

Lynx: Building the Legal Knowledge Graph for Smart Compliance Services in Multilingual Europe

Elena Montiel-Ponsoda, Víctor Rodríguez-Doncel

Ontology Engineering Group, Universidad Politécnica de Madrid
{emontiel, vrodriguez}@fi.upm.es

Abstract

This position paper describes the motivation, objectives and expected results of the recently started European project Lynx (H2020-780602). Lynx aims to provide a set of smart services to assist companies in Europe with compliance needs. The novelty of the proposed compliance services is that they will be built on and exploit a graph of legal and regulatory data – from different jurisdictions and in different languages – duly integrated according to the Linked Data paradigm. The fact that the Legal Knowledge Graph will bring together data from different legal and regulatory traditions in several languages is one of the most challenging aspects that language technologies will help to overcome.

Keywords: compliance, Legal Knowledge Graph, Linked Data, multilingualism, language technologies

1. Motivation

SMEs (Small and Medium Enterprises) introducing a product or service in a new market face fundamental issues related to legal and regulatory aspects that affect different markets. Questions that might have to be solved before going international include (by means of example):

How does country X deal with this aspect regulated in my country by regulation R? Which consequences could that have for the product or service I would like to launch?

How is the technical question Q handled in country Y? How has it been understood by local courts?

Which standards should we implement before launching our products or services?

In order to help companies to deal with these and similar compliance issues, the Lynx solution will create a unique and novel knowledge base related to compliance integrating information from heterogeneous data and content sources in what we have termed the Legal Knowledge Graph (LKG). Since compliance has to do with “the conformance to a set of laws, regulations, policies, or best practices” (Silveira et al., 2010), deeply rooted in each country’s traditions and which are mostly expressed in its own language, the integration task is a challenging one. In this sense, we believe that the combination of Semantic Web approaches, specifically the Linked Data paradigm, and language technologies will help us perform the integration phase. As a result, a set of services will be built upon the LKG to demonstrate the feasibility of the approach and solve some of the current business needs represented by three pilots on the following topics: a) data protection, b) regulation in the oil and gas industry, and c) labour law, as will be explained in Section 5.

Lynx is a European research project funded as an H2020 Innovation Action covering the topic ICT-14: Big Data PPP: *cross-sectorial and cross-lingual data integration and experimentation*. The project began in December 2017 and will run for three years.

2. Main Objectives

The main objective of Lynx is to facilitate compliance of SMEs in internationalisation processes by leveraging existing European legal and regulatory open data. This will

be achieved by providing innovative legal business models by connecting national and international legislation, regulations, standards, case law, and best practices in the same or different languages. Building such a cross-language legal information system is aimed to offer a competitive advantage for European companies against others operating within a single jurisdiction. As such, these companies will manage to reduce the costs and efforts related to organising and monitoring legislation, regulations and sectorial good practices, and even incorporating new jurisdictions into their businesses.

As for the technical objectives, these can be classified into three: 1) to deliver domain-neutral common services to create and exploit the LKG (document annotation, interlinking, term extraction, smart search, translation, etc.), 2) to create a family of business-oriented applications to cover the particular needs of the user cases involved in the project, and 3) to provide a single entry point to interlinked legal information across jurisdictions and languages.

Beyond those business and technical objectives, Lynx also pursues societal objectives. Amongst others, providing European citizens better access to legal and regulatory information from multiple jurisdictions, and an easier involvement in legislative processes.

3. The Lynx Data Value Chain

The Lynx approach relies on a data value chain that consists of three main phases: data acquisition, data integration and data exploitation, as illustrated in Figure 1.

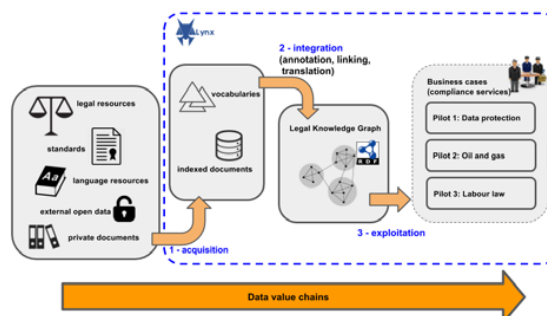


Figure 1: Lynx Data value chain

Step 1. Acquisition: Regulatory data in a broad sense (legislation, jurisprudence, standards, and norms) from multiple jurisdictions and in different languages will be aggregated in the LKG.

