Towards a Linked Open Data Cloud of Language Resources in the Legal Domain

Patricia Martín Chozas[1], Elena Montiel Ponsoda, Víctor Rodríguez-Doncel
Ontology Engineering Group, Universidad Politécnica de Madrid, Spain

Abstract. This paper describes the process of identifying and transforming terminological resources in the legal domain into RDF. A survey of heterogenous legal resources in different languages and applicable to different jurisdictions is made, and a selected group of terminologies is transformed into RDF and published as linked data.

Keywords: language resources, semantic web, linked data, legal knowledge graph

1. Introduction

Most practitioners of the legal profession are pleased to ornate their office with books, and legal dictionaries are never missing in their collections. Legal dictionaries used to be valuable resources in their daily job; but nowadays computers have revoked their usefulness. These computers, nevertheless, still need from language data to work properly, and a new breed of electronic language resources has taken over. In this context, *language resources* are defined as pieces of structured data in a machine-readable form, comprising corpora, terminologies, thesauri, knowledge bases, lexicons and dictionaries. These resources are necessary to train machine translation tools, to automate software localisation systems or to test the natural language processing algorithms of a speech recognition system, to mention but a few.

Examples of language resources in the legal domain are Jurivoc[2], a juridical thesaurus for Swiss regulations; the UNESCO thesaurus[3], which contains terms from various fields including the legal domain; or the STW thesaurus[4], covering the economy domain. These resources were intended for human consumption, but they have been repurposed to be consumed by machines.

Entries in these databases are naturally connected through hyperlinks within the same resource (a dictionary referring to other entries in the same dictionary), across similar resources (you can jump online from an entry in the Random House dictionary to the equivalent in the Merriam Webster) or even across resources of different nature (a corpus of texts with some of the terms linked to entries in another term database). The value of the resources is much higher when connected. We, humans, like hypertext documents in the Web, which enable us to naturally hop from document to document in a form that is connatural with the way we think. Machines perceive many more advantages, and when data is connected to other pieces of data, it is no longer considered as data but as knowledge. *Knowledge graphs* are no other thing that a set of connected pieces of information. A knowledge graph is a structure to represent information, where *entities* are represented as nodes, their *attributes* node labels and the relationship between entities are represented as *edges*.

This paper is the first effort towards the construction of a knowledge graph of language resources in the legal domain. The overall objective of the work where this paper is framed is the construction of a Legal Knowledge Graph (LKG) enabling the provision of compliance-

[1] This work has been funded by the H2020 Lynx project, by a UPM grant, and by a Juan de la Cierva grant. Lynx has received funding from the Horizon 2020 European Union (EU) Research and Innovation programme under Grant Agreement: 780602.
[2] https://www.bger.ch/ext/jurivoc/live/de/jurivoc/Jurivoc.jsp?interfaceLanguage=german
[3] http://skos.um.es/unescothes/?l=en
[4] http://zbw.eu/stw/version/latest/about

related services. This is the main goal of the H2020 Lynx[5] project, and this contribution is a part of it, specifically focusing on language resources. The rest of the abstract is organised as follows: Section 2 describes the goal of this work, a knowledge graph of language resources in the legal domain, together with a first account of identified assets. Section 3 describes the process of transforming existing language resources and adding them to the graph, together with the future work.

2. Linked Open Data Cloud of legal language resources

Many resources in the legal domain can already claim to be connected, like any HTML or XML documents with hyper-references to other documents. However, a good way to describe connected resources on the Web relies on the W3C specifications of the Semantic Web, such as RDF[6], RDFS[7], OWL[8] and SKOS[9]. Linked Data [1] is a particularly sound manner of publishing RDF. Linked data is data published according to the Linked Data Principles [2]: entities should be identified via unique URIs; the URIs should be HTTP URIs, follow standard web protocols, return useful information about the resource and contain links to other related resources. Publishing data as linked data improves the interoperability of data and enables a new breed of tools for data analysis, comparative law studies of systems, regulation checks, etc.

