and futher work are outlined in Section 5.

2. Linked Open Legal Data

The World Wide Web Consortium (W3C) currently describes the Semantic Web as "W3C's vision of the Web of linked data",[1] a Web that exposes the meaning of data in a standard machine-readable form that allows users and applications to access, understand, and connect data, and to discover new information and knowledge through aggregation and inference. On one hand, languages such as RDF,[2] RDFS,[3] and OWL 1 and 2,[4] the SPARQL query language.[5] constitute the backbone of the Semantic Web. In the legal domain, for example, `legislation.gov.uk` and `govtrack.us` offer access to legislative RDF data, UK legislation and US bills, members of Congress, voting records, etc., respectively.

On the other, the application of Linked Data principles, such as the URI naming of resources, assertions about named relationships between resources or between resources and data values, and the possibility to easily extend, update, and modify these relationships and resources, allows integration and aggregation.[6]

These language standards, principles and techniques facilitate both the availability of interrelated data sets on the Web in standard formats, and the development of vocabularies, taxonomies, and ontologies to represent and organize conceptual domain knowledge.[7]

2.1. Legal Vocabularies, Taxonomies and Ontologies

In the legal domain the analysis and use of controlled vocabularies, taxonomies, and ontologies to

[1] W3C Semantic Web documentation: `http://www.w3.org/standards/semanticweb/`.

[2] Resource Description Languge Primer (RDF): `http://www.w3.org/TR/2004/REC-rdf-primer-20040210`.

[3] Resource Description Language Schema Primer (RDFS): `http://www.w3.org/TR/rdf-schema`

[4] Ontology Web Language Primer (OWL): `http://www.w3.org/TR/2009/REC-owl2-primer-20091027`.

[5] SPARQL Query Language: `http://www.w3.org/TR/rdf-sparql-query` and SPARQL Query Language 1.1: `http://www.w3.org/TR/sparql11-query`.

[6] `http://www.w3.org/DesignIssues/LinkedData.html`

[7] For a discussion on the meaning and the evolution of concept of computational *ontology*, see [10].

support legal information search and retrieval is extensive, as the use of Semantic Web ontology languages, which offer machine-readable semantic metadata, enhances storage, search and retrieval of information and knowledge.[8]

Initially, legal ontologies were built mainly at core level, an intermediary level that relates upper to domain ontologies in the legal domain (e.g. the Frame-based Ontology of Law (FBO) [66], the Functional Ontology of Law (FOLaw) [63], the LRI-Core ontology [7], and, most recently, the LKIF Core Ontology [30,5,31]).[9]

However, most legal ontologies constructed up to date are built towards, semantic indexing, search and retrieval, and represent mainly domain-specific knowledge. For example, the CLIME ontology was aimed at improving access to international rules and regulations regarding ship classification [70]. The OCL.NL Ontology of Dutch Criminal Law supported semi-automated information management of transcriptions of criminal trial hearings [6]. Jur-(Ital)Wordnet (Jur-IWN) [61,53] and Core Legal Ontology (CLO) are a terminology and ontology-based Italian extension to the legal domain of EuroWordNet [25], respectively. The European VAT Regulatory Ontology represented the financial forensics domain for several languages [33,72]. The Ontology of French Code Law by [39] was developed to search and retrieval of legal information. The Ontology of Dutch Tort Law or BEST-user Ontology supported laymen access to BATNA (Best Alternative to a Negotiated Agreement) information [69,62]. The Legal Taxonomy Syllabus 2.0 by [3,2] takes a comparative law perspective to the modelling of legal terms and concepts from European Union Directives. The Legal Case Ontology "could be used as a tool to build a database of cases" for case representation and reasoning [71].[10] The Ontology of Professional Judicial Knowledge (OPJK) was developed to enhance the search and retrieval capabililites of a web-based frequent question answering system for Spanish judges in their first appointment [12,64,10].

Some areas of legal knowledge have been heavily targeted such as the representation of intellectual property rights for the development of intelligent digital rights management systems (IPROnto, Intellectual Property Rights Ontology, the Copyright Ontology [21,26], the Generic Ontology for Digital Content Licensing [49], and the ALIS IP ontology [15]), the design of data protection management and compliance applications (OntoPrivacy by [9] or the NEURONA Ontologies by [13,46]), and consumer-related legal knowledge to facilitate information gathering and decision-making support due to non-compliance(the Customer Complaint Ontology or CContology, [32], and the Consumer Protection Ontology developed within the DALOS project by [1].)

Multi-lingual and multi-jurisdictional RDF Dictionary for the legal world by [48], the Dictionary Workgroup at LegalXML by [44] and the European Legal RDF Dictionary are also relevant examples of the formalization of legal terminologies and concepts towards cross-search and retrieval of legal information.[11]

Further, see some commercial legal databases provide semantically-enhanced search in their search engines, although few information is available with regards to the technical details of the knowledge bases: LawMoose,[12] LexisNexis TotalPatent^{TM} for patent research,[13] or La Ley Digital.[14]

Currently, the Simple Knowledge Organization System (SKOS), an RDFS/OWL-based specification, supports the representation of controlled vocabularies, thesauri, taxonomies and folksonomies used in knowledge organization systems.[15] The conversion of existing thesauri into the SKOS specification is an increasingly used technique for the publication of thesauri towards reuse, and Linked Data enabling. In this line, the recent conversion of the EuroVoc Thesaurus opens up a new approach to the cross-jurisdictional retrieval of legal information (as already experimental N-Lex portal)[16] that could be further extended if combined with Open Government Data approaches.[17]

[8] For a complete description of existing legal ontologies see [10, 59].

[9] The LKIF Core Ontology, inspired in the Ontology of Fundamental Legal Concepts by [57] and LRI-Core, is available from: `http://www.estrellaproject.org/lkif-core`.

[10] This ontology is available at: `http://wyner.info/research/ontologies/LegalCaseOntology\_v9.owl`.

[11] Visit: `http://rdfdictionary.sourceforge.net`, `http://www.legalxml.org`, and `http://www.lexml.de/eu/index.htm`.

[12] MooseBoost: `http://www.lawmoose.com`.

[13] TotalPatent^{TM} `http://www.lexisnexis.com/semantic-search-1`.

[14] Wolters Kluwer Spain (La Ley): `http://www.atencionclientes.com/FAQ/LALEY/FAQ_Buscar_Sinonimos.htm`.

[15] SKOS: `http://www.w3.org/2004/02/skos`.

[16] N-Lex portal: `http://eur-lex.europa.eu`.

[17] The EuroVoc Thesaurus: `http://eurovoc.europa.eu`. See also [54].

2.2. The Code of Federal Regulations

U.S. Federal regulations are compiled annually in the Code of Federal Regulations (CFR).[18] These regulations are compiled in titles according to their subject matter; currently, the CFR is divided in 50 titles that represent regulatory areas, such as agriculture, finance and tax, food and drugs, judicial administration, energy, etc.[19]

This codification represents a final step for rules produced in the *rulemaking* process, the process by which Federal government agencies and departments formulate, amend, or repeal rules [45] according to their delegated authority and area of activity. Therefore, regulatory-related information and materials are not only available within this compilation, but they are also made available by different sources in different formats at different stages of this rulemaking process.

Moreover, regulatory information is neither produced in the vacuum nor isolated; it is necessarily related to the ongoing work of the issuing agencies (e.g. guidance documents, audit datasets, etc.), the boundaries set up by legislation (published in the Public Law and in the United States Code (USCode), the modifications prescribed by judicial decisions revising regulated issues, and shaped by other relevant documentation (e.g. news, scientific publications, etc.).

Aside from these sources, there are many other regulation-related publications and finding aids. For example, the Unified Agenda of Federal Regulations (Regulatory Agenda) lists the regulations expected to be reviewed or developed in the next year;[20] the List of CFR Sections Affected includes proposed, new, and amended Federal regulations that have been published in the Federal Register since the most recent revision date of a CFR title;[21] the U.S. Government Manual is the official handbook of the Federal Government and it provides comprehensive information on agencies of the legislative, judicial, and executive branches, including quasi-official agencies, international organizations in which the United States participates, and boards, commissions, and committees;[22] the Thesaurus of Indexing Terms "includes indexing terms that describe the specific program regulations of individual agencies as well as general administrative regulations common to all agencies", and it is used by Federal agencies to prepare the List of Subjects included in rule and proposed rule;[23] or the Parallel Table of Authorities (PTOA) lists the rulemaking authority for the regulations codified in the CFR, within others.[24]

Thus, on one hand, linked machine readable data regarding the structure of these materials and their structural relations, together with further relations derived from the above-mentioned finding aids and tables, and the improvement of concept and term-based finding aids that aggregate information regarding regulated objects (including special definitions, obligations, etc.) could offer better support search and retrieval and information aggregation of the regulatory corpus. Towards this end, some preliminary work has been proposed towards the reuse in RDF of the Parallel Table of Authorities [56].

3. The Code of Federal Regulations SKOS Vocabulary

The reuse, conversion or development of existing content-related thesauri, controlled vocabularies, or taxonomies in a machine-readable form could allow semantic search and retrieval enhancement of the Code of Federal Regulations; "to *intelligently browse and retrieve* relevant regulations utilizing familiar terms and vocabularies" [17]. Moreover it would allow the combination of ontology supported search, free text search, or facet search, together with the exploitation of the CFR structural information currently modeled and published in XML. Previous research in this direction includes, for example, the mapping of regulations from several U.S states with existing industry-specific taxonomies (contruction) by keyword matching and structure reuse to cluster relevant sectons from multiple regulations [16,17].

[18] The CFR is updated once per year on a regular basis. For more information see: http://www.archives.gov/federal-register/cfr/about.html.

[19] The Government Printing Office (GPO) publishes an XML version of the CFR as bulk data for download. See 2011 complete CFR at: http://www.gpo.gov/fdsys/bulkdata/CFR/2011.

[20] The Unified Agenda may be consulted at http://www.gpoaccess.gov/ua/index.html.

[21] See the List at: http://www.gpoaccess.gov/lsa/index.html.

[22] The Manual may be consulted at: http://www.gpoaccess.gov/gmanual/about.html.

[23] A reduced 1995 version of the Thesaurus is available in plain text from: http://www.archives.gov/federal-register/cfr/thesaurus-alpha.txt

[24] The PTOA revised as of January 1st, 2011 may be found at: http://www.access.gpo.gov/nara/cfr/parallel/parallel_table.html.

