[LLD]

Linked Legal Data

Improving Access to Regulatory Information

Sarah Bouwman, Dallas Dias, Jie Lin, Sharvari Marathe, Krithi Rai, Ankit Singh, Debraj Sinha, Sanjna Venkataraman

Department of Computer Science

Núria Casellas

Legal Information Institute, Cornell Law School

Introduction

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 legal knowledge to individual citizens, businesses and government agencies and administrations, and allow sharing and reuse of legal information across applications, organizations and jurisdictions.

Goal

With this project, we will enhance access to the Code of Federal Regulations (a text with 96.5 million words in total; ~823MB XML file size) with an RDF dataset created with a number of semantic-search and retrieval applications and information extraction techniques based on the development and the reuse of RDF product taxonomies, the application of semantic matching algorithms between these materials and the CFR content (Syntactic and Semantic Mapping), the detection of product-related terms and relations (Vocabulary Extraction), obligations and product definitions (Definition and Obligations Extraction).

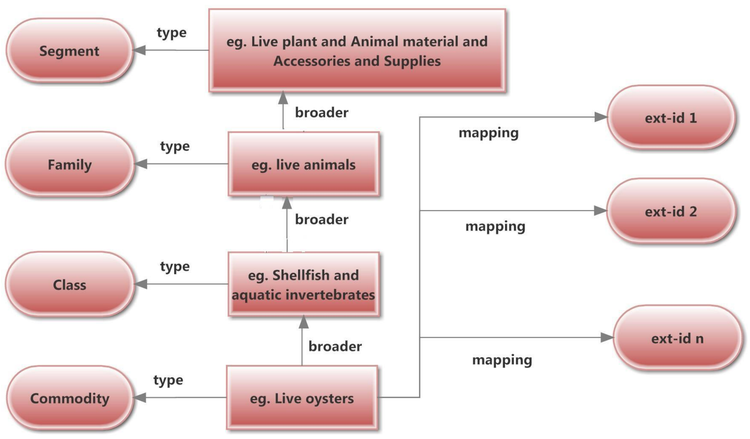
Methods and Initial Results

Mapping Products: What regulations apply to my product?

The reuse of product classifications, such as NAICS (North American Industry Classification System) and UNSPSC (United Nations Standard Products and Services Code), can support the discovery of tailored product regulatory information. How do we map these codes to the relevant sections of the CFR? For example, the term "sculpture" is represented as two different UNSPSC commodity codes; 1) 60121002 of the UNSPSC family "Arts and crafts equipment and accessories and supplies", and 2) 86131503 of the UNSPSC Family "Specialized educational services".

In order to create sets of RDF statements <code match section> for each product and industry label, we are exploring three different and successive mapping strategies [1]:

1. Mapping of sections which contain an exact match of all the words in the label;
2. Sections which contain all of the words in the label in any order.
3. Use the structural hierarchy inherent in RDF/SKOS (both at CFR structure and NAICS/UNSPSC levels) to disambiguate meaning and locate those sections which contain the label in the correct context.



Example of UNSPSC to CFR mapping

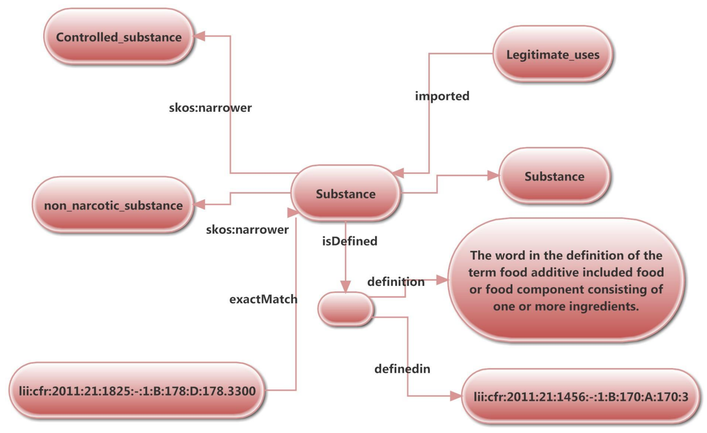
Vocabulary Extraction: How can search be improved?

The Code of Federal regulations is divided in 50 titles that represent broad areas subject to Federal regulation: agriculture, food and drugs, energy, etc. The improvement of finding aids that support full-text search and information aggregation can be immensely helpful to a broad population.

The goal of this task is to semi-automatically extract relevant concepts and relations from the CFR to create a vocabulary to support term expansion. We follow a bottom up approach with the Stanford Parser to obtain structured trees and dependencies from the CFR text. Stanford typed dependencies are used to exploit grammatical relations to construct the vocabulary. Hearst patterns support the identification of taxonomical relationships between the terms [2].

From these inputs, we are creating sets of RDF statements based on the following relationships:

1. skos:narrower/ skos:broader
2. skos:related
3. liivoc:*predicates* from the structure of the sentences N-V-N.



