Received: November 28, Date of acceptance: December 12, 2020, Date of publication: December 20 2020, DOI: https://doi.org/10.26826/law-in-context.v37i1.129

Research Note

LYNX: Towards a Legal Knowledge Graph for Multilingual Europe

By **Víctor Rodríguez-Doncel,** Associate Professor at the Artificial Intelligence Department of Universidad Politécnica de Madrid, **Orcid:** *https://orcid.org/0000-0003-1076-2511*

and **Elena Montiel-Ponsoda,** Associate Professor at the Applied Linguistics Department of Universidad Politécnica de Madrid, **Orcid:** *https://orcid.org/0000-0003-3263-3403*

Universidad Politécnica de Madrid, Madrid, Spain

ABSTRACT

Lynx is an innovation project in Europe whose objective is to develop services for legal compliance. A legal knowledge graph is built over multilingual, multijurisdictional documents using semantic web technologies. A collection of services implementing natural language techniques enables better legal information retrieval, cross-lingual answering of questions and information discovery. Three use cases are discussed, as well as the overall impact of the project.

Keywords – *Lynx Legal Knowledge Graph, compliance services, European legislation, multilingualism*

Acknowledgements: *This work has been funded by the project Lynx, which has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement no. 780602. For more information: http://www.lynx-project.eu.*

Disclosure statement – *No potential conflict of interest was reported by the authors.*

Suggested citation: Rodríguez-Doncel, V. and Montiel-Ponsoda, E. 2021. "Lynx:Towards a Legal Knowledge Graph for Multilingual Europe", *Law in Context, 37(1): 175-178, DOI: https://doi.org/10.26826/law-in-context.v37i1.129*

Summary

ISSN: 1839-4183

1. INTRODUCTION

The European Union (EU), post-Brexit, is comprised of 27 member states populated by approximately 450 million people, who speak over 60 indigenous languages with different legal status. The EU has adopted 24 official languages, and every EU national has the right to use any of these 24 languages to contact the EU institutions, and institutions are obliged to reply in the same language. Although the European Commission favors three of these languages as 'procedural languages' (English, French and German), EU law and many other legislative documents are published in all official languages except Irish. However, member states do not translate their legislation into foreign languages and finding the applicable norms in any location in one's own language is not easy.

The double fragmentation (of Law and languages) hampers the development of a unified market with fluid commercial exchanges in Europe. This has been recognized by the EU authorities, who set the completion of the Digital Single Market as one of their 10 political priorities for the term 2014-2019[1]. The actions required to implement this goal include disparate measures such as the abolishment of roaming charges for mobile telephones across Europe and the funding of applied research towards lowering barriers. Thus, the EU funded the now expiring Horizon 2020 Framework Program (H2020)[2] with 77,000 M€, and many of the projects included in the program have focused on language technologies.

The H2020 Lynx project ('Building the Legal Knowledge Graph for Smart Compliance Services in Multilingual Europe') is one of these efforts. This article describes how the combination of Semantic Web technologies and natural language processing techniques applied in the legal domain by the Lynx project opens the door to a new breed of legal information systems, cross-lingual and cross-jurisdictional, whose opportunities are yet to be fully explored.

2. RESULTS OF LYNX PROJECT

The Lynx project began in December 2017 with the objective of developing services to facilitate compliance in multilingual, multijurisdictional scenarios. The hypotheses were that semantic web technologies were mature and that natural language processing techniques had experienced a decisive leap forward that made their joint exploitation timely (Montiel-Ponsoda and Rodríguez-Doncel 2018). Technology would provide affordable services for compliance to be consumed by small and medium enterprises (SMEs). Funded with 3M€ until April 2021, the Lynx project was run by a consortium of 11 partners integrating universities (doing fundamental research), IT companies (providing real-scale infrastructure), language experts (machine translation companies, editors of dictionaries) and law firms (which are experiencing heavy transformations in their digitization processes). The main results of the project are the following:

2.1. A MULTILINGUAL LEGAL KNOWLEDGE GRAPH

A knowledge graph is a collection of interconnected entities with semantic types and properties, represented in a machine-readable form (Kroetsch and Weikum 2016). Much of the progress made by knowledge-related information systems in the last few years is owed to this technique, applied by large companies in private knowledge graphs or uploaded into the commons as linked open data by a myriad of parties (Bizer 2009). The cornerstone of the Lynx project is a knowledge graph of legal and regulatory information that brings together legislation and other legal resources from several jurisdictions in Europe. The technology that enables this integration of multiple sources is Resource Description Framework (RDF), a framework specified by the World Wide Web Consortium (W3C) to represent data in graph structures. Using the specifications of the W3C for the Semantic Web, data becomes more connected and is more easily interoperable. Application to the legal domain has been described by Casanovas et al. (2016). So far, the Lynx project has identified the major legal resources emerging from the European Union, Spain, Austria, Germany and the Netherlands, necessary to provide exploitable compliance services for Lynx business cases (Conejero et al. 2018). An ontology or data model has been created to identify and describe legal documents with metadata, structuring the content in as minimal way as possible. The so-called 'Legal Knowledge Graph Ontology' defined in Lynx is based on the European Legislation Identifier (ELI) and NLP Interchange Format (NIF). The ELI initiative,[3] adopted by the EU and several EU member states, specified 'a system to make legislation

[1] *A The end-of-term assessment of these priorities has been evaluated in a report: European Commission (2019). The Juncker Commission's ten priorities. Doi: 10.2861/618373*

[2] *http://ec.europa.eu/newsroom/horizon2020/document.cfm?doc_id=17607*

[3] *Council conclusions inviting the introduction of the European Legislation Identifier (ELI). DO C 325 de 26.10.2012*

available online in a standardized format' and includes an ELI ontology[4] which has proven useful in harmonizing the metadata of norms. The NIF ontology is aimed at providing text annotations with linguistic information (Hellman et al. 2014).