A SKOS vocabulary could incorporate the terms regarding CFR regulated objects and offer a basis for its integration with other extractable information: definitions, obligations, etc. Moreover, linked data approaches applied on such vocabulary would also allow cross-jurisdictional search based on thesaurus matching (e.g. Eurovoc) and other term-based extensions. In this section the development of a SKOS vocabulary for the Code of Federal Regulations is explored, taking into account the possibility to reuse existing materials, the conversion of regulatory-related thesauri, and, finally, or the application of techniques to extract a terminology from text.

3.1. SKOS Vocabulary Reuse

There are many existing vocabularies and taxonomies in reusable machine-readable formats, see table 1, that cover the aspects regulated in the CFR. For example, on one hand, the Agricultural Thesaurus of the U.S. National Agricultural Library includes agricultural terms in English and Spanish;[25] and the AGROVOC thesaurus[26] contains concepts in 21 different languages in the food, nutrition, agriculture, fisheries, forestry, or environmental domains.

On the other, the GLIN Subject Term Index includes the terms used by the Global Legal Information Network database of official texts of laws, regulations, judicial decisions, and other complementary legal sources contributed by governmental agencies and international organizations; the EuroVoc thesaurus is a multilingual thesaurus that includes terms about all the activities of the European Union, and it is used by the Eur-Lex application to enable keyword search for all legal documents produced in the EU;[27] and the Government of Canada Core Subject Thesaurus includes terms from domains included in any information resources of the Government of Canada.

As shown in table 1, there are multiple SKOS vocabularies that can be reused to improve search in multiple domains. However, appropriate coverage and domain representation of the content of the CFR need to be addressed. With regards to coverage, taken individually, few of these vocabularies or taxonomies cover the many domains of interest regulated in the content of the Code of Federal Regulations. As mentioned in section 3, the CFR contains regulatory information regarding all the areas of activity of Federal agencies and departments, from the Animal and Plant Health Inspection Service to the Antitrust Division:

AGROVOC and NAL cover the agricultural domain, although AGROVOC could extend its content to other related areas; DrugBank could be reused for pharmaceutical and drug related terms,[29] Linked Life Data could provide terminologies for the biomedical domain,[30] aerospace-related terms could be reused from the NASA taxonomy,[31] economic terms could be reused from the STW Thesaurus for Economics,[32] etc. Therefore, in order to be able to evenly cover most of the content of the CFR we would require the use of multiple integrated thesauri and taxonomies. Provenance and mapping issues aside, many areas of the Code of Federal Regulations would be still left uncovered (see, for example, table 2).

In order to avoid the coverage problem, vocabularies that are varied in nature, such as LSCH,[33] GLIN, the Government of Canada Core Subject Thesaurus,[34] Dbpedia categories or NY Times subjects,[35] could be reused. However, most of these vocabularies present isses regarding the accuracy of the domain representation; the relation between the vocabulary (the domain it represents and the knowledge acquisition strategy) and the textual source, the CFR, to be enhanced. Neither of these vocabularies contain terms extracted solely from regulatory sources, and some are originated in different legal jurisdictions.

Therefore, the variety of the subject matter and the use of specific terminology in the CFR require tailored solutions: the reuse of specific Federal regulatory vocabularies, if available, or the development of the CFR SKOS vocabulary form the text of the Code of Federal Regulations.

3.2. Existing Thesauri Conversion

The Federal Register Thesaurus of Indexing Terms, as mentioned in section 2.2, is an indexing vocabulary

[25] NAL Thesaurus: `http://agclass.nal.usda.gov`.
[26] AGROVOC: `http://aims.fao.org/agrovoc/lod`.
[27] EuroVoc: `http://eurovoc.europa.eu`.

[29] DrugBank database (XML): `http://www.drugbank.ca`. An RDF version of the DrugBank database is provided by the Free University of Berlin `http://www4.wiwiss.fu-berlin.de/drugbank`.
[30] Linked Life Data: `http://linkedlifedata.com`.
[31] NASA taxonomy: `http://nasataxonomy.jpl.nasa.gov`.
[32] STW Thesaurus for Economics: `http://zbw.eu/stw`.
[33] LSCH: `http://id.loc.gov`.
[34] Government of Canada Core Subject Thesaurus: `http://www.thesaurus.gc.ca`.
[35] NY Times subjects: `http://data.nytimes.com`.

Table 1
List of SKOS vocabularies for possible reuse

SKOS Vocabulary	Domain	Languages	Source
AGROVOC	agriculture, forestry, fisheries, environment	20	Food and Agriculture Organization (FAO) of the United Nations
EuroVoc	Varied (activities EU)	22	EU Office of Publications
Agricultural Thesaurus	agriculture	2	National Agriculture Library
DrugBank	FDA approved drugs	1	DrugBank (RDF version by
Librarly of Congress Subject Headings (LCSH)	Varied (bibliographic)	1	U.S. Library of Congress
Dbpedia Categories	Varied (Wikipedia entries)	$\sim$ 100	Dbpedia (community project)
New York Times Subjects	Varied (news)	1	New York Times
NASA Taxonomy	Varied (NASA web content: locations, missions, etc.)	1	U.S. National Aeronautics and Space Administration
Linked Life Data	biomedical, biotechnology, pharmaceutical data[28]	N/A	Linked Life Data (LarKC EU project)
Government of Canada Core Subject Thesaurus	Varied (information sources from government)	2	Library and Archives Canada
GLIN Subject Term Index	legal (law, regulations, judicial decisions at US and international level	1	Global Legal Information Network
STW Thesaurus for Economics	economy	1	Leibniz Information Center for Economics

Table 2
Example of some uncovered subjects extensively regulated in the CFR

Commerce	Education	Energy	Elections
Employment	Firearms	Food	Housing
Indians	Labor	Nationality	Natural Resources
Patents	Pensions	Products	Property
Security	Telecommunications	Transportation	Wildlife

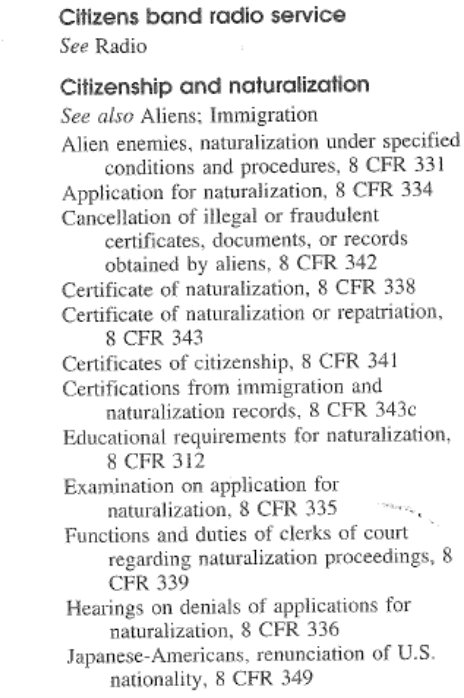
that "includes indexing terms that describe the specific program regulations of individual agencies as well as general administrative regulations common to all agencies. The indexing terms included are intended to express and organize the often technical regulatory concepts in research terms familiar to laypersons." This list of indexing terms is also used by the Office of the Federal Register "as the basis for the subject entries in the Code of Federal Regulations Index which is published annually as of January 1" (see Figure 1). Although little information is available regarding the curation of the Thesaurus and its quality control processes, agencies and staff members of the Office of the Federal Register (National Archives and Records Administration) suggest additions and changes that might be incorporated (55 Fed. Reg 38443, 1990).

Moreover, Federal agencies are required (1 CFR § 18.20) to use the Thesaurus to prepare the "List of Subjects" that is included in the publication of the rules and proposed rules in the Federal Register. This thesaurus is currently made available in printed format and can be requested from Office of the Federal Register, although a plain text 1995 version is available online.[36]

```
§ 18.20
Identification of subjects in agency regulations.
(a) Federal Register documents. Each agency that submits
a document that is published in the Rules and
Regulations section or the Proposed Rules section of
the Federal Register shall--
  (1) Include a list of index terms for each Code of
  Federal Regulations part affected by the document; and
  (2) Place the list of index terms as the last item in
  the Supplementary Information portion of the preamble
  for the document.
(b) Federal Register Thesaurus. To prepare its list of
index terms, each agency shall use terms contained in
the Federal Register Thesaurus of Indexing Terms. Agencies
may include additional terms not contained in the Thesaurus
as long as the appropriate Thesaurus terms are also used.
[...]
```

The conversion of this plain text thesaurus into a machine-readable format could not only allow se-

[36]Thesaurus of Indexing Terms: `http://www.archives.gov/federal-register/cfr/thesaurus.html`.



Citizens band radio service
See Radio

Citizenship and naturalization
See also Aliens; Immigration
Alien enemies, naturalization under specified conditions and procedures, 8 CFR 331
Application for naturalization, 8 CFR 334
Cancellation of illegal or fraudulent certificates, documents, or records obtained by aliens, 8 CFR 342
Certificate of naturalization, 8 CFR 338
Certificate of naturalization or repatriation, 8 CFR 343
Certificates of citizenship, 8 CFR 341
Certifications from immigration and naturalization records, 8 CFR 343c
Educational requirements for naturalization, 8 CFR 312
Examination on application for naturalization, 8 CFR 335
Functions and duties of clerks of court regarding naturalization proceedings, 8 CFR 339
Hearings on denials of applications for naturalization, 8 CFR 336
Japanese-Americans, renunciation of U.S. nationality, 8 CFR 349

Fig. 1. Example of CFR Index entry

mantic search and retrieval enhancement through the annotation and indexing of the content of the text of the Code of Federal Regulations, but also facilitate Linked Data efforts when mapped (e.g. `skos:exactMatch`) to other RDF/SKOS available materials: EuroVoc, NAL Agricultural Thesaurus, Library of Congress Subject Headings, or AGROVOC, etc.

There are a number of methods and tools to explore the conversion of vocabularies and taxonomies into SKOS,[37] According to [65], a three-step approach was followed.

3.2.1. Step A: analyze thesaurus

The plain text version of the Thesaurus of Indexing Terms (or Thesaurus) includes "an alphabetic list of all indexing terms with a series of notations under each term to refer users to preferred or related terms".[38].

[37]OBO to SKOS: `http://www.cs.man.ac.uk/~sjupp/skos`, Zthes to SKOS Converter: `http://www.w3.org/2001/sw/wiki/Zthes_to_SKOS_Converter`, MeshToSKOS: `http://code.google.com/p/hive-mrc/wiki/MeshToSKOS`, a spreadsheet to SKOS method: `http://topquadrantblog.blogspot.com/2010/12/how-to-convert-spreadsheet-to-skos.html`.