CFR RDF graph

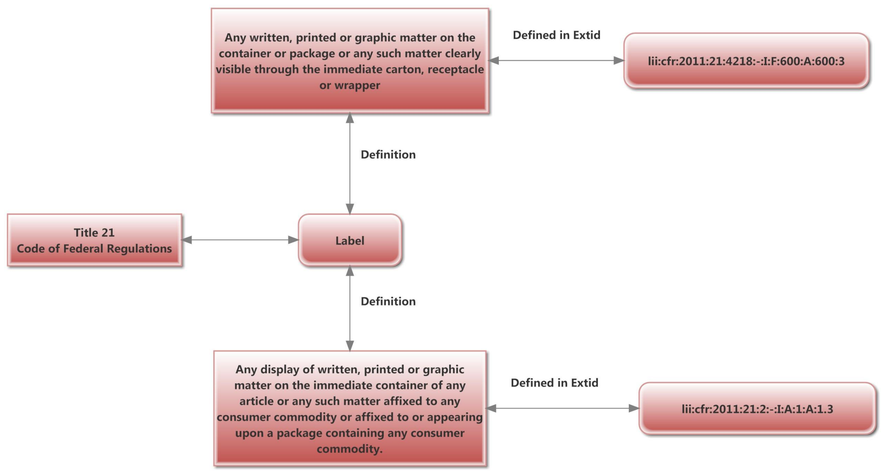
Definition Extraction: What is the meaning of "product"?

Many titles, chapters and parts include definitions for terms used within their descriptions. Some terms contain different definitions for different sections. Some definitions are relevant for more than one section. Some terms are defined in other sections, and their given definitions are only references.

We have taken the following approach to definition extraction from CFR text:

1. Detect sections whose titles include the word definitions in the CFR XML.
2. Use regular expressions to extract well formed definitions and generate an XML file.
3. Use an XML definition file to generate RDF statements that capture the relationship between the defined term, its definition and the section of the CFR that contains the definition of that defined term.

We are currently exploring the detection of the scope of the definition [3].



Definition extraction from CFR

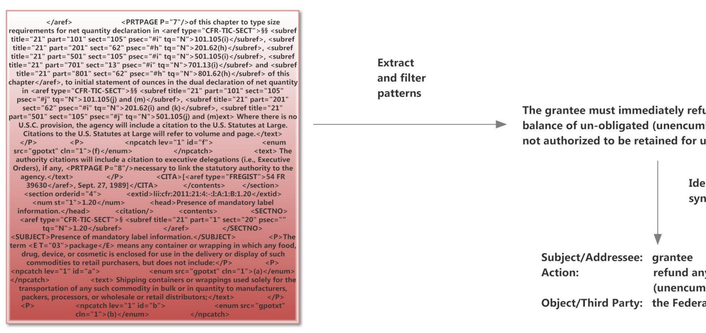
Obligation Extraction: What shall I do?

The goal of this task is to identify and extract obligations from the CFR, as well as the arguments of an obligation and represent it in an RDF format [4]. Given CFR text that contains an obligation of the form “X is obliged to do Y with regards to Z”, our system attempts to:

- Identify that this text is an obligation
- Identify the addressee of the obligation.
- Identify the action of the obligation.
- Identify the object/third party.

To achieve this, we use a multi-step process:

1. We identify sentences within the CFR that look like obligations using a pattern matching approach (presence of modals (shall/must) and other stemmed words that imply the presence of obligations: responsibility, obligation, require etc.) ;
2. To avoid false positives, we introduced further pattern matching constraints by identifying the semantic roles in a sentence.
3. Then, we use syntactic dependencies in the sentence to identify the addressee, action and object of the obligation.



Example of obligation extraction

Conclusions

We are currently evaluating and testing the results of the ongoing tasks of the Linked Legal Data project for the Code of Federal Regulations. Future work will focus on the semantic improvement of product and industry mappings, on the refinement of vocabulary extraction and ontology learning, on the detection of scope in definitions, and on the extraction of the bearer and the object of the obligations.

From these inputs we'll create an RDF dataset of the Code of Federal Regulations (structure, vocabulary, definitions, obligations) and link its contents to other collections of data: DrugBank, DBpedia, etc. From these results, we will build LOD-based applications to improve navigation, discovery and aggregation of the material in the CFR, enabling

References

[1] Pavel Shvaiko and Jérome Euzenat (2005). “A Survey of Schema-based Matching Approaches”, Journal of Data Semantics IV: 146-171.
[2] Philipp Cimiano and Johanna Vöelker, (2005): Text2Onto. A Framework for Ontology Learning and Data-driven Change Discovery Proceedings of the Applications of Natural Language to Information Systems (NLDB)' , pp. 227-238Roberto Navigli and Paola Velardi (2008). “From Glossaries to Ontologies: Extracting [3] Semantic Structure from Textual Definitions”, Proceeding of the 2008 conference on Ontology Learning and Population: Bridging the Gap between Text and Knowledge, 71--87, 2008, IOS Press
[4] Carlo Biagioli, Enrico Francesconi, Andrea Passerini, Simonetta Montemagni, and Claudia Soria. (2005). Automatic semantics extraction in law documents. In Proceedings of the 10th international conference on Artificial intelligence and law (ICAIL '05). ACM, New York, NY, USA, 133-140.

Acknowledgements

We would like to acknowledge the collaboration and work of Thomas R. Bruce (Director), Sara Frug, Daniel Nagy, David Shetland, and Wayne Weibel at the Legal Information Institute, and Professor Claire Cardie at the Department of Computer Science.