Lynx and the Legal Knowledge Graph: Integrating lexical and terminological resources with legal data

Elena Montiel-Ponsoda, Víctor Rodríguez-Doncel, Patricia Martín-Chozas, Ilan Kernerman

1. Introduction

December 2017 saw the kick-off of the European H2020 Innovation Action entitled *Lynx: Building the Legal Knowledge Graph for Smart Compliance Services in Multilingual Europe* (http://www.lynx-project.eu/). This project will run for three years and one of its main objectives is to create a unique and novel knowledge graph related to compliance, integrating information from heterogeneous data and content sources in various languages in what has been termed the Legal Knowledge Graph (LKG). Based on this knowledge base, the Lynx project will be demonstrated in three pilot studies where compliance services would be of much interest: labour law, data protection, and compliance with standards in the energy and oil & gas industries.

By *compliance* we mean "the conformance to a set of laws, regulations, policies, or best practices" (Silveira et al., 2010), that are deeply rooted in each country's traditions and which are mostly expressed in its own language(s). The main idea is thus that by connecting or interlinking legal data coming from different jurisdictions and institutions at various levels (locally, regionally, nationally, internationally), companies and public bodies involved in internationalization processes can have a more efficient access to legal information. This is especially relevant for companies in Europe due to the linguistic variety and different jurisdictions in the common shared market.

In order to bring such data together, Lynx will rely on public open data, on the one hand, and on the formalisms and technologies provided by the Linked Data paradigm, on the other. The latter enables to publish data of varying nature in standardised formats that permit to establish fine-grained relations between single data elements in a machine processable format. In this sense, a term mentioned in a European regulation can be related (i) to a lexical or terminological resource in which its meaning is specified, (ii) to a multilingual terminological database to obtain translations of legal terms, and (iii) to other documents that refer to the transposition of European law in several national legislations.

For this purpose, a set of domain-neutral common services will be developed to perform tasks such as the linking itself, but also term extraction, conversion of resources to the linked data formats, or document annotation. These services will facilitate the development of a set of business-oriented applications to cover the needs of the use cases involved in the project, including smart search, recommendation and alert systems, document translation and document summarization. All in all, Lynx will provide a single entry point to interlinked legal information across jurisdictions and languages, which will enable users a better access to legal and regulatory data.

The typology of data being interconnected in the LKG will be of a different nature. In this contribution, we offer an overview of the data sources considered relevant for the purposes of Lynx (Section 2), focussing on linguistic resources. Then, we analyse the role that lexical and terminological data, specifically, will play in the LKG (Section 3). Finally, we conclude with a brief overview of the partners involved in the Lynx project.

2. Data sources for the LKG

The European Directive on Public Sector Information (PSI), in force since 2003 and amended in 2013, was followed by national developments in the Member States, which unleashed huge amounts of high quality data in government data portals. Central to the domain of compliance are the European Open Data Portal[1] and the Eur-Lex portal[2], maintained by the EU Publications Office, and hub of European Union law and other official public documents. Apart from the access to European legislation through the Eur-Lex portal, the European Open Data Portal provides access to N-Lex[3], a portal for the national legislation, to EuroVoc[4], the EU multilingual thesaurus, or to the Translation Memory of the DGT (Directorate General for Translation)[5], to mention some of them.

The Commission Decision of 12 December 2011 on the reuse of Commission documents mandated that all documents and data should be available for reuse without charge "in machine-readable format where possible and appropriate" and "through a data portal as a single point of access to its structured data so as to facilitate linking and reuse". Whereas the Eur-Lex portal has indisputably succeeded to fulfil these mandates, the machine-readability and the linking aspects can still be taken further to broader contexts, and this is exactly where the Lynx project would like to advance the state of the art.