[38]Available for download from: `http://www.archives.gov/federal-register/cfr/thesaurus-alpha.txt`.

There are four types of possible relations between the terms: *sa*, *see*, *x*, and *xx*:

```
Agricultural research
     xx
          Agriculture
          Research
AIDS/HIV
    see
          HIV/AIDS
Airmen (13, 19)
      x
          Aircraft pilots
          Pilots
     xx
          Air transportation
Alcohol abuse
     sa
          Alcoholism
```

In this file, also certain numerical codes (from 1 to 19) are assigned to some of the entries; these codes refer to the grouping of these terms in 19 different subject categories contained in a different text file.[39]

From the available documentation, however, there is no established definition of the meaning of the different relation types. Regarding the top categories, although there are 19 broad subject categories appended to some terms in the Thesaurus, most terms do no have such reference. Furthermore, the 19 terms are not included in the Thesaurus.

3.2.2. Step B: map data items to SKOS

Terms were modeled as `skos:Concept` [14], although the documentation contained no clear definitions of the relation types, from the overall analysis the following assumptions were established. First, the use of *see* in "A see B" relationships generally referred to the use of a preferred term and a non-preferred term, mapping to the SKOS `skos:prefLabel` and `skos:altLabel` properties. In this representation, one of the terms is, in the end, included in the vocabulary only as an alternative label (only one `skos:Concept` is created from this structure). Second, the usage of *sa*, as see also, could be mapped to the SKOS `skos:related` object property. Third, the authors assumed that *xx* and *x* stood for the `skos:broader` and `skos:narrower` relationships, respectively.

Finally, the list of broad subject categories were taken into account to establish the top concepts through `skos:hasTopConcept` within the CFRT (Code of Federal Regulations Thesaurus) concept scheme `skos:ConceptScheme`.

[39]Grouping of Terms: `http://www.archives.gov/federal-register/cfr/thesaurus-categories.txt`.

Table 3
Mapping of Thesaurus features into SKOS properties

Data item	Feature/function	Property
A see B	Preferred term and non-preferred term	`skos:prefLabel=B` `skos:altLabel=A`
A sa B	Related term	`skos:Concept` with `rdf:about=A` and `A skos:related B`
A xx B	Broader term	`skos:Concept` with `rdf:about=A` and `A skos:broader B`
A x B	Narrower term	`skos:Concept` with `rdf:about=A` and `A skos:narrower B`
A {grouping}	Broader/Narrower term	`A skos:narrower grouping terms` and `grouping term skos:narrower A`
19 subject categories	Concept scheme	CFRT concept scheme `skos:hasTopConcept` 19 subject categories

3.2.3. Step C: create conversion program

A JAVA program was created that parsed the plain text file, stored the information in arrays and hashtables according to the mappings established in Step B (see table 3), and outputs a SKOS RDF file. A validator was also developed to detect incompleteness and inconsistencies within the output, in order to refine the initial SKOS convertor program.

3.3. Evaluation and Results

Although during the analysis of the documentation some inconsistencies in the usage of the relationships had been noticed, a detailed analysis of the content and the automatic validation of the SKOS conversion results detected a list of incompleteness and inconsistency cases. For example, although most relationships in the Thesaurus also contained their inverses, there were incomplete sets (e.g. Grains x Cereals was present, but Cereals xx Grains was not). Also, some orphan concepts existed outside the top concept relationships.

Moreover, there was a clash between associative links and hierarchical links in some resources, due to an inconsistent use of inverse properties. In particular, it appeared the the pattern "A see B" was generally followed by the inverse "B x A". While the term A was being represented as the content of a`skos:altLabel` for term B (with `skos:prefLabel=B`), a `B skos:narrower A` statement was created at the same time.

Finally, several cyclic loops were detected, mostly due to the introduction of the top concept hierarchy. And, although the existence of `A skos:broader B` together with `B skos:broader A` is consistent with the SKOS data model, "for many applications where knowledge organization systems are used, a cycle in the hierarchical relation represents a potential problem".[40] See table 4 for some the results of the evaluation also performed with SKOS evaluation tools such as qSKOS [41][41] and skosify [60][42] These tools provide evaluation for some features such as: orphan concepts, hierarchical cycles, associative and hierarchical clashes, label conflicts, etc. Skosify also offers the possibility to break hierarchical cycles and solve associative and hierarchical clashes.

For example, many of the abovee-mentioned issues can be already observed with the observation of a related set of entries:

```
Additives
     see
          color additives
          food additives
                    fuel additives
Food additives (01)
     sa
          Color additives
      x
          Additives
          Food ingredients
          Generally Recognized as Safe
          (GRAS) food ingredients
     xx
          Foods
Color additives (01, 09)
      x
          Additives
     xx
```

[40] See SKOS: `http://www.w3.org/TR/skos-reference/`.

[41] For the sake of heterogenity, features are treated similarly for the different methods, however, each method defines the included characteristics under analysis differently. qSKOS: `https://github.com/cmader/qSKOS`.

[42] skosify: `http://code.google.com/p/skosify`.

```
          Food additives
Fuel additives (05)
      x
          Additives
          Gasoline additives
     xx
          Petroleum
Gasoline additives
    see
          Fuel additives
```

Most of these problems could be resolved by creating inverse relations when they are incomplete, creating top concepts out of orphan concepts (or abandoning the use of the grouping of subject categories), favoring hierarchical relationships over associative or label-based relationships, or eliminating a broader/narrower link between nodes in cyclic relationships.

However, the results obtained raised several issues regarding the nature of the Federal Register Thesaurus of Indexing Terms, its curation and quality control processes. On the one hand, the Thesaurus provides an organized list of terms, on the other, it serves as a classification scheme for federal agencies and authorities to establish relations between these terms and *parts* of the Code of Federal Regulations that are affected by their regulations. Moreover, the fact that the terms are used and extended by more than 200 different agencies in their varied subject-matter and particular domains of application may result in some loose curation and control of the organization of the terms.

Finally, the Thesaurus has been mainly developed and organized as a finding aid available in print and, thus, it inherits some of these characteristics in the meaning of its relationships. For example, `A see B`, in print, requires the user of the index to turn pages towards a different entry in the print material. While the establishment of a `skos:altLabel=A` could be reasonable in this scenario, it overlooks the fact that, in this case, the Thesaurus in its plain text version also tries to maintain information regarding the fact that A is somewhat included in or related to B, generating the corresponding inconsistent `B x A` relationship.

These issues not only recommended the curation for digital purposes of the content of the thesurus before attempting a direct conversion into a machine-readable SKOS-based vocabulary, but also supported an approach based on the extraction of a vocabulary from the text of the Code of Federal Regulations.

3.4. CFR Vocabulary Extraction from Text

Terminology extraction and ontology learning from text apply natural language processing, statistical analysis, and machine learning techniques to the automatic discovery and development of vocabularies, taxonomies, and ontologies from textual corpora, supporting the extraction of terms, synonyms (and multilinguistic variants), concepts,[43] taxonomical or non-hierarchical relations, and rules [8].[44] For example, statistical frequencies (e.g. TFIDF, multiterm detection, C-value, etc.), named entity recognition, the use of existing domain vocabularies or ontologies, syntactic parsing (e.g. "chunking"), and pattern-based extraction (e.g. Hearst patterns) are widely used techniques. [50,51,24,35,19]. And several tools that integreate or combine these techniques have been developed for ontology extraction, population and semantic indexing: such as GATE's ANNIE (General Architecture for Text Engineering),[45] TerMine,[46] the KIM Platform,[47] or Text2Onto.[48]

In the legal domain, natural language processing techniques have been applied, for example, towards the extraction of case factors [71], term extraction for ontology enrichment [52,22], ontology learning from Spanish legal texts [68,10], terminology analysis of the French Civil Code [38], support deep semantic analysis interpretation of legal texts [43], support the e-discovery process,[49]syntactic and lexical comparison of Italian legal corpora [67], within others.[50]

3.4.1. Method

Terminology extraction and vocabulary development from the Code of Federal Regulations follows a bottom-up approach based on a combination of syntactic analysis and lexico-syntactic pattern matching of the text contained in CFR parts. The complete text of the Code of Federal Regultations contains over 96.5 million words, therefore, these techniques are applied

[43] While terms may be generally understood as "linguistic realizations of domain-specific concepts", "[t]he extraction of concepts from text is controversial" [8].

[44] Extensive accounts and comparisons between several ontology learning and textual analysis tools may be found in [8,27,28,18,58]. See [67] for a brief overview of techniques on legal corpora.

[45] GATE: `http://gate.ac.uk`.

[46] TerMine: `http://www.nactem.ac.uk/software/termine`.

[47] KIM: `http://www.ontotext.com/kim`.

[48] Text2Onto: `http://code.google.com/p/text2onto`.

[49] TREC (Text Retrieval Conference) Legal Track: `http://trec-legal.umiacs.umd.edu`.

[50] See [23] and the proceedings of SPLeT (Semantic Processing of Legal Texts) 2012 for recent applications on the area: `http://www.lrec-conf.org/proceedings/lrec2012/index.html`.