Datasets published as linked data are part of the *Linked Open Data (LOD) cloud[10]*, a diagram representing connected linked data resources. The *Linguistic Linked Open Data* (LLOD) *cloud*[11] [3] is a subset of the former, restricted to datasets in the linguistic domain. The first objective of the work presented here is the identification of existing linked open data language resources in the legal domain. This *Linguistic Legal Linked Open Data* (LLLOD) cloud shaped here would be the inner core of a broader Legal Knowledge Graph, where other non-RDF documents are also referenced.

Despite the existence of language resource portals, there is no good catalogue focused on language legal resources and identifying the relevant resources in the domain is already a first contribution of this work. In order to identify relevant resources, three different paths were explored: (a) general web search; (b) lookup of resources described in papers from the specialized literature and (c) search in data portals specialized in language resources. The latter includes ELRC-SHARE repository (used for documenting language resources by the European Language Resource Coordination), ReTeLe Catalogue (for language resources in Spain), CLARIN (European research infrastructure for language resources), the OLAC Language Resource Catalogue[12] (unified portal for language resource search) and the ELRA Catalogue (European Language Resources).

Each of the resources of interest was described in terms of (a) a general description; (b) whether the dataset is RDF or not and if it is available as linked data and (c) which other resources were connected to this one. Table 1 shows the initial compilation of resources, whereas Figure 1 illustrates the interconnections between some of them.

[5] http://www.lynx-project.eu/
[6] https://www.w3.org/RDF/
[7] https://www.w3.org/TR/rdf-schema/
[8] https://www.w3.org/TR/owl2-syntax/
[9] https://www.w3.org/TR/skos-reference/
[10] http://lod-cloud.net/
[11] http://linguistic-lod.org/llod-cloud
[12] https://www.elrc-share.eu/, http://catalogo.retele.linkeddata.es/, https://www.clarin.eu/, http://www.language-archives.org/, http://catalog.elra.info/en-us/, respectively.

ID	Name	Description	Language
iate	IATE	EU terminological database.	EU languages
eurovoc	Eurovoc	EU multilingual thesaurus.	EU languages
eur-lex	EUR-Lex	EU legal corpora portal.	EU languages
conneticut-legal-glossary	Connectict Legal Glossary	Bilingual legal glossary.	en, es
unesco-thesaurus	UNESCO Thesaurus	Multilingual multidisciplinary thesaurus.	en, es, fr, ru
library-of-congress	Library of Congress	Legal corpora portal.	en
imf	International Monetary Fund	Economic multilingual terminology.	en, de, es
eugo-glossary	EUGO Glossary	Business monolingual dictionary.	es
cdisc-glossary	CDISC Glossary	Clinical monolingual glossary.	en
stw	STW Thesaurus for Economics	Economic monolingual thesaurus.	en
edp	European Data Portal	EU datasets.	EU languages
inspire	INSPIRE Glossary (EU)	General terms and definitions in English.	en
saij	SAIJ Thesaurus	Controlled list of legal terms.	es
calathe	CaLaThe	Cadastral vocabulary.	en
Gemet	GEMET	General multilingual thesauri.	en, de, es, it
informea	InforMEA Glossary (UNESCO)	Monolingual glossary on environmental law.	en
copyright-termbank	Copyright Termbank	Multi-lingual termbank of copyright-related terms.	en, es, fr, pt
gllt	German labour law thesaurus	Thesaurus with labour law terms.	de
jurivoc	Jurivoc	Juridical terms from Switzerland.	de, it, fr
termcat	Termcat	Terms from several fields including law.	ca, en, es, de, fr, it
termcoord	Termcoord	Glossaries from EU institutions and bodies.	EU languages
agrovoc	Agrovoc	Controlled general vocabulary.	29 languages

Table 1. Some features of relevant language resources for the legal domain.