The EU PSI Directive, 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 Data Portal¹ and the Eur-Lex² portal, maintained by the EU Publications Office, and hub of European Union law and other official public documents.

Additionally, a number of public and private-led initiatives are publishing online many valuable data sets and resources, openly licensed, but not yet interlinked; missing, thus, opportunities for data analytics. Only dedicated efforts like openlaws³ are fully exploiting this potential. The LKG will comprehensively integrate these resources assuming the linked data paradigm, containing both structured and unstructured content, the latter being annotated with terminologies and links to other documents. Most legal documents at European (e.g., legislation in Eur-Lex) and national levels (e.g. in CENDOJ⁴, Spanish case law from the Supreme Court, National High Court and other organs) already have some unstructured text (e.g., the text of a directive) and structured metadata (e.g., document's authorship, date of creation).

Step 2. Integration: In the second step, the basic LKG management infrastructure will be implemented, as well as the set of common services which will include (i) linking and extraction services, for the interlinking and annotation of documents and data; (ii) language services, for the domain specific machine translation and summarisation of documents, and (iii) information retrieval and smart services, for the intelligent search, curation, and comparison of documents, as well as a service of customizable notifications and alerts.

Step 3: Exploitation: Three pilot B2B applications will be developed on top of the LKG and the common services to demonstrate the technical feasibility and business potential of the Lynx platform. In each of these pilots, customised solutions will be created to meet the specific needs of three business cases. These pilots are: (i) legal compliance solution for data protection, where data protection related documents are innovatively managed, analysed, and visualised across different jurisdictions; (ii) compliance assurance services in the Oil & Gas and Energy sectors, where the Lynx platform helps understanding regulatory regimes (norms and standards) related to operations; (iii) compliance solution for strategy design in labour law, where legal provision, case law, administrative resolutions, and expert literature are interlinked, analysed, and compared to define the strategy that is applicable or of interest for the case.

¹ <https://www.europeandataportal.eu/>

² <http://eur-lex.europa.eu/homepage.html>

4. Expected Outcomes

The outcomes of this project will be:

Business outcomes: The companies involved in the project will be able to support their customers more effectively in the compliance-related services they provide, considering faster, more complete information, and reducing risks. This novel approach will transform the way these companies operate. The feasibility of using open-data based services for compliance will be demonstrated and other law firms and compliance assurance companies will be encouraged to adopt the Lynx approach, thus multiplying the impact and reducing the language and legal barriers in the fragmented EU markets and beyond.

Societal outcomes: EU citizens will have access to information related to compliance for the first time integrating legislation and standards in a single multilingual portal. Citizens moving or operating across borders will enjoy more opportunities because legal uncertainty and risks will be reduced.

Technical and innovation outcomes: New legal linked data, publicly offered through APIs and highly connected to external datasets. A set of core services that can be reused by third parties external to the project and three running pilots covering in depth specific domains, with a user interface designed to address specific business cases. The new point to access open legal information, different in business models to the traditional legal information providers, and a proof that open data properly curated and connected suffices in many of the standard needs of lawyers in the context of compliance SMEs will benefit from Lynx in two different flavours: companies registered in the Lynx platform and companies in the portfolio of customers of law firms and consultancy agencies registered in Lynx.

5. The Lynx platform and Business Cases

The Lynx platform is an information system based on semantic and multilingual data technologies and aimed at assisting companies in internationalisation processes through a set of compliance-oriented digital services.

As can be seen in figure 2, the platform will consist of a knowledge graph with legal and regulatory data named Legal Knowledge Graph, (block 1, at the bottom of the figure), a collection of common services for compliance (block 2, in the middle) and the user interfaces for three specific-domain pilots (block 3, at the top). The Lynx platform will also be offered as a web portal accessible to the general public who can search, browse, and access open legal and regulatory documents.

A number of data source dependant converters will be used to ingest the documents and data in the Lynx platform. Such data will be very diverse in nature, scope and formats, and it will comprise: legal resources (legislation and case law), standards, language resources both domain-specific and domain-neutral (terminologies, dictionaries, vocabularies, etc.), external open data sources (such as

³ <https://openlaws.com/>

⁴ <http://www.poderjudicial.es>

those contained in the cloud of linked open data), and company private documents.

The core of the Lynx platform is a set of common services that are built upon existing and well tested technologies developed by the different technological partners: (i) the PoolParty semantic middleware suite; (ii) the FREME framework for multilingual and semantic enrichment of digital content; (iii) DKT API for digital curation processes; (iv) TILDE custom machine translation and cloud terminology services APIs; (v) OEG-UPM' APIs for ontology engineering and linked data publication, and (vi) K Dictionaries dictionary APIs.