Table 4
Some evaluation results

Features	Own method	qSKOS	skosify
orphan concepts	N/A	8	12
hierarchy cycles	92	21	62
hierarchy vs associative links	780	709	743
label conflicts	266	86	N/A

at initially Title level, considering the extraction for each CFR Title as a particular concept scheme of a wider collection of vocabularies. In this section, we describe the vocabulary extraction for Title 21 (Food and Drugs), from the XML version of the materials curated by the Legal Information Institute at Cornell University.[51]

Title 21 is divided into chapters, subchapters, parts, subparts and sections, and the latter contain, generally, the text of the compiled rules and regulations. From this structure, first, the text contained in the sections is extracted, and pre-processed. At this step, special characters and numbers are removed, and anaphors are resolved with JavaRAP, an implementation of Resolution Anaphora Procedure (RAP) by [55]. Anaphora resolution determines the antecedent of a reference in the text that points at a previous token, and JavaRAP resolves, in particular, "third person pronouns, lexical anaphors, and identifies pleonastic pronouns".[52]

Then, the Stanford Parser[36] for English language, is a lexicalized probabilistic parser that also provides grammatical relations between the words of a sentence or typed dependencies (Stanford Dependencies, [20]) and is used to tokenize, sentence-split and parse this input, and to output part-of-speech tagged text with a phrase structure grammar representation and its typed dependencies.[53]

```
The/DT purpose/NN of/IN this/DT part/NN is/VBZ
to/TO establish/VB restrictions/NNS on/IN the/DT
sale/NN ,/, distribution/NN ,/, and/CC use/NN
of/IN cigarettes/NNS and/CC smokeless/NNS
tobacco/NN in/IN order/NN to/TO reduce/VB
the/DT number/NN of/IN children/NNS and/CC
adolescents/NNS who/WP use/VBP these/DT
products/NNS ,/, and/CC to/TO reduce/VB
the/DT life-threatening/JJ consequences/NNS
associated/VBN with/IN tobacco/NN use/NN ./.

(ROOT
  (S
    (NP
      (NP (DT The) (NN purpose))
      (PP (IN of)
        (NP (DT this) (NN part))))
    (VP (VBZ is)
      (S
        (VP (TO to)
          (VP (VB establish)
            (NP (NNS restrictions))
            (PP (IN on)
              (NP
                (NP (DT the) (NN sale) (, ,)
                  (NN distribution) (, ,)
                  (CC and)
                  (NN use))
                (PP (IN of)
                  (NP (NNS cigarettes)
                    (CC and)
                    (NNS smokeless) (NN tobacco)))))
            (SBAR (IN in) (NN order)
              (S
                (VP
                  (VP (TO to)
                    (VP (VB reduce)
                      (NP
                        (NP (DT the) (NN number))
                        (PP (IN of)
                          (NP (NNS children)
                            (CC and)
                            (NNS adolescents)))
                        (SBAR
                          (WHNP (WP who))
                          (S
                            (VP (VBP use)
                              (NP (DT these) (NNS products))
                                ))))))
                  (, ,)
                  (CC and)
                  (VP (TO to)
                    (VP (VB reduce)
                      (NP
                        (NP (DT the) (JJ life-threatening)
                          (NNS consequences))
                        (VP (VBN associated)
                          (PP (IN with)
                            (NP (NN tobacco) (NN use)))))))))))))
    (. .)))
```

The following is the list of typed dependencies extracted from the previous sentence parsing:

```
det(purpose-2, The-1)
nsubj(is-6, purpose-2)
prep(purpose-2, of-3)
det(part-5, this-4)
pobj(of-3, part-5)
root(ROOT-0, is-6)
aux(establish-8, to-7)
xcomp(is-6, establish-8)
dobj(establish-8, restrictions-9)
prep(establish-8, on-10)
det(use-17, the-11)
nn(use-17, sale-12)
conj(use-17, distribution-14)
cc(use-17, and-16)
pobj(on-10, use-17)
prep(use-17, of-18)
pobj(of-18, cigarettes-19)
cc(cigarettes-19, and-20)
nn(tobacco-22, smokeless-21)
conj(cigarettes-19, tobacco-22)
mark(reduce-26, in-23)
dep(reduce-26, order-24)
```

[51] LII: http://www.law.cornell.edu.

[52] JavaRAP: http://aye.comp.nus.edu.sg/~qiu/NLPTools/JavaRAP.html.

[53] Stanford Parser: http://nlp.stanford.edu/software/lex-parser.shtml.

```
aux(reduce-26, to-25)
dep(establish-8, reduce-26)
det(number-28, the-27)
dobj(reduce-26, number-28)
prep(number-28, of-29)
pobj(of-29, children-30)
cc(children-30, and-31)
conj(children-30, adolescents-32)
nsubj(use-34, who-33)
rcmod(number-28, use-34)
det(products-36, these-35)
dobj(use-34, products-36)
cc(reduce-26, and-38)
aux(reduce-40, to-39)
conj(reduce-26, reduce-40)
det(consequences-43, the-41)
amod(consequences-43, life-threatening-42)
dobj(reduce-40, consequences-43)
partmod(consequences-43, associated-44)
prep(associated-44, with-45)
nn(use-47, tobacco-46)
pobj(with-45, use-47)
```

From this output, typed dependencies are used to identify certain grammatical patterns that relate to different types of SKOS relationships: noun modifiers, adjectival modifiers, prepositional modifiers, conjunctions, and verbal complementation patterns.[54] For example, **nn** and **amod** are patterns that indentify noun modifiers, noun and adjectival respectively. This dependencies suggest the extraction of `skos:narrower` (with inverse `skos:broader`) properties. The **prep** dependency identifies a prepositional modifier, that used in conjunction with its corrensponent **pobj** property object, can also be used to extend `skos:narrower` properties and create further `skos:related` properties. This pattern, however, is only extracted when a noun is the governer of the dependency. Finally, the **conj** conjunct typed dependency represents a relationship between to elements connected by a coordinating conjunction (e.g. and, or, etc.). This dependency pattern can be use to express the `skos:related` relationship between vocabulary terms. Table 5 depicts a list of examples and converstion types, inverses are not included for brevity.

More complex lexico-syntactic structures are extracted using Hearst patterns, which support the identification of hypernymic and hyponymic relationships between the terms [29].

```
{ such NP as (NP,)* (or|and) NP
{ NP (,)? (such|like) (NP,)* (or|and) NP
{ NP (,)? (including|especially) (NP,)* (or|and) NP
```

These relationships can also be expressed by the use of `skos:narrower` and `skos:broader` properties.

[54] A list of stopwords is used to avoid the extraction of patterns from terms related to the structure of the Code of Federal Regulations or other related legal materials (e.g. part, schedule, section, paragraph, etc.).

```
"A public body, such as a municipality, county,
 district, authority, or other political
 subdivision of a state".

public body skos:narrower municipality
public body skos:narrower county
public body skos:narrower district
public body skos:narrower authority
```

Finally, the extractioin of subject-predicate-object patterns was also experimentally explored, through the analysis of typed dependencies related to the union of nominal subject and direct object using the same governer. For example, for the sentence: "A practitioner may sign a paper prescription in the same manner as he would sign a check or legal document", the following triple would be created: `medical_practioner liivoc:sign paper_prescription`. Once this and all the above-mentioned lexico-syntactic patterns have been extracted, the SKOS RDF statements are generated as output.

3.4.2. Preliminary Results and Evaluation

The bottom-up unsupervised extraction of the vocabulary from Title 21 of the Code of Federal Regulations contains currently 375,000 statements.[55] Upon the analysis of specific concepts, the extraction yields interesting results and captures relevant terminological inforamtion (see, for example, the formalization (appendix A) and a visualization of the concept "milk" in figure 2).

In comparison to the conversion of the Thesaurus of Indexing Terms, in this case, orphan concepts, hierarchical and associative clashes, and label-related issues can be controlled during the extraction. qSKOS evaluation detects few hierarchy cycles, but it is able to assess the existence of associative vs. hierarchical relation clashes taking into account the requirement that `skos:related` is disjoint with the property `skos:broaderTransitive`. Hierarchy cycles continue to appear problematic in this extraction.

Overall, and in comparison to the conversion of the Thesaurus of Indexing Terms (subsection 3.2), the quality of the vocabulary extraction in these ares, see table 6 for results on the complete Title 21 vocabulary,

[55] Although the frequency or relevance of terms is not taken into account to control the set of terminology extracted (e.g. as in Text2Onto [19], in a training set for Title 21 percentage of terms extracted using the part-of-speech tagging (NN/NNPS/NNP/NNS terms) that were incorporated in the vocabulary was of 72.6%. Also, the evaluation of the retrieval and inclusion of frequent multi-word terms (C-Value algorithm [24]) has been evaluated on the same training set; 71.2% of total frequent terms are incorporated in the vocabulary.

Table 5
List of typed dependencies and derived SKOS properties

Typed dependency	Example	SKOS conversion
nn: nound compound modifier	nn(tobacco-22, smokeless-21)	`tobacco skos:narrower smokeless tobacco`
amod: adjectival modifier	amod(consequences-43, life-threatening-42)	`consequences skos:narrower life-threatening consequences`
prep: prepositional modifier & *pobj*: object of preposition	prep(use-17, of-18) & pobj(of-18, cigarettes-19)	`use skos:narrower use of cigarretes` & `cigarretes skos:related use of cigarretes`
conj: conjunct	conj(children-30, adolescents-32)	`children skos:related adolescents`

are greatly improved. However, the universal consideration of adjectival,and prepositional modifiers, the extraction of subject-predicate-object patterns, the parsing of named entities, together with the specificity and complexity of the regulatory text itself (long sentences >70 words, sentences splitted in lists, incorrect use of punctuation, varied use of capitalization, etc.) result in defective and uneven output.

The generic approach taken to the use of typed dependencies of adjectival and prepositional modifiers for vocabulary extraction presents significant drawbacks. On one hand, the adjectival modifier extraction is able to detect relevant vocabulary entries, such as, "transgenic animal", "exotic animal" or "milk-producing animal", while, at the same time, it would also extract "complete animal" and "adequate milk". On the other, the extraction of structures based on prepositional modifiers seems to render mixed results (from "diet for animal", "size of animal", or "edible product from treated animal" to "animal per head" and "number of animal"). A more granular revision of Treebank's pos-tagger [42] with regards to the use of adjectives and prepositions could improve the final results for the extraction of properties based on **amod** and **prep** dependencies.[56] For example, the **JJ** Treebank tag for adjectives includes ordinal numbers, and although most comparative adjectives and superlatives are included within the **JJR** and **JJS** tags, some are also being included within the more generic **JJ** tag.

While a stopword list is already taken into account to control terminology extraction, the improvement of the quality of this list through evaluation of the current results, together with the introduction of a human-in-the-loop for expert validation and vocabulary control could greatly benefit the output of the extraction. This semi-automatic approach could also offer support to a method for hierarchy cycle control and a frequency-based method for vocabulary trimming.

[56]Treebank: `http://www.cis.upenn.edu/~treebank`.

