Spanish Legislation as Linked Data

Víctor RODRÍGUEZ-DONCEL [a], María NAVAS-LORO [a], Elena MONTIEL-PONSODA [a] and Pompeu CASANOVAS [b]
[a] *Universidad Politécnica de Madrid*
[b] *School of Law, La Trobe University*

Abstract. Legislation is officially published in Spain as HTML, PDF and XML. In the next few months, metadata will also be published as RDF, following the guidelines of the European Legislation Identifier (ELI) and using metadata records supported by the ELI ontology. The work presented here is an independent effort to publish Spanish consolidated legislation strongly linked to other external resources. In the published dataset, text is structured in articles; key terms are related to external terminological databases, named entities are identified, and links between internal and external documents have been automatically identified. The dataset is publicly available in a SPARQL endpoint.

Keywords. official journal, government gazette, publication of law, RDF, BOE

1. Introduction

Legislation is published in Spain by the Official State Gazette (*Boletín Oficial del Estado*, BOE), an official journal that makes available laws, regulations and other acts and documents approved by the Spain's Parliament and the Autonomous Communities.
The 'Technical Specification for the implementation of the ELI in Spain'[1], adopted in March 2018, lays down the guidelines for the implementation of the ELI in Spain, the European Legislation Identifier, which is harmonizing the way legislation is published in Europe. Every piece of legislation will be identified by an HTTP URI, and homogeneously described with a common minimum set of metadata elements supported by the ELI Ontology[1]. The structure of the URI, known as URI template (RFC6570) has been well determined[2] for each document, and precise instructions have been given on how to mint the actual URIs for the Spanish case. Thus, the specification says that ISO 3166 codes will be used for the 'jurisdiction' field, and a set of acronyms have been defined for the different types of legislation that are published by BOE in Spain. Also, 23 properties of the ELI Ontology have been chosen for the description of the documents, such as `eli:jurisdiction` or `eli:title`. The ELI Ontology does not provide the means for structuring the content, instead, other formats, such as the OASIS Akoma Ntoso enable the representation of executive, legislative and judiciary documents in a structured manner as XML. The benefits brought by this initiative are multiple. Citizens and companies will have a better access to legislation by finding homogeneous practices

[1] http://publications.europa.eu/mdr/resource/eli/eli.owl

[2] /eli/{jurisdiction}/{type}/{year}/{month}/{day}/{number}/{version}/{version_date}/{language}/{format}

throughout all Europe; companies will be able to run legal information systems more smoothly, and documents will be unequivocally identified and better described.

However, the benefits for computer programs would be larger if documents were massively annotated and linked to other documents, creating a connected *legal knowledge graph* (LKG). This LKG will enable new applications and will enhance existing legal information systems.

The work presented in this paper has been made in the framework of the H2020 Lynx project (*Building the Legal Knowledge Graph for Smart Compliance Services in Multilingual Europe*), which aims at building a LKG for providing compliance services. This paper describes the transformation of Spanish legislation into RDF and the linkage of the legal resources with entities present in external databases. The dataset is available for download as a raw data[3] file and it is also accessible in a SPARQL endpoint[4].

2. Methodology

2.1. Data harvesting and transformation

Legislation is published in Spain in the *Boletín Oficial del Estado* (BOE) website[5] in different formats, usually HTML, PDF and XML, and both in a gazette and in a consolidated version. Documents in the BOE are not available for bulk download, but an iterative access via scripts[6] was made to retrieve the XML version of all documents pertaining to a collection of specific categories of norms (*Ley, Ley Orgánica, Decreto Ley* and *Real Decreto Ley* among others) between 1978-2018, and being limited to legislation in force. Some of these documents are available in the different co-official languages of Spain, but the search was limited to the Spanish version.

A second batch of scripts transformed the downloaded XML documents into its RDF form, following standard practices inasmuch as possible. Thus, the data model was based on the ELI Ontology, complemented by properties from other *de iure* or *de facto* standard ontologies (FOAF, DublinCore, EU Common Data Model[7]). Whenever no well acknowledged entity type or property existed, new ontology classes and properties were defined within a LKG ontology. The ELI Ontology provides most of the elements for a good metadata description of the legislative documents. The FRBR model in which it lies is suitable to represent multiple expressions and languages of the same legal resource. Texts can be at least minimally structured with ELI elements (`eli:LegalResourceSubdivision`, `eli:has_part`), although other models such as the EU official terms for subdivisions in legal documents[8] or those in the Metalex ontology could be more flexible. Information about the status of the Act can also be given with the elements of the ELI Ontology[9]– whether it is in force, partially in force, or deprecated. Some figures related to the extracted triples are shown in Table 1.