The first steps in this regard have consisted of identifying the existing resources in machine-processable formats that could be directly integrated into the LKG. Here we can distinguish between datasets in the regulatory domain, and datasets in the linguistic domain. As for the regulatory domain, the following datasets are of interest for the project:

- Eur-Lex: Database of legal information containing EU law (treaties, directives, regulations, decisions, consolidated legislation, etc.), preparatory acts (legislative proposals, reports, green and white papers, etc.), case-law (judgments, orders, etc.), international agreements, etc. Eur-Lex is a huge database updated daily with some texts dating back to 1951.
- Openlaws: Austrian laws (federal and of the 9 regions)

1 https://www.europeandataportal.eu/en/homepage
2 https://eur-lex.europa.eu/
3 http://eur-lex.europa.eu/n-lex/
4 http://eurovoc.europa.eu/
5 https://ec.europa.eu/jrc/en/language-technologies/dgt-translation-memory

and rulings (from 10 different courts), German federal laws, European laws (regulations, directives) and rulings (general court, European Court of Justice). It includes Eur-Lex, 11k national acts and 300k national cases in a neo4j graph.
- DNV-GL: Standards, regulations and guidelines to the public, usually in PDF.

Regarding open linguistic datasets, several sites and repositories have been surveyed. One of the sources of most interest for linguistic open data is the Linked Open Data cloud (LOD cloud)[6], due to its open nature and adequate format as linked data or RDF. In particular, the Linguistic Linked Open Data Cloud[7] is a subset of the LOD cloud which provides exclusively linguistic resources sorted by typology. Different types of datasets in the Linguistic Linked Open Data Cloud are: corpora; terminology, thesauri and knowledge bases; lexicons and dictionaries; linguistic resource metadata; linguistic data categories; and typological databases.

For the purposes of this project, the three first types of resources have been shortlisted as the most useful, and four resources have been identified as of particular interest for Lynx. They are:

- STW Thesaurus for Economics[8]: a thesaurus that provides a vocabulary on any economic subject. It also contains terms used in law, sociology and politics. (English)
- Copyright Termbank[9]: a multilingual term bank of copyright-related terms that has been published to connect WIPO definitions, IATE terms and definitions from Creative Commons licenses. (multilingual)
- EuroVoc[10]: a multilingual and multidisciplinary thesaurus covering the activities of the EU. It is not specifically legal, but it contains pertinent information about the EU politics and law. (multilingual)
- IATE[11]: a terminological database developed by the EU which is constantly updated by translators and terminologists. Amongst other domains, terms belong to the domains of law and EU governments. A transformation of a part of the public IATE was published as RDF in 2015. (multilingual)

Resources published in other formats have been considered as well. Structured formats include TBX (used for term bases), CSV and XLS. Exceptionally, resources published in non-machine-readable formats might be considered. Consequently, the following resources published by the EU have also been listed as usable, although they are not included in the Linguistic Linked Open Data Cloud:

- INSPIRE Glossary[12]: a term base developed by the INSPIRE Knowledge Base of the EU. Although this project is related with the field of spatial information, the glossary contains general terms and definitions that specify the common terminology used in the INSPIRE Directive and in the INSPIRE Implementing Regulations. (English)
- EUGO Glossary[13]: a term base addressed to companies and entrepreneurs that need to comply with administrative or professional requirements to perform a remunerated economic activity in Spain. This glossary is part of a European project and contains terms about regulations that are valuable for Lynx purposes. (Spanish)
- GEMET[14]: a general thesaurus, conceived to define a common general language to serve as the core of general terminology for the environment. This glossary is available in RDF and it shares terms and structures with EuroVoc. (multilingual)
- Termcoord[15]: a portal supported by the EU that contains glossaries developed by the different institutions. These glossaries cover several fields including law, international relations and government. Although the resources are available in PDF, at some point these documents could be treated and transformed into RDF if necessary. (multilingual)

In the same way, the United Nations also counts with consolidated terminological resources. Given their intergovernmental domain, the following resources have been selected:

- UNESCO Thesaurus[16]: a controlled list of terms intended for the subject analysis of texts and document retrieval. The thesaurus contains terms on several domains such as education, politics, culture and social sciences. It has been published as a SKOS thesaurus and can be accessed through a SPARQL endpoint. (multilingual)
- InforMEA Glossary[17]: a term bank developed by the UN and supported by the EU with the aim of gathering terms on Environmental Law and Agreements. It is available as RDF and it will be upgraded to a thesaurus during the following months. (multilingual)
- International Monetary Fund Glossary[18]: a terminology list containing terms on economics and public finances related with the EU. It is available as a PDF downloadable file; however, it may be transformed as part of future work. (multilingual)

In addition, other linguistic resources (not supported by the EU nor the UN) have been spotted. Some of them are already converted into RDF:

- Termcat (Terminologia Oberta)[19]: a set of terminological databases supported by the government of Catalonia. They contain term equivalents in several languages. Part of these databases were converted into RDF previously and are part of the TerminotecaRDF project, where they can be accessed through a SPARQL endpoint. (multilingual)

6 http://lod-cloud.net/clouds/lod-cloud.svg
7 http://linguistic-lod.org/
8 http://zbw.eu/stw/
9 https://termbank.com/es/espanol-ingles
10 http://eurovoc.europa.eu/
11 http://iate.europa.eu/
12 http://inspire.ec.europa.eu/glossary/
13 https://www.eugo.es/
14 https://www.eionet.europa.eu/gemet/en/
15 http://termcoord.eu/
16 http://vocabularies.unesco.org/browser/thesaurus/en/
17 https://www.informea.org/en/terms
18 https://www.imf.org/external/np/term/eng/index.htm
19 http://www.termcat.cat/es/TerminologiaOberta/

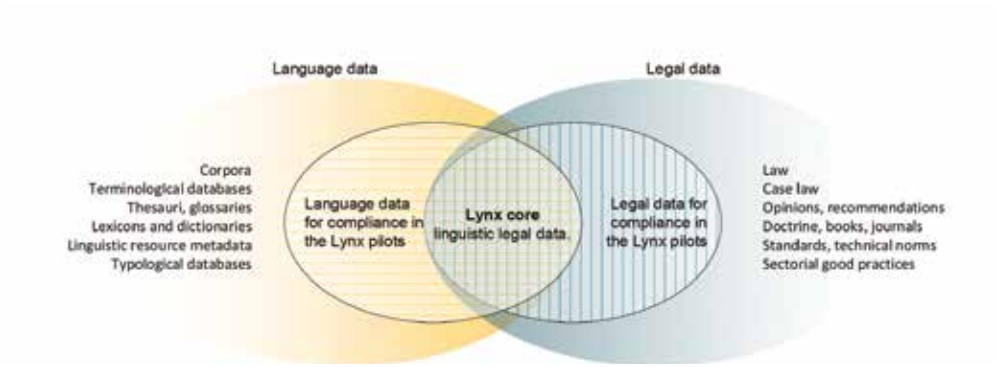

Figure 1. Scope of the multilingual Legal Knowledge Graph

- German Labour Law Thesaurus[20]: a thesaurus covering all main areas of labour law, such as the roles of employee and employer and legal aspects of labour contracts. It is available through a SPARQL endpoint and as RDF downloadable files. (German)
- Jurivoc[21]: a juridical thesaurus developed by the Federal Supreme Court of Switzerland in cooperation with Swiss legal libraries. It contains juridical terms arranged in a monohierarchic structure. (multilingual)
- SAIJ Thesaurus[22]: a thesaurus that organises legal knowledge through a list of controlled terms which represent concepts. It is available in RDF and is intended to ease users' access information related to the Argentinian legal system that can be found in a file or in a documentation centre. (Spanish)
- CaLaThe[23]: a thesaurus for the domain of cadastre and land administration that provides a controlled vocabulary. It is interesting because it shares structures and terms with AGROVOC and the GEMET thesaurus, and it can be downloaded as an RDF file. (English)
- CDISC Glossary[24]: a glossary containing definitions of terms and abbreviations that can be relevant for medical laws and agreements and is available in several formats, including OWL. (English)

Finally, we would like to refer to one multilingual terminology database that is not open, but that will be provided by Tilde, a member of the Lynx consortium.