In the same line, although pre-processing takes already into account a list of named entities such as agencies, departments and acts, there is a need to detect multiple-term named entities to improve the results of grammar parsing. Also, although interesting results regarding regulatory procedural knowledge are extracted from the subject-predicate-object patterns in the analysis of the union of nominal subject and direct objects, further evaluation of the implications for the overall structure of the vocabulary is necessary (e.g. larger retrieval of terms). Finally, the specificity of regulations and the granularity of their content affects the structure of sentences contained in the Code of Federal Regulations, as shown in the example below. Further, tailored parser training and pre-processing of regulatory text pose challenging tasks, common to the particularities of legal text, in general [4,37,39,67].

```
(a) Records for manufacturers. Each person registered or
authorized to manufacture controlled substances shall
maintain records with the following information:
  (1) For each controlled substance in bulk form to be
      used in, or capable of use in, or being used in,
      the manufacture of the same or other controlled or
      noncontrolled substances in finished form,
   (i) The name of the substance;
   [...]
        (v) The quantity used to manufacture the same
        substance in finished form, including:
       (A) The date and batch or other identifying
             number of each manufacture;
          (B) The quantity used in the manufacture;
          [...]
          (H) The theoretical and actual yields; and
       (I) Such other information as is necessary to
            account for all controlled substances used
                in the manufacturing process;
          [...]
```

4. Linked CFR Data: DrugBank

As outlined in section 2, the application of Linked Open Data (LOD) principles to legal information (URI naming of resources, assertions about named relationships between resources or between resources and data values, and the possibility to easily extend, update and modify these relationships and resources) could offer better access and understanding of regulatory information to individual citizens, businesses and government

Table 6
Some evaluation results

Features	Own method	qSKOS
orphan concepts	0	0
hierarchy cycles	1114 (x2)	4
hierarchy vs associative links	0	7885
label conflicts	0	0

agencies and administrations, and allow its sharing and reuse across applications, organizations and jurisdictions.

Title 21 of the Code of Federal Regulations, titled "Food and Drugs", contains most of the enacted Federal rules and regulations related to medical devices, chemicals, (manufacturing, labeling, and packaging of) pharmaceutical products, prescriptions, cosmetics, medical records, clinical trials, exportation and importation of controlled substances, and procedures and functions of the Food and Drug Administration (FDA), within others.

The vocabulary developed for Title 21, together with other CFR-based generated datasets, could be related to other relevant datasets and vocabularies; for example the ones analyzed in subsection 3.1. In particular, Title 21's CFR vocabulary could be easily related to the DrugBank dataset (currently developed by the Departments of Computing Science and Biological Sciences of the University of Alberta, Canada). This dataset "contains 6711 drug entries including 1447 FDA-approved small molecule drugs, 131 FDA-approved biotech (protein/peptide) drugs, 85 nutraceuticals and 5080 experimental drugs".[57]

The extension of the CFR SKOS vocabulary with DrugBank Linked Data, which, in turn has been already linked to other resources, such as Dbpedia.org, through exact string matching of the labels and the formalization of `owl:sameAs` statements [47], could offer an initial testbed for the use and integration of regulatory data in applications that require regulatory knowledge for the development and maintenance of drug inventories, or requirement compliance for chemical storage and pharmaceutical product development, etc.

Although the extraction of the vocabulary is not yet perfected, and the retrieval of drug terms poses certain text pre-processing demands (e.g. they are generally listed in tables), for demonstration purposes, `owl:sameAs` mapping relationships can already be discovered through the string matching between some DrubBank drugs (`rdf:type drug`) and CFR terms. Table 7 describes a small amount of the possible mappings.[58]

```
<skos:Concept rdf:about="http://liicornell.org/isopropamide">
  <owl:sameAs rdf:resource="http://www4.wiwiss.fu-berlin.de/
   drugbank/resource/drugs/DB01625"/>
</skos:Concept>
<skos:Concept rdf:about="http://liicornell.org/prednisolone">
  <owl:sameAs rdf:resource="http://www4.wiwiss.fu-berlin.de/
   drugbank/resource/drugs/DB00860"/>
</skos:Concept>
<skos:Concept rdf:about="http://liicornell.org/immune_globulin">
  <owl:sameAs rdf:resource="http://www4.wiwiss.fu-berlin.de/
   drugbank/resource/drugs/DB00028"/>
</skos:Concept>
<skos:Concept rdf:about="http://liicornell.org/mupirocin">
  <owl:sameAs rdf:resource="http://www4.wiwiss.fu-berlin.de/
   drugbank/resource/drugs/DB00410"/>
</skos:Concept>
```

5. Conclusions and Further Work

In this paper, the development of a SKOS vocabulary for the Code of Federal Regulations (Title 21 in particular) has been explored from three different approaches: through the reuse existing materials, the conversion of regulatory-related thesauri, and, finally, or the application of natural language processing techniques to extract a terminology from text.

After the revision of several available vocabularies, the reuse of non-regulatory vocabularies was abandoned due to the variety of the subject matter and the use of specific terminology in the Code of Federal Regulations. Although the conversion of the Federal Register Thesaurus of Indexing Terms was attempted, the results obtained were inadequate and recommended the revision and curation for digital pur-

[57] DrugBank database (XML): `http://www.drugbank.ca`. The availability of linked pharmaceutical regulatory information and data could support the development of applications to monitor the safety requirements of certain chemicals, the changes in the regulatory environment for the development of pharmaceutical products, or facilitate an entry point to regulations for concerned consumers of an FDA approved drug (e.g. the conversion of a brand name into its pharmaceutical components). The RDF dataset of the DrugBank database is currently maintained by the Free University of Berlin `http://www4.wiwiss.fu-berlin.de/drugbank`.

[58] Due to space constraints the namespaces have been modified.

Table 7
Exact matching between DrugBank drugs and CFR vocabulary terms

Drug ID	CFR term ID	Drug ID	CFR term ID
DB01625	isopropamide	DB01134	desoxycorticosterone_pivalate
DB00860	prednisolone	DB01075	diphenhydramine
DB00028	immune_globulin	DB00878	chlorhexidine
DB00410	mupirocin	DB00312	pentobarbital
DB04272	citric_acid	DB00119	pyruvic_acid
DB04183	methylmalonic_acid	DB00518	albendazole
DB00684	tobramycin	DB00971	selenium_sulfide
DB01677	fumarate	DB00446	chloramphenicol
DB05245	silver_sulfadiazine	DB02579	acrylic_acid
DB01093	dimethyl_sulfoxide	DB00479	amikacin
DB01160	dinoprost_tromethamine	DB04626	apramycin
DB03733	ethylene_dichloride	DB00821	carprofen
DB04566	inosinic_acid	DB01396	digitoxin
DB00919	spectinomycin	DB01213	fomepizole
DB00281	lidocaine	DB02640	fumagillin
DB00148	creatine	DB04077	glycerol
DB01174	phenobarbital	DB00602	ivermectin
DB01592	iron	DB01009	ketoprofen
DB00986	glycopyrrolate	DB00814	meloxicam
DB00107	oxytocin	DB00826	natamycin
DB00730	thiabendazole	DB01345	potassium
DB00440	trimethoprim	DB03904	urea
DB00825	menthol	DB00595	oxytetracycline
DB00121	biotin	DB00729	diphemanil_methylsulfate
DB04257	palmitoleic_acid	DB04829	lysergic_acid_diethylamide

poses of the content of the Thesaurus before conversion. The current CFR SKOS vocabulary is developed from a bottom-up approach for the extraction of terminology from texts, through the use of a combination of syntactic analysis and lexico-syntactic pattern matching. The SKOS vocabulary is described per Title of the Code of Federal Regulations and, this paper, describes the extraction of the vocabulary for Title 21, Food and Drugs.

Although the preliminary results are promising, several issues (a method for hierarchy cycle control, expert evaluation/curation/control support, named entity detection, and adjective and prepositional modifier reduction) require improvement and revision before the release of the final version of the vocabulary. The improvement of evaluation techniques and the design of vocabulary quality control measures (e.g. expert evaluation, term reduction, quality stopword lists, etc.) is currently in progress.

The Code of Federal Regulations vocabulary, integrated by title-based concept schemes, will support the enhancement of search, retrieval, navigation, discovery and aggregation of regulatory materials at the Legal Information Institute. Also as an example of the Linked Data possibilities offered by such a vocabulary, an exploratory interlinking with the DrugBank database materials is suggested.

This development is part of a larger Linked Legal Data project [11], that aims at the integration of the Code of Federal Regulations, other related materials and finding aids. The use of Semantic Web technologies for the specification in machine-readable format of regulatory concepts, regulated objects, obligations and legal definitions could provide, for example, a means to investigate agency behavior in rulemaking (e.g. querying for the agencies involved in CFR section modification and regulation through time); could enable search and retrieval of requirements, obligations, etc. with regards to a regulated product (e.g. linking the vocabulary to pharmaceutical data from the DrugBank database); could allow the reuse of structural information (both hierarchy and citations) to enable cross-search between legislation, regulations and case-law; could enable thesauri enhanced search through estab-

lishing relationships between thesauri and CFR content, and offering regulated concepts (and its legal definitions) as Linked Data to be reused by other domain applications.

Thus, similar techniques used in the development of this vocabulary are taken into account for the detection and extraction of defined terms (together with their definitions and its scope), and obligations (e.g. addressee). These materials are formalized also in RDF to facilitate integration with the current vocabulary and enable access to regulatory linked data.

References

[1] Tommaso Agnoloni, Lorenzo Bacci, Enrico Francesconi, Wim Peters, Simonetta Montamegni, and Giulia Venturi. A two-level knowledge approach to support multilingual legislative drafting. In Joost Breuker, Pompeu Casanovas, M. Klein, and Enrico Francesconi, editors, *Law, Ontologies and the Semantic Web. Channelling the Legal Information Flood*, volume 188 of *Frontiers in Artificial Intelligence and Applications*, pages 177–198. IOS PRESS, ios press edition, 2009.

[2] Gianmaria Ajani, Guido Boella, Leonardo Lesmo, Marco Martin, Alessandro Mazzei, Daniele P. Radicioni, and Piercarlo Rossi. Multilevel legal ontologies. In *Semantic Processing of Legal Texts*, pages 136–154, 2010.

[3] Gianmaria Ajani, Guido Boella, Leornardo Lesmo, Marco Martin, Alessandro Mazzei, Daniele P. Radicioni, and Piercarlo Rossi. Legal taxonomy syllabus version 2.0. In *In Proceedings of LOAIT 2009 at ICAIL 2009*, volume 2 of *IDT Series*, pages 9–17. IDT/Huygens Editorial, 2009.

[4] Trevor J. M. Bench-Capon and Pepijn R. S. Visser. Open texture and ontologies in legal information systems. In *DEXA Workshop*, pages 192–197, 1997.

[5] Alexander Boer, Radboud Winkels, , and Fabio Vitali. Metalex xml and the legal knowledge interchange format. In Pompeu Casanovas, Giovanni Sartor, Núria Casellas, and Rossella Rubino, editors, *Computable Models of the Law*, volume 4884 of *Lecture Notes in Artificial Intelligence*, pages 21–41. Springer-Verlag, Berlin Heidelberg, 2008.

[6] Joost Breuker, Abdullatif Elhag, Emil Petkov, and Radboud Winkels. Ontologies for legal information serving and knowledge management. In T.J.M. Bench-Capon, A. Daskalopulu, and R.G.F. Winkels, editors, *Legal Knowledge and Information Systems. Jurix 2002: The Fifteenth Annual Conference*, pages 73–82, Amsterdam, 2002. IOS Press.

[7] Joost Breuker and Rinke Hoekstra. Core concepts of law: Taking common-sense seriously. In *Proceedings of Formal Ontologies in Information Systems FOIS-2004*, pages 210–221. IOS-Press, 2004.

[8] Paul Buitelaar, Philipp Cimiano, and Bernardo Magnini. Ontology learning from text: An overview. In Paul Buitelaar, Philipp Cimiano, and Bernardo Magnini, editors, *Ontology Learning from Text: Methods, Evaluation and Applications*, volume 123 of *Frontiers in Artificial Intelligence and Applications Series*. IOS Press, July 2005.

[9] A. Cappelli, V. Bartalesi Lenzi, R. Sprugnoli, and C. Biagioli. Modelization of domain concepts extracted from the italian privacy legislation. In *Proceedings of the Workshop on Computational Semantics (IWCS-7), Tilburg*, 2007.

[10] Núria Casellas. *Legal Ontology Engineering*, volume 3 of *Law, Governance and Technology*. Springer-Verlag, 2011.

[11] Nuria Casellas, Thomas R. Bruce, Sara. S. Frug, Sarah Bouwman, Dallas Dias, Jie Lin, Sharvari Marathe, Krithi Rai, Ankit Singh, Debraj Sinha, and Sanjna Venkataraman. Linked legal data: improving access to regulations. In John Carlo Bertot, Luis F. Luna-Reyes, and Sehl Mellouli, editors, *DG.O*, pages 280–281. ACM, 2012.

[12] Núria Casellas, Pompeu Casanovas, Joan-Josep Vallbé, Marta Poblet, Mercedes Blázquez, Jesús Contreras, José Manuel López-Cobo, and V. Richard Benjamins. Semantic enhancement for legal information retrieval: Iuriservice performance. In *Proceedings of the Eleventh International Conference on Artificial Intelligence and Law. ICAIL 2007, June 4-8, Stanford Law School, California*, pages 49–57. Association for Computing Machinery, 2007.

[13] Núria Casellas, Juan Emilio Nieto, Albert Meroño, Antoni Roig, Sergi Torralba, Mario Reyes de los Mozos, and Pompeu Casanovas. Ontological semantics for data privacy compliance: the neurona ontology. In *AAAI Spring Symposium Series Technical Reports SS-10-05 (Intelligent Information Privacy Management), Stanford 23rd-25th of March 2010*, pages 34–38. AAAI Press, 2010.

[14] Núria Casellas, Joan-Josep Vallbé, and Thomas R. Bruce. From legal information to open legal data: A case study in u.s. federal legal information. In *AAAI Fall Symposium Series (Open Government Knowledge: AI Opportunities and Challenges) Technical Report*, 2011.

[15] Claudia Cevenini, Giuseppe Contissa, Migle Laukyte, Contact Information, Régis Riveret, and Rossella Rubino. Development of the alis ip ontology: Merging legal and technical perspectives. In *Computer-Aided Innovation (CAI)*, volume 277 of *IFIP International Federation for Information Processing*, pages 169–180. Springer Boston, 2008.

[16] Chin Pang Cheng, Gloria T. Lau, Kincho H. Law, Jiayi Pan, and Albert Jones. Regulation retrieval using industry specific taxonomies. volume 16, pages 277–303, Hingham, MA, USA, September 2008. Kluwer Academic Publishers.

[17] Chin Pang Cheng, Jiayi Pan, Gloria T. Lau, Kincho H. Law, and Albert Jones. Relating taxonomies with regulations. In *Proceedings of the 2008 international conference on Digital government research*, dg.o '08, pages 34–43. Digital Government Society of North America, 2008.

[18] Philipp Cimiano, Johana Völker, and Rudi Studer. Ontologies on demand? a description of the state-of-the-art, applications, challenges and trends for ontology learning from text. *Information Wissenschaft &*, 57(6-7):315–320, 2006.

[19] Philipp Cimiano and Johanna Völker. Text2onto - a framework for ontology learning and data-driven change discovery. In Andres Montoyo, Rafael Munoz, and Elisabeth Metais, editors, *Proceedings of the 10th International Conference on Applications of Natural Language to Information Systems (NLDB)*, volume 3513 of *Lecture Notes in Computer Science*, pages 227–238, Alicante, Spain, 2005. Springer.

[20] Marie-Catherine de Marneffe, Bill MacCartney, and Christopher D. Manning. 2006. Generating typed dependency parses from phrase structure parses. In *Proceedings of the Inter-*

national Conference on Language Resources and Evaluation (LREC 2006), 2006.

[21] J. Delgado, I. Gallego, S. Llorente, and R. García. Ipronto: An ontology for digital rights management. In D. Bourcier, editor, *Legal Knowledge and Information Systems. Jurix 2003: The Sixteenth Annual Conference*, 111-120, Amsterdam, 2003. IOS Press.

[22] Enrico Francesconi, Simonetta Montemagni, Wim Peters, and Daniela Tiscornia. Integrating a bottom-up and top-down methodology for building semantic resources for the multilingual legal domain. In *Semantic Processing of Legal Texts*, pages 95–121, 2010.

[23] Enrico Francesconi, Simonetta Montemagni, Wim Peters, and Daniela Tiscornia, editors. *Semantic Processing of Legal Texts: Where the Language of Law Meets the Law of Language*, volume 6036 of *Lecture Notes in Computer Science*. Springer, 2010.

[24] Katerina T. Frantzi, Sophia Ananiadou, and Jun-ichi Tsujii. The c-value/nc-value method of automatic recognition for multi-word terms. In *Proceedings of the Second European Conference on Research and Advanced Technology for Digital Libraries*, ECDL '98, pages 585–604, London, UK, UK, 1998. Springer-Verlag.

[25] Aldo Gangemi, Maria-Teresa Sagri, and Daniela Tiscornia. A constructive framework for legal ontologies. In V. R. Benjamins, P. Casanovas, J. Breuker, and A. Gangemi, editors, *Law and the Semantic Web. Legal Ontologies, Methodologies, Legal Information retrieval, and Applications*, volume 3369 of *Lecture Notes in Computer Science*, pages 97–124. Springer-Verlag Berlin Heidelberg, 2005.

[26] Roberto García. *A Semantic Web Approach to Digital Rights Management*. Doctorate in computer science and digital communication, Department of Technologies, Universitat Pompeu Fabra, Barcelona, November 2006.

[27] Asunción Gómez-Pérez and David Manzano-Macho. A survey of ontology learning methods and techniques. Ist-2000-29243 deliverable 1.5, OntoWeb: Ontology-based Information Exchange for Knowledge Management and Electronic Commerce IST EU Project, 2003.

[28] Asunción Gómez-Pérez and David Manzano-Macho. An overview of methods and tools for ontology learning from texts. *Knowl. Eng. Rev.*, 19(3):187–212, 2004.

[29] Marti A. Hearst. Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 14th conference on Computational linguistics - Volume 2*, COLING '92, pages 539–545, Stroudsburg, PA, USA, 1992. Association for Computational Linguistics.

[30] Rinke Hoekstra, Joost Breuker, Marcello Di Bello, , and Alexander Boer. The lkif core ontology of basic legal concepts. In Pompeu Casanovas, Maria Angela Biasiotti, Enrico Francesconi, and Maria Teresa Sagri, editors, *Proceedings of the Workshop on Legal Ontologies and Artificial Intelligence Techniques (LOAIT 2007) at the International Conference on AI and Law (ICAIL'07) Stanford, USA, June 4*, pages 43–63, 2007.

[31] Rinke Hoekstra, Joost Breuker, Marcello Di Bello, and Alexander Boer. LKIF core: Principled ontology development for the legal domain. In Joost Breuker, Pompeu Casanovas, Michel C. A. Klein, and Enrico Francesconi, editors, *Law, Ontologies and the Semantic Web. Channelling the Legal Information Flood*, volume 188 of *Frontiers in Artificial Intelligence and Applications*, pages 21–52. IOS Press, 2009.

[32] Mustafa Jarrar. *Towards Methodological Principles for Ontology Engineering*. Doctor of philosophy, Vrije Universiteit Brussel, May 2005.

[33] Koen Kerremans and Gang Zhao. Topical ontology for vat. Deliverable of The FFPOIROT IP project (IST-2001-38248) Deliverable D2.3 (WP 2), STARLab VUB, July 2005.

[34] C. M. Kerwin. *The management of regulation development: out of the shadows*. Presidential Transition Series. IBM Center for The Business of Government, Washington, DC, 2008.

[35] Atanas Kiryakov, Borislav Popov, Ivan Terziev, Dimitar Manov, and Damyan Ognyanoff. Semantic annotation, indexing, and retrieval. *J. Web Sem.*, 2(1):49–79, 2004.

[36] Dan Klein and Christopher D. Manning. Accurate unlexicalized parsing. In *Proceedings of the 41st Meeting of the Association for Computational Linguistics*, pages 423–430, 2003.

[37] Guiraude Lame. *Construction d'ontologie a partie de textes. Une ontologie du droit dediée à la recherche d'information sur le Web*. PhD thesis, L'école des Mines de Paris, Décembre 2002.

[38] Guiraude Lame. Using nlp techniques to identify legal ontology components: Concepts and relations. *Artificial Intelligence and Law*, 12(4), 2004.

[39] Guiraude Lame. Using nlp techniques to identify legal ontology components: concepts and relations. In V. Richard Benjamins, Pompeu Casanovas, Joost Breuker, and Aldo Gangemi, editors, *Law and the Semantic Web. Legal Ontologies, Methodologies, Legal Information retrieval, and Applications*, volume 3369 of *Lecture Notes in Computer Science*, pages 169–184. Springer-Verlag Berlin Heidelberg, 2005.

[40] Gloria T. Lau, Kincho H. Law, and Gio Wiederhold. Legal information retrieval and application to e-rulemaking. In *Proceedings of the 10th international conference on Artificial intelligence and law*, ICAIL '05, pages 146–154, New York, NY, USA, 2005. ACM.

[41] Christian Mader, Bernhard Haslhofer, and Antoine Isaac. Finding quality issues in skos vocabularies. *CoRR*, abs/1206.1339, 2012.

[42] Mitchell P. Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini. Building a large annotated corpus of english: the penn treebank. *Comput. Linguist.*, 19(2):313–330, June 1993.

[43] L. Thorne McCarty. Deep semantic interpretations of legal texts. In *Proceedings of the 11th international conference on Artificial intelligence and law*, ICAIL '07, pages 217–224, New York, NY, USA, 2007. ACM.

[44] John McClure. The legal-rdf ontology. a generic model for legal documents. In Pompeu Casanovas, Maria Angela Biasiotti, Enrico Francesconi, and Maria Teresa Sagri, editors, *Proceedings of LOAIT 2007 at ICAIL'07, Stanford, USA, June 4*, pages 25–42, 2007.

[45] Richard J. McKinney. A research guide to the federal register and the code of federal regulations. *Law Library Lights*, 46:10–15 [updated November 2006], 2002.

[46] Albert Meroño-Peñuela, Núria Casellas, Sergi Torralba, Mario Reyes de los Mozos, and Pompeu Casanovas. Legal compliance in an ontology-based information system (poster). In *EKAW 2010 - Knowledge Engineering and Knowledge Management by the Masses 11th October-15th October 2010 - Lisbon, Portugal*, 2010.

[47] Ahsan Morshed, Caterina Caracciolo, Gudrun Johannsen, and Johannes Keizer. Thesaurus alignment for linked data publish-

ing. In *DCMI International Conference on Dublin Core and Metadata Applications DC-2011.*, 2011.

[48] M. Muller. Legal rdf dictionary. In R.G.F. Winkels, editor, *Proceedings of the Second International Workshop on Legal Ontologies (LEGONT) in JURIX 2001, Amsterdam (Netherlands)*, pages 20–21, Amsterdam, Netherlands, 2001.

[49] Nadia Nadah, Mélanie Dulong de Rosnay, and Bruno Bachimont. Licensing digital content with a generic ontology: Escaping from the jungle of rights expression languages. In *Proceedings of the Eleventh International Conference on Artificial Intelligence and Law (ICAIL 2007), June 4-8, Palo Alto, CA, USA*, pages 65–69. ACM Press, New York, 2007.

[50] Roberto Navigli and Paola Velardi. From glossaries to ontologies: Extracting semantic structure from textual definitions. In *Proceedings of the 2008 conference on Ontology Learning and Population: Bridging the Gap between Text and Knowledge*, pages 71–87, Amsterdam, The Netherlands, The Netherlands, 2008. IOS Press.

[51] Patrick Pantel and Dekang Lin. A statistical corpus-based term extractor. In *Proceedings of the 14th Biennial Conference of the Canadian Society on Computational Studies of Intelligence: Advances in Artificial Intelligence*, AI '01, pages 36–46, London, UK, UK, 2001. Springer-Verlag.

[52] Wim Peters. Text-based legal ontology enrichment. In Núria Casellas, Enrico Francesconi, Rinke Hoekstra, and Simonetta Montemagni, editors, *3rd Workshop on Legal Ontologies and Arificial Intelligence Techniques joint with 2nd Workshop on Semantic Processing of Legal Text (LOAIT 2009), Co-located with the 12th International Conference on Artificial Intelligence and Law (ICAIL 2009).*, volume 2 of *IDT Series*, pages 55–65. IDT/Huygens Editorial, 2009.

[53] Wim Peters, Maria-Teresa Sagri, and Daniela Tiscornia. The structuring of legal knowledge in lois. *Artificial Intelligence and Law*, 15(2):117–135, 2007.

[54] Luis Polo, José María Álvarez, and Emilio Rubiera Azcona. Promoting government controlled vocabularies to the semantic web: EUROVOC thesaurus and cpv product classification scheme. In *Proceedings of SIEDL2008 at ESWC2008, Tenerife, Spain, June 2, 2008.*, pages 111–122, 2008.

[55] Long Qiu, Min-Yen Kan, and Tat-Seng Chua. (a public reference implementation of the rap anaphora resolution algorithm. In *In proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC 2004)*, volume 1, pages 291–294, 2004.

[56] Robert C. Richards and Thomas R. Bruce. Adapting specialized legal metadata to the digital environment: The code of federal regulations parallel table of authorities and rules. In *Proceedings of the 13th International Conference on AI and Law (ICAIL 2011)*, Pittsburgh, PA, June 2011 2011.

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

[58] David Sánchez-Ruenes. *Domain Ontology Learning from the Web*. Tesi doctoral programa de doctorat en intel·ligència artificial, Departament de Llenguatges i Sistemes Informàtics, Universitat Politècnica de Catalunya, 2007.

[59] Giovanni Sartor, Pompeu Casanovas, Mariangela Biasiotti, and Meritxell Fernández-Barrera. *Approaches to Legal Ontologies. Theories, Domains and Methodologies*. Number 1 in Law, Governance and Technology Series. Springer, 2011.

[60] Osma Suominen and Eero Hyvönen. Improving the quality of skos vocabularies with skosify. In *Proceedings of the 18th International Conference on Knowledge Engineering and Knowledge Management (EKAW 2012), Springer-Verlag, Galway, Ireland, October, 2012. [To be published]*, 2012.

[61] Daniela Tiscornia. The lois project: Lexical ontologies for legal information sharing. In Carlo Biagioli, Enrico Francesconi, and Giovanni Sartor, editors, *Proceedings of the V Legislative XML Workshop*, pages 189–204. European Press Academic Publishing, 2007.

[62] Elisabeth M. Uijttenbroek, Arno R. Lodder, Michel C.A. Klein, Gwen R. Wildeboer, Wouter Van Steenbergen, Rory L.L. Sie, Paul E.M. Huygen, and Frank van Harmelen. Retrieval of case law to provide layman with information about liability: Preliminary results of the best-project. In Pompeu Casanovas, Giovanni Sartor, Núria Casellas, and Rossella Rubino, editors, *Computable Models of the Law*, volume 4884 of *Lecture Notes in Artificial Intelligence*, pages 291–311. Springer-Verlag, Berlin Heidelberg, 2008.

[63] A. Valente. *Legal Knowledge Engineering; A Modelling Approach*. IOS Press, Amsterdam, The Netherlands., 1995.

[64] Joan-Josep Vallbé. *Facing Uncertainty: Institutional and Cognitive Constraints in Spanish Junior Judges Decision-Making*. Phd thesis, Department of Constitutional Law and Political Science, Universitat de Barcelona, Barcelona (Spain), July 2009.

[65] Mark van Assem, Véronique Malaisé, Alistair Miles, and Guus Schreiber. A method to convert thesauri to skos. In York Sure and John Domingue, editors, *ESWC*, volume 4011 of *Lecture Notes in Computer Science*, pages 95–109. Springer, 2006.

[66] Robert W. van Kralingen. *Frame-Based Conceptual Models of Statute Law*. Kluwer Law Intl, 1995.

[67] Giulia Venturi. Semantic processing of legal texts. chapter Legal language and legal knowledge management applications, pages 3–26. Springer-Verlag, Berlin, Heidelberg, 2010.

[68] Johanna V´olker, Sergi Fernández Langa, and York Sure. Computable models of the law. chapter Supporting the Construction of Spanish Legal Ontologies with Text2Onto, pages 105–112. Springer-Verlag, Berlin, Heidelberg, 2008.

[69] Gwen R. Wildeboer, Michel C.A. Klein, and Elisabeth Uijttenbroek. Explaining the relevance of court decisions to laymen. In A. R. Lodder and L. Mommers, editors, *Legal Knowledge and Information Systems. JURIX 2007: The Twentieth Annual Conference*, volume 165 of *Frontiers in Artificial Intelligence and Applications*, pages 129–138. IOS Press, 2007.

[70] R. Winkels, A. Boer, and R. Hoekstra. Clime: Lessons learned in legal information serving. In *Proceedings of the 15th European Conference on Artificial Intelligence (ECAI 2002), Lyon, France*, pages 230–234, Amsterdam, The Nederlands, 2002. IOS Press.

[71] Adam Wyner and Rinke Hoekstra. A legal case owl ontology with an instantiation of popov v. hayashi. *The Knowledge Engineering Review*, 14(2):1–24, 2010.

[72] Gang Zhao and Richard Leary. Topical ontology of fraud. Deliverable of The FFPOIROT IP project (IST-2001-38248) Deliverable D2.3 (WP 2), STARLab VUB, August 2005.