[3] `http://lkg.linkeddata.es` and `http://lynx-project.eu/data2/spain`

[4] `http://sparql.lynx-project.eu`

[5] `http://www.boe.es`

[6] `https://gitlab.com/superlynx/upm`

[7] `https://publications.europa.eu/en/web/eu-vocabularies/cdm`

[8] `http://publications.europa.eu/mdr/resource/authority/subdivision`

[9] `http://data.europa.eu/eli/ontology`

Number of documents:	3617
Number of triples:	162786
Size of the bulk file:	200Mb

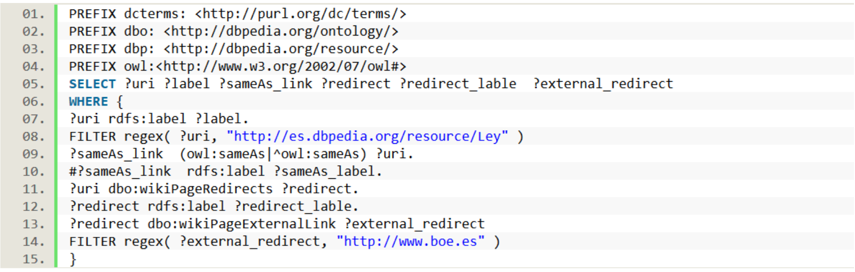
Table 1. Some relevant figures of the dataset

The next stage, described in the following sections, consists of the data linking, namely, the identification of resources in external RDF datasets to connect to. Whereas automated entity extraction and linking sometimes can be done in real time, there are advantages for the execution of this task in a previous stage: information is available faster at runtime, more complex queries can be built linking different databases, curation by other parties is possible, indexing of extracted information improves the search, or the use of the dataset in a simpler manner by non NLP-experts.

2.2. Named Entity Recognition and Linking

The entities that have been recognized, disambiguated and linked include:

Nicknames of laws, Wikipedia, Wikidata — A number of queries were made (available online, see a sample in Figure 1) with the goal of linking some popular laws to their Wikipedia and Wikidata pages. This effort resulted in 73 links from legal Acts to the Wikipedia in Spanish, plus a dozen of links to Wikidata. Whereas the relation is not one to one (one Wikipedia article may describe several Acts and vice versa), the connection of a legal resource to a popular description such as Wikipedia/DBpedia enables queries that were not possible before. Derived from this effort, a number of additional nicknames, or popular terms with which some laws are known, have been added. For example, much as the "Bolkenstein Directive" refers to the "Directive 123/2006/CE", the "Ley de Economía Sostenible" is informally referred in Spain as "Ley Sinde". These nicknames were obtained from Wikipedia redirections and coreferences manually revised.



```
PREFIX dcterms: <http://purl.org/dc/terms/>
PREFIX dbo: <http://dbpedia.org/ontology/>
PREFIX dbp: <http://dbpedia.org/resource/>
PREFIX owl:<http://www.w3.org/2002/07/owl#>
SELECT ?uri ?label ?sameAs_link ?redirect ?redirect_lable  ?external_redirect
WHERE {
?uri rdfs:label ?label.
FILTER regex( ?uri, "http://es.dbpedia.org/resource/Ley" )
?sameAs_link  (owl:sameAs|^owl:sameAs) ?uri.
#?sameAs_link  rdfs:label ?sameAs_label.
?uri dbo:wikiPageRedirects ?redirect.
?redirect rdfs:label ?redirect_lable.
?redirect dbo:wikiPageExternalLink ?external_redirect
FILTER regex( ?external_redirect, "http://www.boe.es" )
}
```

Figure 1. Sample SPARQL query (courtesy of I. Badji)

Corporate bodies — The Publications Office of the EU maintains a Metadata Registry with metadata elements, named authority lists (NALs), schemas, etc. used by different European Institutions. Information in this registry, such as a list of corporate bodies[10], is also offered as SKOS RDF. Spanish public institutions and administrative units are also normalized (*Directorio Común de Unidades Orgánicas y Oficinas)*, but no RDF is provided. The existence of mentions to these bodies and departments has been detected

[10] http://publications.europa.eu/resource/authority/corporate-body

for each document. A simple regular expression search was made to find these links; disambiguation was not needed due to the capitalization and long extension of the elements.

Legal References, Persons, Localizations and others[11] — A general Named Entity Recognition algorithm has been implemented by Badji [12] following a rule-based approach and using different NLP frameworks. The collection of rules is specific to the Spanish jurisdiction (e.g. `Ley xx/yyyy`) and is published along with the dataset. Among the specific entities able to be detected by the system [12] Judgments, Articles, Constitutions, Organic Laws and other institutions have been considered. For this system, some resources were specifically created to identify different ways to refer to Spanish laws. Next to this effort, the alignment of BOE keywords to Eurovoc terms has been carried out using the Lemon models in SKOS/RDF

3. Related work

The publication of legislation online in a structured form is not new. Whereas a first wave of publishers adopted different flavors of XML to publish legislation [18][24][25] and [11], including the Spain jurisdiction [13], Linked Data is gaining adoption in the last few years. Some examples of XML-based systems are NormeInRete in Italy, LexDania in Dennmark [26], Akoma Ntoso in Brazil, CHLexML in Switzerland or LexML in Austria.