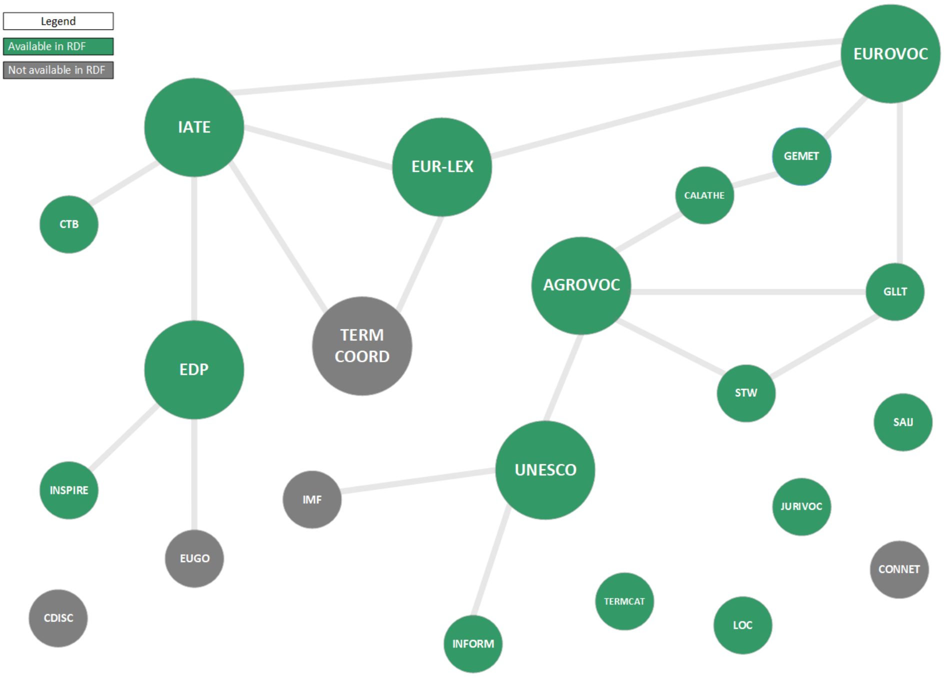

Figure 1. Relations between identified resources sorted by format.

Each of these datasets has been described with the DCAT vocabulary and published in the CKAN-based open data portal of the Lynx project[13], where they can be browsed using facets

[13] http://data.lynx-project.eu/

(language, domain format, jurisdiction, etc.). Further description of the language resources and the research material of this paper is online[14].

3. Population of the LOD and future work

Some relevant datasets with language data useful for the legal domain are already RDF and some of them are not (green and grey bubbles in Figure 1). In order to connect disparate RDF resources, and in order to transform non-RDF resources to RDF, the existence of common data models is of great help. Some of the most important data models in this domain are SKOS, Ontolex and NIF. The first represents concepts and ontological relation as is common in thesauri in a structured form. The latter second was conceived to model linguistic information relative to ontologies and expose lexical resources in the Semantic Web. In addition, since Lynx is developed in a multilingual environment, the vartrans module of Ontolex is relevant. NIF is of interest to represent annotations in annotated corpora etc.

In order to transform non-RDF resources to SKOS, Ontolex and NIF, the OpenRefine[15] tool has been chosen. The SKOS and Ontolex properties have been chosen according to the information that needs to be represented (lexical and terminological information) and to the structure of the RDF resources gathered in order to ease the linking stage. Accordingly, some of the most relevant properties spotted are contained in Table 2.

SKOS	Ontolex
skos:prefLabel	ontolex:lexicalEntry
skos:altLabel	ontolex:canonicalForm
skos:definition	ontolex:language
skos:note	ontolex:writtenRep
skos:broader	ontolex:sense
skos:topConcept	

Table 2. SKOS and Ontolex properties considered (non-equivalent).

The discovery and growth of the LLLOD cloud is an endeavour hardly started. There are many steps to be taken: the work described here is just a first approach to a higher scope. More datasets need to be identified with their subsequent transformation to RDF. On the other hand, specific language resources for Lynx project are to be generated. Finally, the linking of all the resources needs to be performed, contributing in this manner to the enrichment of the Semantic Web.

References

[1] T. Berners-Lee, “Design issues: Linked data,” 27 07 2006. [Online]. Available: https://www.w3.org/DesignIssues/LinkedData.html. [Accessed 13 05 2018].

[2] C. Bizer, T. Health and T. Berners-Lee, “Linked Data - The Story So Far,” *International Journal on Semantic Web and Information Systems,* 2011.

[3] C. Chiarcos, J. McCrae, P. Cimiano and C. Fellbaum, “Towards Open Data for Linguistics: Linguistic Linked Data,” *New Trends of Research in Ontologies and Lexical Resources,* 2013.

[14] http://lynx-project.eu/data/language-resources

[15] http://openrefine.org/