- Tilde Terminology Database: A database of terminological data, including standardised terminology and statistical database of term candidates and translation equivalents in all EU official languages. It contains more than 4 million authoritative terms and over 17 million automatically extracted terms and translation equivalents in XML and RDF formats. (multilingual in 24 European languages)

Apart from these thesaurus and terminological databases that are domain-specific, some linguistic resources covering general language will be integrated as well. In this sense, we have focused on open resources, and the ones we consider as candidates for the LKG include:

- DBpedia[25]: A crowd-sourced community initiative to extract structured information from Wikipedia and make it available on the Web, to enable sophisticated queries and link different data sets, in the aim of facilitating the utilization of the huge amount of information in Wikipedia in novel ways and inspiring new mechanisms for navigating, linking, and improving Wikipedia itself. It contains 4.5M concepts published in RDF.
- JRC-Names[26]: A multilingual entity resource for person and organisation names consisting of large lists of names and their spelling variants (including across scripts), also available as linked data with additional information such as frequencies per language, titles found with the entities, and date ranges. It has been published both as RDF and CSV.
- BabelNet[27]: This is both a multilingual encyclopaedic dictionary, with lexicographic and encyclopaedic coverage of terms, and a semantic network which connects concepts and named entities in a large network of semantic relations, including 14M entries in RDF and other formats.

As in the case of the domain-specific resources, Lynx also counts on another partner, K Dictionaries, to provide general linguistic resources for the four languages of the project (English, German, Italian, Spanish) that could be linked in the LKG, including:

- Global Series: Lexicographic data sets including semantic and grammatical information in varying degrees of monolingual, bilingual and multilingual combinations. (XML and RDF)
- Password and MultiGloss: English semi-bilingual dictionaries with translation equivalents in the other three languages, along with semi-automat multilingual glossaries for these languages linked to the English (multilingual) network. (XML)
- Random House Webster's College Dictionary: A comprehensive (American) English monolingual dictionary. (XML)

3. The role of lexical and terminological data in the LKG

As described above, Lynx embraces the LOD paradigm and its integration of language and legal data for generating added value to be exploited by NLP services built on top of it. This added value is at the intersection of language and legal data, and is represented in Figure 1. Lynx will use language data to annotate documents of different sorts in the legal and regulatory domains (law, case law, opinions and recommendations, doctrines, books, journals, standards, technical norms, sectorial good practices) of relevance to the business cases

20 http://vocabulary.wolterskluwer.de/
21 https://www.bger.ch/ext/jurivoc/live/fr/jurivoc/
22 https://www.bartoc.org/es/
23 http://cadastralvocabulary.org/
24 https://www.cdisc.org/standards/semantics/glossary
25 https://wiki.dbpedia.org/
26 https://ec.europa.eu/jrc/en/language-technologies/jrc-names
27 https://babelnet.org/

involved in the project (labour law, data protection, and compliance with standards in the energy and oil & gas industries). The integration of these two realms could derive in the creation of what we dubbed the Lynx core linguistic legal data.

As for the language resources, we can distinguish between those that are not specific of the legal domain but document and describe language in general, that is general dictionaries and lexicons, and those that gather and describe the specific vocabulary of experts in the various domains of knowledge, that we name for the purposes of this work terminological resources (also known as terminologies, terminological databases, terminological glossaries, thesauri or specialized dictionaries).

Language resources which are not domain-specific are also of interest for processing legal documents as the information they contain (words, definitions, synonyms, hypernyms-hyponyms, linguistic categories, etc.) may serve in word sense disambiguation tasks when identifying relevant information in a text and support the classification or annotation of texts. Such resources can also serve more complex tasks in language processing, like those involved in machine translation and summarization.

As for resources that cover specific areas of specialized knowledge, they contain the definitions of terms used in their expert-to-expert communication. In the case of Lynx, we refer to *legal terminologies*. A classic terminological resource contains information related to the terms used in a certain domain, their definitions, a reference to the domain or sub-domain in which the term is used, the term and definition sources, variants used in the same domain, equivalents in other languages (translation equivalents), and sentences that exemplify the use of the term or usage notes. In the case of IATE, for instance, we see a rating that refers to the degree of reliability of the information contained in the terminology entry, which is of great value to translators who will be using these resources to look up for translation candidates.

The role of such terminological resources in the LKG can be of different nature: they can simply serve to provide definitions of the terms used in a legal document for readers to better understand, or they can be used in other services to classify texts, identify topics (key words that define a text), or annotate texts. As in the case of general language resources, they are highly useful in the translation of specialized documents and in summarization tasks. If terminologies are multilingual, they can also help in the identification of equivalent terms in documents in multiple languages, thus contributing to the establishment of links among documents referring to similar topics.

4. The partners

The Lynx consortium is composed of ten partners from seven countries, with complementary skills. It is led by the Ontology Engineering Group from Universidad Politécnica de Madrid, which coordinates the project, leads the Data Acquisition and Management work package and also contributes to the development of services, given its expertise in semantic technologies and data-driven language technologies. The other academic partner, the Autonomous University of Barcelona, represented by the Institute of Law and Technology, brings in its expertise in the application of new technologies to the legal domain, and leads the Industry requirements elicitation process as well as the dissemination and exploitation of the project results.

The German Research Centre for Artificial Intelligence (DFKI), as one of the leading institutions in Europe for advanced IT applications dealing with human language, leads the development of a set of curation tools, technologies and services to bridge between the Lynx core platform services and the use case specific pilots. Semantic Web Company is an Austrian SME offering ICT consulting services and solutions in the fields of semantic information and data management. It leads the development of the Lynx platform core services, bringing in proprietary software components and semantic tools. The specification of the technical architecture as well as the integration of all platform components is performed by Alpenite, an IT software consulting and system integration company with headquarters in Italy.

The Latvian SME Tilde brings in its expertise in multilingual natural language data processing technologies. It provides custom machine translation services and cloud terminology services, and is also involved in other technological, management, dissemination and exploitation tasks. The main provider of lexical data is K Dictionaries, an Israeli technology-driven-content creator of crosslingual lexical data.

Openlaws, another Austrian SME, operates in the legal tech domain and contributes core technology to the data acquisition and management tasks. In addition, it represents the use case on data protection, and leads the development and integration of the two other pilot studies. The use case on industry standards is led by DNV.GL, an international standards certifying company with headquarters in the Netherlands, Norway and Germany. It contributes to the requirements elicitation and specification for its pilot and to the creation of the regulatory graph within the LKG. Finally, the labour law pilot is led by Cuatrecasas, a prominent Spanish law firm with presence in over ten countries, which also contributes with functional specification requirements.

Bibliographical References

Rodríguez-Doncel, V., Martín-Chozas, P., and Montiel-Ponsoda, E. 2018. D2.1 Initial Data Management Plan (Version 1.0). Zenodo (10.5281/zenodo.1256834).

Silveira, P., Rodríguez, C., Casati, C., Daniel, F., D'Andrea, V., Worledge, C. and Taheri, Z. 2010. On the design of compliance governance dashboards for effective compliance and audit management. In *Service-Oriented Computing, ICSOC 2009*, pp.208-217. Berlin and Heidelberg: Springer.

http://lynx-project.eu

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 780602.