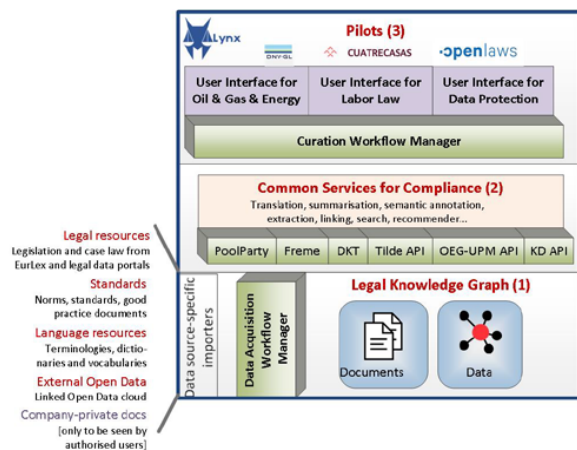


Figure 2. High level architecture of the Lynx platform

The common services provide core functionalities for annotating, linking, translating, and processing documents and data. Some of these services are invoked in order to treat new content arriving in the LKG (e.g., documents are annotated, linked or translated), and some of these services are used by the pilots, which are web-based applications to solve specific problems in three different domains. The invocation of these services for ingesting new content is coordinated by the Data Acquisition Workflow Manager, whereas their invocation for providing functionalities to the pilots is orchestrated by the Curation Workflow Manager. The common services are the backbone of a modular architecture which allows for the growing of input data sources and applications, as well as the dynamic scaling of the services in order to accommodate both surges and future expansion.

The platform will be instantiated in three pilots for each of the business cases considered in the project. The pilots will exploit different parts of the LKG and will make use of different common services according to their necessities. Each pilot will develop a specialised user interface that will capture the necessities of every business case according to their respective requirements.

These pilots support the vision of Lynx: that companies belonging to every sector of activity face compliance challenges when crossing the borders, and that an integrated approach to law and standards across different languages would increase competitiveness for companies in Europe reducing costs and corporate risks.

First, the Data Protection pilot will evidence the benefits of connecting legislation and case law from the EU and Members States. Second, the Oil & Gas and Energy pilot will demonstrate how the management of norms and industry standards is simplified by aggregating, comparing, and harmonising heterogeneous sources. Third, the Labour Law pilot will validate the appropriateness of the solution across different jurisdictions and languages, especially in such a complex domain as the Labour Law one, where each member state has different regimes, procedures, and standards.

6. Lynx Consortium

Ten partners from 7 different countries and complementary skills will work together to attain the objectives described in the previous sections. The Ontology Engineering Group from the Universidad Politécnica de Madrid is coordinating the project and is leading the Data Acquisition and Management work package. It is also contributing 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, will bring in its expertise in application of new technologies to the Legal Domain, and will lead the Industry requirements elicitation process, and the Dissemination and exploitation of 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, will lead the development of a set of curation tools, technologies and services to bridge between the 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. This company will lead the development of the Lynx platform core services, bringing in at the same time proprietary software components and semantic tools.

The specification of the technical architecture as well as the integration of all platform components will be performed by Alpenite, an IT software consulting and system integration company with headquarters in Italy.

The Latvian SME Tilde will bring in its expertise in multilingual natural language data processing technologies. Specifically, it will provide custom machine translation services and cloud terminology services, and will be also involved in other technological tasks, management, dissemination, and exploitation tasks.

The main provider of lexical data will be KDictionaries, an Israeli technology-driven-content creator that has developed quality lexicographic data for 50 languages.

Openlaws, another Austrian SME operating in the legal tech domain, will represent the use case on data protection, and will lead the development and integration of the two other pilots defined in the project. It will also contribute with core technology to the data acquisition and management tasks.

The use case on industry standards is led by DNV.GL, a standards certifying company with headquarters in Netherlands, Norway and Germany. This company will contribute 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 will be led by Cuatrecasas, a Spanish law firm with presence in over ten countries. It will also contribute with functional specification requirements and the development of the pilot.

9. Acknowledgements

Lynx has received funding from the Horizon 2020 European Union (EU) Research and Innovation programme under Grant Agreement: 780602. We thank all partners of the Lynx consortium for their contribution to the project's inception and proposal document. Special thanks to Ilan Kernerman and Georg Rehm for their suggestions to improve the paper.

7. Bibliographical References

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