2.2. SOFTWARE

In addition to having defined data models and legal knowledge graphs, the Lynx project has also developed software. This software has taken the form of a service-oriented architecture and a collection of interoperable compliance-related services. These services are capable of (i) neural machine translation, using models specifically trained for the legal domain; (ii) annotation of documents, identifying named entities such as companies, temporal entities or persons; (iii) summarization of documents; (iv) question answering and cross-lingual search (Khvalchik et al. 2019) and (v) terminologically-related services.

Population workflows orchestrating these services have been put in place (Schneider et al. 2018 and 2020): original documents have been annotated with these NLP-based services, uniformly structured and linked to internal and external references. A major role in providing intelligence to these services is provided by the use of domain-specific terminologies, which support the information retrieval operations in cross lingual scenarios (Martín-Chozas et al. 2019).

2.3. DEMONSTRATION IN DIVERSE SCENARIOS

Three different application scenarios have been considered by the Lynx project. The aim of the first use case was to answer labor law-related questions posed by lawyers in plain language, considering both legislation and collective bargaining agreements. The second use case, focused on contract analysis, aimed at extracting the key information in text contracts of any kind, making them available for further processing in a structured form, thus reducing costs, and corporate and personal risks. The third use case, related to geothermal energy projects, aimed at retrieving the applicable legislation and technical standards that may apply in a certain location at a certain moment in time. These three heterogenous commercialized cases, prove the versatility of the Lynx solution.

3. IMPACT OF THE PROJECT

The Lynx project was modest in its objectives but has had some impact, both from the technological perspective and from the socio-legal perspective.

1) Lynx has demonstrated that there is a gap between the public legal information offered by official sources and the legal information provided by commercial systems, and has proved that the gap can be covered fairly cheaply. The maturity of semantic web technologies and the considerable improvement in the quality of open natural language techniques have created an opportunity for leveraging enriched public legal information to develop compliance by design solutions.
2) Lynx has made use of recently created resources in the linguistic linked open data cloud (Cimiano et al. 2020), demonstrating their value in a new application domain. Legal thesauri, terminologies and ontologies help to improve cross information retrieval tasks in different subdomains, such as data protection (Pandit et al. 2018), intellectual property (Rodríguez-Doncel et al. 2015), economics (Neubert 2009), and criminology (Schmidt et al. 2020).
3) Lynx has demonstrated the need to further develop uniform standards across Europe. Much as the ELI harmonized the way legislation is identified and described in the EU and some of its member states, there is a need for institutions to adopt uniform content structuring for legislative documents, using standards such as CEN Metalex[5] and OASIS LegalDocML.[6] The legislator is thereby encouraged to issue legislation not in traditional formats such as PDF, but also in machine-readable forms.
4) Access to legal information is democratized, and citizens have access to information previously available only to law firms paying fees. Although Lynx has used some commercial software systems (such as PoolParty[7] or Tilde's translation services[8]), and some non-open data (such as KDictionaries' Lexicala[9]), a large part of the source code and resulting data have been licensed openly.

[4] *https://publications.europa.eu/en/web/eu-vocabularies/eli*

[5] *http://www.metalex.eu/.*

[6] *http://www.legalxml.org/*

[7] *http://poolparty.biz*

[8] *http://tilde.com*

[9] *https://www.lexicala.com/*

ISSN: 1839-4183

4. CONCLUSIONS

Europe has invested in technologies to build a digital single market that overcome language and legal heterogeneities, and NLP and semantic web technologies have much to offer, as the Lynx project has shown –although much effort is still required. The legal knowledge graph developed in this project covers only very specific domains, is extended to limited jurisdictions and gives coverage to just a few languages. Moreover, it ignores case law (although other existing projects made great progress, like Boella et al. 2015) and the accrual methods to update the graph are limited to the project term. However, the Lynx project has set a precedent for others. The same needs covered by Lynx appear in other regions of the globe and the overall prospect of a new way of publishing and interlinking legislation will be beneficial for SMEs and citizens alike.

5. REFERENCES

1. Bizer, C. 2009. "The Emerging Web of Linked Data." *IEEE Intelligent Systems*, 24 (5): 87-92.
2. Boella, G., Di Caro, L., Graziadei, M., Cupi, L., Salaroglio, C. E., Humphreys, L., and Simov, K. 2015. "Linking legal open data: breaking the accessibility and language barrier in European legislation and case law". *In Proc. of the 15th Int. Conf. on Artificial Intelligence and Law*, ACM, pp. 171-175.
3. Casanovas, P., Palmirani, M., Peroni, S., Van Engers, T., and Vitali, F. 2016. "Semantic web for the legal domain: the next step". *Semantic Web*, 7 (3): 213-227.
4. Cimiano, P., Chiarcos, C., McCrae, J. P., and Gracia, J. 2020. "Linguistic linked open data cloud". In *Linguistic Linked Data* (pp. 29-41). Cham: Springer Nature, pp. 29-41.
5. González-Conejero, J., Casanovas, P., and Teodoro, E. 2018. "Business Requirements for Legal Knowledge Graph: the LYNX Platform". In *Technologies for Regulatory Compliance TERECOM@ JURIX*, pp. 31-38. http://ceur-ws.org/Vol-2309/03.pdf
6. Hellmann, S., Lehmann, J., Auer, S., and Brümmer, M. 2013. "Integrating NLP using linked data." In H. Alani et al. (eds.), *International Semantic Web Conference*: ISWC 2013, Part II, LNCS 8219, Berlin, Heidelberg: Springer, pp. 98-113.
7. