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 Tecnologia, 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 enhancement 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, linked data, 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].

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 regulatory materials (e.g. DrugBank database).

After a brief 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, 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 further 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”, 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, RDFS, and OWL 1 and 2, or the SPARQL query language 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. 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.

2.1. Legal Vocabularies, Taxonomies and Ontologies

In the legal domain the analysis and use of controlled vocabularies, taxonomies, and ontologies to support legal information search and retrieval is exten-

1W3C Semantic Web documentation: http://www.w3.org/standards/semanticweb/.
4OWL 2 Web Ontology Language: http://www.w3.org/TR/owl2-primer.
5SPARQL Query Language: http://www.w3.org/TR/sparql11-query.
6http://www.w3.org/DesignIssues/LinkedData.html.
7For a discussion on the meaning and the evolution of concept of computational ontology, see [10].
sive, 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-(Ita)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 codified legislative 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]. The Ontology of Professional Judicial Knowledge (OPJK) was developed to enhance the search and retrieval capabilities 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 in non-compliance cases (the Customer Complaint Ontology or CContology, [32], and the Consumer Protection Ontology developed within the DALOS project by [11]).

The 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, some commercial legal databases provide semantically-enhanced search, although little information is available with regards to the technical details of the knowledge bases: LawMoose,12 LexisNexis TotalPatentTM 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

8For a complete description of existing legal ontologies see [10, 59].
9The 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.
10This ontology is available at: http://wyner.info/research/ontologies/LegalCaseOntology_v9.owl.
15SKOS: http://www.w3.org/2004/02/skos.
17The 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). 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 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), which 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. For example, towards this end, some preliminary work has been proposed towards the reuse in RDF of the Parallel Table of Authorities.\(^24\)

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 faceted 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 (e.g. building and construction) by keyword matching and structure reuse to cluster relevant sections from multiple regulations.\(^16,17\).

\(^19\)The CFR is updated once per year on a regular basis. For more information see: http://www.archives.gov/federal-register/cfr/about.html.
\(^20\)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 domains 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, and the AGROVOC thesaurus 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 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 extends its content to other related areas; DrugBank could be reused for pharmaceutical and drug related terms. Linked Life Data could provide terminologies for the biomedical domain, aerospace-related terms could be reused from the NASA taxonomy, economic terms could be reused from the STW Thesaurus for Economics, 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 LCSH, GLIN, the Government of Canada Core Subject Thesaurus, Dbpedia categories or NY Times subjects, could be reused. However, most of these vocabularies present issues 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 contains 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

---

26AGROVOC: http://aims.fao.org/agrovoc/1od.
33LCSH: http://id.loc.gov.
34Government of Canada Core Subject Thesaurus: http://www.thesaurus.gc.ca.
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

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

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

Table 1
List of SKOS vocabularies for possible reuse

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

§ 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. [...]
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 some terms in the Thesaurus are appended with a broad subject category, most terms do not contain such reference. Furthermore, the 19 terms are not included in the Thesaurus.

3.2.2. **Step B: map data items to SKOS**

Terms in entries were modeled as skos:Concept types \(^{[14]}\). Although the documentation did not contain clear definitions of the meaning of the different types of relationships between the terms, from the overall analysis of the Thesaurus, the following assumptions were established. First, the use of *see* in “A see B” relationships appeared to generally refer to the use of preferred and non-preferred terms, thus mapping to the SKOS skos:prefLabel and skos:altLabel properties. In this mapping, one of the terms (the non-preferred) is, in the end, included in the vocabulary as an alternative label; only one skos:Concept is created from this structure. Second, the usage of *sa*, as *see also*, was 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 and skos:topConceptOf within the Code of Federal Regulations Thesaurus.


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 conversion 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 SKOS conversion results and their automatic validation 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 and hierarchical links in some resources, due to an inconsistent use of inverse properties. In particular, it appeared the pattern “A see B” was generally followed by the inverse “B x A” (has narrower). While the term A was being represented as the string literal of a skos:altLabel for the skos:Concept, a B skos:narrower A statement was created at the same time, generating also a resource for entry A.

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”.

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

---

30 SKOS: http://www.w3.org/TR/skos-reference.
31 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; skosify: http://code.google.com/p/skosify.
Table 3
Mapping of Thesaurus features into SKOS properties

<table>
<thead>
<tr>
<th>Data item</th>
<th>Feature/function</th>
<th>Representation</th>
</tr>
</thead>
<tbody>
<tr>
<td>A see B</td>
<td>Preferred and non-preferred term</td>
<td>A a skos:Concept ; skos:altLabel &quot;A&quot; ; skos:prefLabel &quot;B&quot; .</td>
</tr>
<tr>
<td>A sa B</td>
<td>Related terms</td>
<td>A a skos:Concept ; skos:related B .</td>
</tr>
<tr>
<td>A xx B</td>
<td>Broader term</td>
<td>A a skos:Concept ; skos:broader B .</td>
</tr>
<tr>
<td>A x B</td>
<td>Narrower term</td>
<td>A a skos:Concept ; skos:narrower B .</td>
</tr>
<tr>
<td>N {grouping (A, B, C, etc.)}</td>
<td>Top concept and narrower terms</td>
<td>N a skos:Concept ; skos:topConceptOf S ; skos:narrower A .</td>
</tr>
<tr>
<td>19 subject categories</td>
<td>concept scheme and top concepts</td>
<td>S a skos:ConceptScheme ; skos:hasTopConcept N .</td>
</tr>
</tbody>
</table>

Table 4
Some evaluation results

<table>
<thead>
<tr>
<th>Features</th>
<th>Own method</th>
<th>qSKOS</th>
<th>skosify</th>
</tr>
</thead>
<tbody>
<tr>
<td>orphan concepts</td>
<td>N/A</td>
<td>8</td>
<td>12</td>
</tr>
<tr>
<td>hierarchy cycles</td>
<td>58</td>
<td>21</td>
<td>62</td>
</tr>
<tr>
<td>hierarchy vs associative links</td>
<td>780</td>
<td>709</td>
<td>743</td>
</tr>
<tr>
<td>label conflicts</td>
<td>266</td>
<td>86</td>
<td>N/A</td>
</tr>
</tbody>
</table>

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 thesaurus 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 multilingual variants), concepts,\textsuperscript{42} taxonomical or non-hierarchical relations, and rules \textsuperscript{8}.\textsuperscript{43} 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 integrate or combine these techniques have been developed for ontology extraction, population and semantic indexing: such as GATE’s ANNI (General Architecture for Text Engineering),\textsuperscript{44} TerMine,\textsuperscript{45} the KIM Platform,\textsuperscript{46} or Text2Onto.\textsuperscript{47}

\textsuperscript{42}While terms may be generally understood as “linguistic realizations of domain-specific concepts”, “[t]he extraction of concepts from text is controversial” \textsuperscript{8}.

\textsuperscript{43}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.

\textsuperscript{44}GATE: http://gate.ac.uk.

\textsuperscript{45}TerMine: http://www.nactem.ac.uk/software/termine.

\textsuperscript{46}KIM: http://www.ontotext.com/kim.

\textsuperscript{47}Text2Onto: http://code.google.com/p/text2onto.
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], and to support deep semantic analysis interpretation of legal texts [43], the e-discovery process,\(^48\) or syntactic and lexical comparison of Italian legal corpora [67], within others.\(^49\)

### 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 Regulations contains over 96.5 million words, therefore, these techniques are applied initially at 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.\(^50\)

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”.\(^51\)

Then, the Stanford Parser [36] for English language, a lexicalized probabilistic parser that also provides grammatical relations between the words of a sentence or typed dependencies (Stanford Dependencies, [20]) 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.\(^52\)

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

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

---


\(^{50}\)LII: [http://www.law.cornell.edu](http://www.law.cornell.edu).

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.33 For example, **nn** and **amod** are patterns that identify noun modifiers, noun and adjectival respectively. These 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 correspondent **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 governor of the dependency. Finally, the **conj** conjunct typed dependency represents a relationship between elements connected by a coordinating conjunction (e.g. and, or, etc.). This dependency pattern can be used to express the **skos:related** relationship between vocabulary terms. Table 5 shows a list of examples and conversion types; inverses and complete URIs are not included for brevity.

More complex lexico-syntactic structures are extracted using Hearst patterns, which support the identification of hyponymic and hyponymous 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.

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

```lisp
@prefix liivoc: <http://liicornell.org/> .
liivoc:public_body a skos:Concept ;
    skos:narrower liivoc:authority ;
    skos:narrower liivoc:county ;
    skos:narrower liivoc:district ;
    skos:narrower liivoc:authority .
```

Finally, the extraction 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 governor. 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_practitioner 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, approximately.54 Upon the analysis of specific concepts, the extraction yields interesting results and captures relevant terminological information (see, for example, the visualization of the concept “milk” in figure 2).

In comparison to the conversion of the Thesaurus of Indexing Terms, in this case, orphan concepts, direct 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 re-

54Although the frequency or relevance of terms is not taken into account to control the set of terminology extracted, 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 conversion examples

<table>
<thead>
<tr>
<th>Typed dependency</th>
<th>Example</th>
<th>SKOS conversion</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>nn</em>: noun compound modifier</td>
<td>nn(tobacco-22, smokeless-21)</td>
<td>tobacco a skos:Concept ;</td>
</tr>
<tr>
<td></td>
<td></td>
<td>skos:narrower smokeless_tobacco .</td>
</tr>
<tr>
<td><em>amod</em>: adjectival modifier</td>
<td>amod(consequences-43, life-threatening-42)</td>
<td>consequences a skos:Concept ;</td>
</tr>
<tr>
<td></td>
<td></td>
<td>skos:narrower life-threatening_consequences .</td>
</tr>
<tr>
<td><em>prep</em>: prepositional modifier &amp; <em>pobj</em>: object of preposition</td>
<td>prep(use-17, of-18) &amp; pobj(of-18, cigarettes-19)</td>
<td>use a skos:Concept ;</td>
</tr>
<tr>
<td></td>
<td></td>
<td>skos:narrower use_of_cigarretes .</td>
</tr>
<tr>
<td></td>
<td></td>
<td>cigarettes a skos:Concept ;</td>
</tr>
<tr>
<td></td>
<td></td>
<td>skos:related use_of_cigarretes .</td>
</tr>
<tr>
<td><em>conj</em>: conjunct</td>
<td>conj(children-30, adolescents-32)</td>
<td>children a skos:Concept ;</td>
</tr>
<tr>
<td></td>
<td></td>
<td>skos:related adolescents .</td>
</tr>
</tbody>
</table>

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 areas has improved significantly (see table 6 for results on the complete Title 21 vocabulary). 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.\(^{55}\) 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.

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. Tailored parser training and pre-processing of regulatory text pose challenging tasks, common to legal text analysis, due to the particularities of legal text ([4,37,39,67]).

\(^{55}\)Treebank: \url{http://www.cis.upenn.edu/~treebank}. 

---

(a) Records for manufacturers. Each person registered or authorized to manufacture controlled substances shall maintain records with the following information:

(i) For each controlled substance in bulk form to be used in, or capable of use in, the manufacture of the same or other controlled or noncontrolled substances in finished form, the name of the substance;

 [...] 

(iv) The quantity used to manufacture the same substance in finished form, including:

(A) The date and batch or other identifying
Table 6
Some evaluation results

<table>
<thead>
<tr>
<th>Features</th>
<th>Own method</th>
<th>qSKOS</th>
</tr>
</thead>
<tbody>
<tr>
<td>orphan concepts</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>hierarchy cycles</td>
<td>1114 (x2)</td>
<td>4</td>
</tr>
<tr>
<td>hierarchy vs associative links</td>
<td>0</td>
<td>7885</td>
</tr>
<tr>
<td>label conflicts</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

number of each manufacture;
[B] The quantity used in the manufacture;
[...]
[B] 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 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”. 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 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 or appendices or contained in images), for demonstration purposes, owl:sameAs mapping relationships can already be discovered through the string matching between some DrugBank drugs and CFR terms. Table 7 lists a sample of the possible mappings detected.

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 purposes 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 establishing 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) from the Code of Federal Regulations. These materials being also formalized in RDF to facilitate integration with the current vocabulary and enable access to regulatory linked data.

References

### Table 7

<table>
<thead>
<tr>
<th>Drug ID</th>
<th>CFR term ID</th>
<th>Drug ID</th>
<th>CFR term ID</th>
</tr>
</thead>
<tbody>
<tr>
<td>DB01625</td>
<td>isopropamide</td>
<td>DB01134</td>
<td>desoxycorticosterone_pivalate</td>
</tr>
<tr>
<td>DB00860</td>
<td>prednisolone</td>
<td>DB01075</td>
<td>diphenhydramine</td>
</tr>
<tr>
<td>DB00028</td>
<td>immune_globulin</td>
<td>DB00878</td>
<td>chlorhexidine</td>
</tr>
<tr>
<td>DB00410</td>
<td>mupirocin</td>
<td>DB00312</td>
<td>pentobarbital</td>
</tr>
<tr>
<td>DB04272</td>
<td>citric_acid</td>
<td>DB00119</td>
<td>pyruvic_acid</td>
</tr>
<tr>
<td>DB04183</td>
<td>methylmalonic_acid</td>
<td>DB00518</td>
<td>albendazole</td>
</tr>
<tr>
<td>DB00864</td>
<td>tobramycin</td>
<td>DB00971</td>
<td>selenium_sulfide</td>
</tr>
<tr>
<td>DB01677</td>
<td>fumarate</td>
<td>DB00446</td>
<td>chloramphenicol</td>
</tr>
<tr>
<td>DB05245</td>
<td>silver_sulfadiazine</td>
<td>DB02579</td>
<td>acrylic_acid</td>
</tr>
<tr>
<td>DB01093</td>
<td>dimethyl_sulfoxide</td>
<td>DB00479</td>
<td>amikacin</td>
</tr>
<tr>
<td>DB01160</td>
<td>dinoprost_trermethamine</td>
<td>DB04626</td>
<td>apramycin</td>
</tr>
<tr>
<td>DB03733</td>
<td>ethylene_dichloride</td>
<td>DB00821</td>
<td>carbopren</td>
</tr>
<tr>
<td>DB04566</td>
<td>inosinic_acid</td>
<td>DB01396</td>
<td>digoxitin</td>
</tr>
<tr>
<td>DB00919</td>
<td>spectinomycin</td>
<td>DB01213</td>
<td>fomepizole</td>
</tr>
<tr>
<td>DB00261</td>
<td>lidocaine</td>
<td>DB02640</td>
<td>fumagilin</td>
</tr>
<tr>
<td>DB00148</td>
<td>creatine</td>
<td>DB04077</td>
<td>glycerol</td>
</tr>
<tr>
<td>DB01174</td>
<td>phenobarbital</td>
<td>DB00602</td>
<td>ivermectin</td>
</tr>
<tr>
<td>DB01592</td>
<td>iron</td>
<td>DB01009</td>
<td>ketoprofen</td>
</tr>
<tr>
<td>DB00986</td>
<td>glycopyrrolate</td>
<td>DB00814</td>
<td>meloxicam</td>
</tr>
<tr>
<td>DB00107</td>
<td>oxytocin</td>
<td>DB00826</td>
<td>natamycin</td>
</tr>
<tr>
<td>DB00730</td>
<td>thiacarbendazole</td>
<td>DB01345</td>
<td>potassium</td>
</tr>
<tr>
<td>DB00440</td>
<td>trimethoprim</td>
<td>DB03904</td>
<td>urea</td>
</tr>
<tr>
<td>DB00825</td>
<td>menthol</td>
<td>DB00595</td>
<td>oxetacycine</td>
</tr>
<tr>
<td>DB00121</td>
<td>biotin</td>
<td>DB00729</td>
<td>diphenamid_methylsulfate</td>
</tr>
<tr>
<td>DB04257</td>
<td>palmitoleic_acid</td>
<td>DB04829</td>
<td>lysergic_acid_diethylamide</td>
</tr>
</tbody>
</table>
USA, 2005. ACM.


[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.


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