Linked Legal Data
Improving Access to Regulatory Information

Sarah Bouwman, Dallas Dias, Jie Lin, Sharvari Marathe, Kirthi Rai, Ankht Singh, Debaj Sinha, Sanjna Venkataratnam

Department of Computer Science
Núria Casellas
Legal Information Institute, Cornell Law School

Introduction

The application of Linked Open Data (LOD) principles to legal information (UR) naming of resources, assertions about named relationships between resources and data values, and the possibility to easily extend, update and modify these relationships or 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; ~623MB 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 Code’s content (Syntactic and Semantic Mapping), and 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 Codes), can support the discovery of federal product regulatory information. How do we map these codes to the relevant sections of the CFR? For example, the term "barrel" is represented as two different UNSPSC commodity codes: 1) 50012102 of the UNSPSC Family "Arms and crafts equipment and accessories and supplies", and 2) 61211513 of the UNSPSC Family "Specialized educational services".

In order to create sets of RDF statements/node match section for each product and industry label, we are exploring three different and successful 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/SPARCS (both of 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

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. Identify sections where a term is used.
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 definition of scope of the definition [2].

Example of definition extraction from CFR

Vocabulary Extraction: How can search be improved?

The Code of Federal regulations is divided in 96 files that represent broad areas subject to federal regulation agriculture, food and drug, 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 Power to obtain structured trees and dependencies from the CFR text. Stanford typed dependencies are used to exploit grammatical relations to construct the vocabulary. Hidden patterns support the identification of taxonemical relationships between the terms [2].

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

1. inclusion/union broader
2. inclusion/intersection
3. lexical relationship from the structure of the sentences in N-4N.

Example of obligation extraction

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 for an obligation and represent it in an RDF format [1]. Given CFR text that contains an obligation of the form "It is obligatory to Y with regards to Z", our system attempts to:

1. identify the text as an obligation
2. identify the address of the obligation
3. identify the action of the obligation
4. identify the subject of the obligation

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 modal shall/tell and other legal 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 addresses, action and object of the obligation.

References


Acknowledgements

We would like to acknowledge the collaboration and work of Thomas R. Bruce (Director), Sara Frang, Daniel Nagy, David Shestdren, and Wayne West at the Legal Information Institute, and Professor Claire C. Card of the Department of Computer Sciences.