Many official law publishers all over the world still do not publish legislation in a structured form. For example, the Australian Commonwealth legislation is published by the Federal Register of Legislation (formerly ComLaw), and it is available in PDF and Microsoft Word formats only.

The Spanish Legislation as Linked Data is not the first effort of the sort. For example, The MetaLex Document Server offers legal documents as versioned Linked Data [19], including Dutch national regulations. Other non-official initiatives have also offered Finnish [16] and Greek [15] legislation as Linked Data.

At the European Level, the Publications Office of the EU maintains the CELLAR repository for storing official publications and bibliographic resources produced by the different institutions of the EU [17]. EU member states are striving to publish national legislation in an equivalent manner, supported on the ELI initiative. ELI is based on three ideas (a) using a well defined identifier; (b) defining metadata records and (c) organising them with an ontology. As of today, the implementation is uneven, but some states have finished the change. Legal resources are already identified in the same manner in France, United Kingdom, Italy, Luxembourg, Ireland or Denmark, just to name some.

Not surprisingly, the information at CELLAR and in some of these countries is also available through SPARQL endpoints. However, these systems have the ambition to mirror the legislator's texts and they cannot offer enhanced versions with added contents or links –what falls in the realm of private companies.

Linking legal references and the named entities they contain has become a major challenge in the last years. Although the identification of legal cross-references has been extensively tackled in previous literature [20], not every language has received the same

[11] The integration of this module is still pending to be implemented.

attention. Most works focus on English texts [3][4]. However, some efforts have also been done for other languages, such as Dutch [5], Italian [6], French [7], Japanese [9], and Spanish [14]. The approach and the kind and depth of the Named Entity classification differ from one work to another. While most of them use a pattern-based approach [14][5][7], machine-learning has been previously applied as well [9]. Regarding the Named Entities targeted, most systems focus on rules and references to other documents, but extraction of entities such as judges or jurisdictions has also been tackled. Finally, some systems for document linking have been proposed too [2][10].

Regarding the connection of legal resources to language resources, most of the documents published by gazettes are described by keywords (as in the Spanish case at the BOE), possibly linked to controlled vocabularies or terminological databases[12] [21]. A proper identification and definition of the terms used in a legal document is essential to a) properly understand the document, b) properly identify the topic (and subtopics) addressed in the document, c) classify it, and d) establish equivalences between terms in other jurisdictions (in the same or different languages). Besides the different types of matches present in terminological databases [22], legal relations are to be considered as well in this domain. Legal taxonomies, like the one by Ajani et al. [23], have considered these aspects but have not been fully embraced by public authorities –with some exceptions such as Jurivoc[13] in Switzerland.

4. Conclusion

This paper has introduced a novel resource: the Spanish legislation published as RDF. This new resource is the result of an independent effort, and no sanction of any official body was intended. Moreover, the importance of this dataset is limited, as the dataset captures a collection of Acts in force as of 2018 but does not include a mechanism for automatically updating the contents. Yet, this is an important step towards building a true Legal Knowledge Graph, which is not an aggregation of legal resources but a strongly connected resource, available to all at no cost, and enabling new applications. Our effort has led to some additional results, such as the table of laws and nicknames by which they are commonly referred in informal contexts or by the mass media. These results can shed a new light about how the network effect of knowledge graphs can leverage the existing resources through innovative applications.

Acknowledgements

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 780602. Authors also thank Ines Badji, whose implementation of the NER system has been key for this work.

[12] The EU Publications Office maintains a collection of language resources at https://publications.europa.eu/en/web/eu-vocabularies

[13] https://bartoc.org/en/node/345

References

[1] Grupo de Trabajo Identificador Europeo de Legislación. (2018). Proyecto ELI: European Legislation Identifier. Especificación Técnica para la Implementación del Identificador Europeo de Legislación en España. Ministerio de Hacienda y Función Pública, *online*.

[2] Palmirani, M. and Benigni, F. (2007). Norma-system: A legal information system for managing time. In Proceedings of the V legislative XML workshop, pages 205-223.

[3] Bruckschen, M., Northeet, C., Silva, D., Bridi, P., Granada, R., Vieira, R., Rao, P., and Sander, T. (2010). Named entity recognition in the legal domain for ontology population. In Proceedings of the 3rd Workshop on Semantic Processing of Legal Texts (SPLeT 2010).

[4] Dozier, C., Kondadadi, R., Light, M., Vachher, A., Veeramachaneni, S., and Wudali, R. (2010). Named Entity Recognition and Resolution in Legal Text, pages 27-43. Springer Berlin Heidelberg, Berlin, Heidelberg.