Appendix

A. Excerpt RDF representation of "milk" in the CFR SKOS vocabulary[59]

```
<skos:Concept rdf:about="condensed_sweetened_milk">
  <skos:prefLabel>condensed sweetened milk</skos:prefLabel>
  <skos:inScheme rdf:resource="Title21"/>
  <skos:broader  rdf:resource="sweetened_milk"/>
</skos:Concept>
<skos:Concept rdf:about="reconstituted_milk">
  <skos:prefLabel>reconstituted milk</skos:prefLabel>
  <skos:inScheme rdf:resource="Title21"/>
  <skos:broader  rdf:resource="milk"/>
</skos:Concept>
<skos:Concept rdf:about="slaughter_milk">
  <skos:prefLabel>slaughter milk</skos:prefLabel>
  <skos:inScheme rdf:resource="Title21"/>
  <skos:broader  rdf:resource="milk"/>
</skos:Concept>
<skos:Concept rdf:about="contamination_of_milk">
  <skos:prefLabel>contamination of milk</skos:prefLabel>
  <skos:inScheme rdf:resource="Title21"/>
  <skos:broader  rdf:resource="contamination"/>
      <skos:related  rdf:resource="milk"/>
</skos:Concept>
<skos:Concept rdf:about="special_milk">
  <skos:prefLabel>special milk</skos:prefLabel>
  <skos:inScheme rdf:resource="Title21"/>
  <skos:broader  rdf:resource="milk"/>
</skos:Concept>
<skos:Concept rdf:about="milk_for_feeding">
  <skos:prefLabel>milk for feeding</skos:prefLabel>
  <skos:inScheme rdf:resource="Title21"/>
  <skos:broader  rdf:resource="milk"/>
  <skos:related  rdf:resource="water"/>
</skos:Concept>
<skos:Concept rdf:about="reconstituted_skim_milk">
  <skos:prefLabel>reconstituted skim milk</skos:prefLabel>
  <skos:inScheme rdf:resource="Title21"/>
  <skos:broader  rdf:resource="skim_milk"/>
</skos:Concept>
<skos:Concept rdf:about="nondairy_milk">
  <skos:prefLabel>nondairy milk</skos:prefLabel>
  <skos:inScheme rdf:resource="Title21"/>
  <skos:broader  rdf:resource="milk"/>
  <skos:related  rdf:resource="nondairy_product"/>
</skos:Concept>
<skos:Concept rdf:about="dry_milk_cream">
  <skos:prefLabel>dry milk cream</skos:prefLabel>
  <skos:inScheme rdf:resource="Title21"/>
  <skos:broader  rdf:resource="milk_cream"/>
</skos:Concept>
<skos:Concept rdf:about="nonfat_milk_cream">
  <skos:prefLabel>nonfat milk cream</skos:prefLabel>
  <skos:inScheme rdf:resource="Title21"/>
  <skos:broader  rdf:resource="milk_cream"/>
</skos:Concept>
<skos:Concept rdf:about="concentrated_skim_milk">
  <skos:prefLabel>concentrated skim milk</skos:prefLabel>
  <skos:inScheme rdf:resource="Title21"/>
  <skos:broader  rdf:resource="skim_milk"/>
  <skos:related  rdf:resource="sweet_skim_milk"/>
  <skos:related  rdf:resource="nonfat_milk"/>
  <skos:related  rdf:resource="dry_milk"/>
</skos:Concept>
<skos:Concept rdf:about="sulfonamide_milk">
  <skos:prefLabel>sulfonamide milk</skos:prefLabel>
  <skos:inScheme rdf:resource="Title21"/>
  <skos:broader  rdf:resource="milk"/>
</skos:Concept>
<skos:Concept rdf:about="animal_milk">
  <skos:narrower  rdf:resource="producing_animal_milk"/>
  <skos:prefLabel>animal milk</skos:prefLabel>
  <skos:inScheme rdf:resource="Title21"/>
  <skos:broader  rdf:resource="milk"/>
</skos:Concept>
<skos:Concept rdf:about="milk_production">
  <skos:narrower  rdf:resource="increased_milk_production"/>
  <skos:prefLabel>milk production</skos:prefLabel>
  <skos:inScheme rdf:resource="Title21"/>
  <skos:broader  rdf:resource="production"/>
</skos:Concept>
<skos:Concept rdf:about="buttermilk_coating">
  <skos:narrower  rdf:resource="chocolate_buttermilk_coating"/>
  <skos:prefLabel>buttermilk coating</skos:prefLabel>
  <skos:inScheme rdf:resource="Title21"/>
  <skos:broader  rdf:resource="coating"/>
</skos:Concept>
<skos:Concept rdf:about="dry_milk">
  <skos:prefLabel>dry milk</skos:prefLabel>
  <skos:inScheme rdf:resource="Title21"/>
  <skos:broader  rdf:resource="milk"/>
  <skos:related  rdf:resource="sweet_skim_milk"/>
  <skos:related  rdf:resource="identity_for_dry_milk"/>
  <skos:related  rdf:resource="skim_milk"/>
  <skos:related  rdf:resource="milk_with_water"/>
  <skos:related  rdf:resource="flavored_milk"/>
  <skos:related  rdf:resource="sufficient_milk"/>
  <skos:related  rdf:resource="concentrated_skim_milk"/>
</skos:Concept>
<skos:Concept rdf:about="human_milk">
  <skos:prefLabel>human milk</skos:prefLabel>
  <skos:inScheme rdf:resource="Title21"/>
  <skos:broader  rdf:resource="milk"/>
  <skos:related  rdf:resource="risk_to_infant"/>
  <skos:related  rdf:resource="associated_risk"/>
  <skos:related  rdf:resource="drug_in_human_milk"/>
  <skos:related  rdf:resource="excretion_in_human_milk"/>
</skos:Concept>
<skos:Concept rdf:about="milk">
  <skos:prefLabel>milk</skos:prefLabel>
  <skos:inScheme rdf:resource="Title21"/>
      <skos:narrower  rdf:resource="evaporated_milk"/>
  <skos:narrower  rdf:resource="calf_milk"/>
  <skos:narrower  rdf:resource="milk_for_feeding"/>
  <skos:narrower  rdf:resource="cattle_milk"/>
  <skos:narrower  rdf:resource="pasteurized_milk"/>
  <skos:narrower  rdf:resource="feed"/>
  <skos:narrower  rdf:resource="skim_milk"/>
  <skos:narrower  rdf:resource="holding_milk"/>
  <skos:narrower  rdf:resource="nonfat_milk"/>
  <skos:narrower  rdf:resource="milk_for_human_consumption"/>
  <skos:narrower  rdf:resource="dried_milk"/>
  <skos:narrower  rdf:resource="whole_milk"/>
  <skos:narrower  rdf:resource="optional_dairy_ingredient"/>
  <skos:narrower  rdf:resource="special_milk"/>
  <skos:narrower  rdf:resource="sweetened_milk"/>
  <skos:narrower  rdf:resource="nondairy_milk"/>
  <skos:narrower  rdf:resource="concentrated_milk"/>
  <skos:narrower  rdf:resource="veterinarian_milk"/>
  <skos:narrower  rdf:resource="condensed_milk"/>
  <skos:narrower  rdf:resource="sulfonamide_milk"/>
  <skos:narrower  rdf:resource="malted_milk"/>
  <skos:narrower  rdf:resource="reconstituted_milk"/>
  <skos:narrower  rdf:resource="feeding_milk"/>
  <skos:narrower  rdf:resource="human_milk"/>
  <skos:narrower  rdf:resource="flavored_milk"/>
  <skos:narrower  rdf:resource="dry_milk"/>
  <skos:narrower  rdf:resource="animal_milk"/>
  <skos:narrower  rdf:resource="sufficient_milk"/>
  <skos:narrower  rdf:resource="buffalo_milk"/>
  <skos:narrower  rdf:resource="post-parturition_milk"/>
  <skos:narrower  rdf:resource="treatment_milk"/>
  <skos:narrower  rdf:resource="solids-corrected_milk"/>
  <skos:narrower  rdf:resource="marketable_milk"/>
  <skos:related  rdf:resource="cream"/>
  <skos:related  rdf:resource="egg-producing_animal"/>
  <skos:related  rdf:resource="residue_in_milk"/>
  <skos:related  rdf:resource="milkfat_from_milk"/>
  <skos:related  rdf:resource="fat_in_milk"/>
  <skos:related  rdf:resource="cattle_for_milk"/>
  <skos:related  rdf:resource="chocolate"/>
  <skos:related  rdf:resource="uncooked_tissue"/>
  <skos:related  rdf:resource="edible_product"/>
  <skos:related  rdf:resource="time_for_milk"/>
  <skos:related  rdf:resource="tissue"/>
  <skos:related  rdf:resource="cheese_from_milk"/>
  <skos:related  rdf:resource="milkfat"/>
      [...]
</skos:Concept>
```

[59] English language tags and namespaces have been eliminated due to page width limitations.

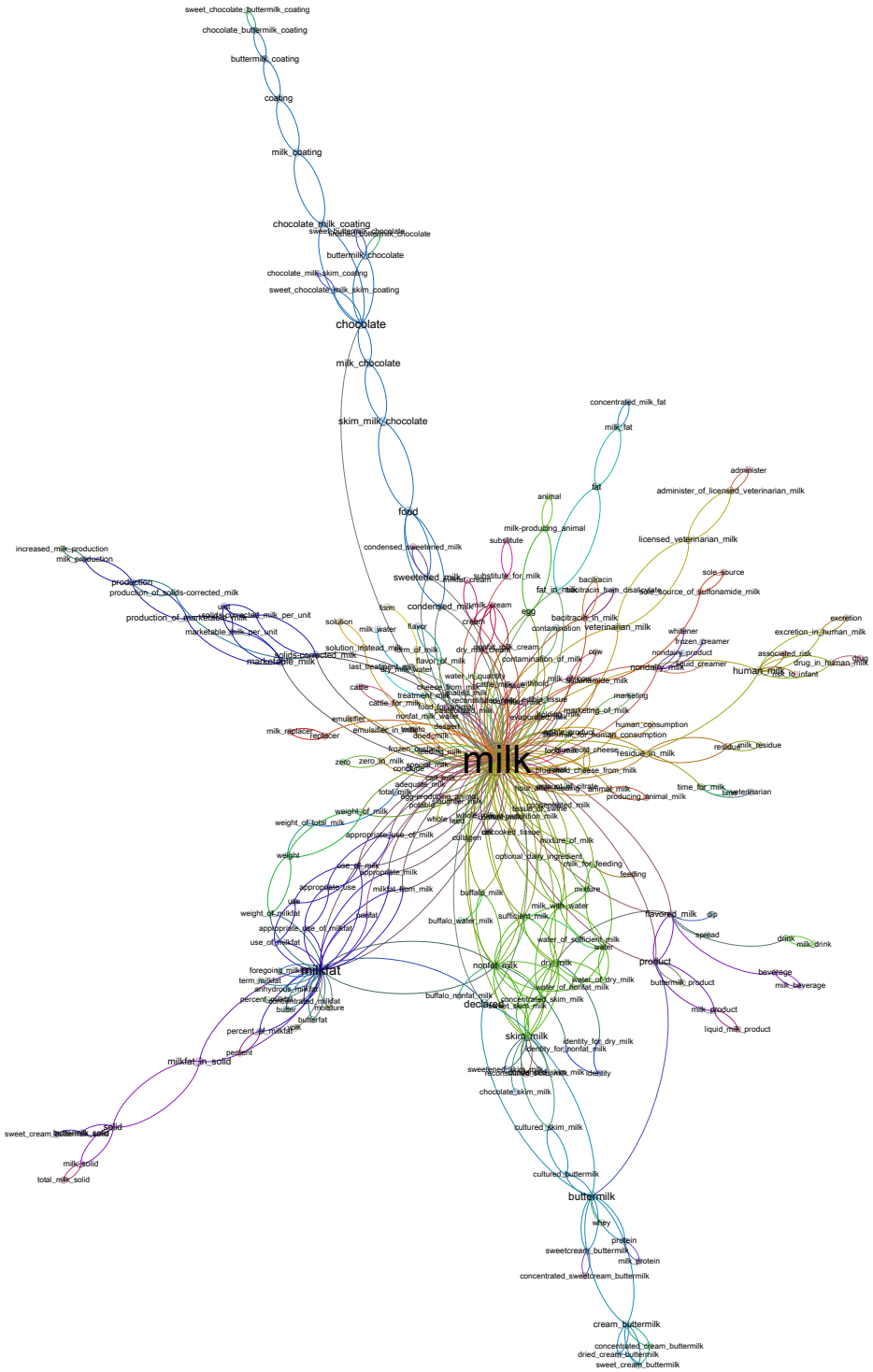


Fig. 2. A graph visualization of the concept “milk” and related concepts