[5] de Maat, E., Winkels, R., and van Engers, T. (2006). Automated detection of reference structures in law. Frontiers in Artificial Intelligence and Applications, 152:41-50.

[6] Palmirani, M., Brighi, R., and Massini, M. (2004). Processing normative references on the basis of natural language questions. In Proceedings 15th International Workshop on Database and Expert Systems Applications, 2004., pages 9-12.

[7] Adedjouma, M., Sabetzadeh, M., and Briand, L. C. (2014). Automated detection and resolution of legal cross references: Approach and a study of luxembourg's legislation. In 2014 IEEE 22nd International Requirements Engineering Conference (RE), pages 63-72.

[8] Sannier, N., Adedjouma, M., Sabetzadeh, M., and Briand, L. (2017). An automated framework for detection and resolution of cross references in legal texts. Requirements Engineering, 22(2):215-237.

[9] Tran, O. T., Ngo, B. X., Nguyen, M. L., and Shimazu, A. (2014). Automated reference resolution in legal texts. Artificial Intelligence and Law, 22(1):29-60.

[10] Opijnen, M., Verwer, N., and Meijer, J. (2015). Beyond the experiment: the extendable legal link extractor. In Workshop on Automated Detection, Extraction and Analysis of Semantic Information in Legal Texts, held in conjunction with the 2015 International Conference on Artificial Intelligence and Law (ICAIL), At San Diego, CA, USA.

[11] Iniesta Delgado, J. J.; Martínez González, M. M.; Vicente Blanco, D. J., eds. (2012). «Representación, organización y gestión del conocimiento en el ámbito jurídico». Scire: representación y organización del conocimiento 18 (1 [Número monográfico]). ISSN 1135-3716.

[12] Badji, I. (2018). Legal entity extraction with ner systems. MSc Thesis, Universidad Politécnica de Madrid.

[13] Alvite Díez, M. L. (2009). Las bases de datos jurídicas y el uso del lenguaje XML en España. Scire, 15(1):33-57

[14] Martínez-González, M., de la Fuente, P., and Vicente, D.-J. (2005). Reference extraction and resolution for legal texts. In International Conference on Pattern Recognition and Machine Intelligence, pages 218{221. Springer.

[15] Chalkidis, I., Nikolaou, C., Soursos, P., & Koubarakis, M. (2017). Modeling and querying greek legislation using semantic web technologies. In European Semantic Web Conference (pp. 591-606). Springer, Cham.

[16] Frosterus, M., Tuominen, J., Wahlroos, M., & Hyvönen, E. (2013). The Finnish law as a linked data service. In Extended Semantic Web Conference (pp. 289-290). Springer, Berlin, Heidelberg.

[17] Francesconi, E., Küster, M. W., Gratz, P., & Thelen, S. (2015). The ontology-based approach of the publications office of the EU for document accessibility and open data services. In International Conference on Electronic Government and the Information Systems Perspective (pp. 29-39). Springer, Cham.

[18] General XML format(s) for legal Sources, Estrella Deliverable 3.1, Caterina Lupo et al. (2007)

[19] Hoekstra, R. (2011). The MetaLex document server. In International Semantic Web Conference (pp. 128-143). Springer, Berlin, Heidelberg.

[20] Palmirani, M., Brighi, R., and Massini, M. (2003). Automated extraction of normative references in legal texts. In Proceedings of the 9th international conference on Artificial intelligence and law, pages 105-106. ACM.

[21] Alvite Díez, M. L. (2012). El uso de vocabularios controlados en los sistemas de información jurídica. Evolución y tendencias actuales de representación. Scire: representación y organización del conocimiento, 18(1), 29-39.

[22] Cabré, M. T. (2008). El principio de poliedricidad: la articulación de lo discursivo, lo cognitivo y lo lingüístico en Terminología (I). IBÉRICA 16:9-36.

[23] Ajani, G., Boella, G., Lesmo, L., Martin, M., Mazzei, A., Radicioni, D. P., & Rossi, P. (2009). Legal taxonomy syllabus version 2.0. IDT, 9.

[24] Martínez González, M. M.; Vicente Blanco, D. J., eds. (2009). XML legislativo: representación y organización de la información jurídica a través de la tecnología XML. SCIRE. Representación y Organización del Conocimiento 15 (1 [Número monográfico]). ISSN 1135-3716.

[25] Martínez González, M. M., ed. (2015). Derecho y Sistemas de Datos. El uso del XML jurídico. Valencia, España: Tirant Lo Blanch. ISBN 978-84-9086-010-6

[26] Petersen, K. E. (2011). Experiences with "Lex Dania Live". From Information to Knowledge: Online Access to Legal Information: Methodologies, Trends and Perspectives, 236, 69.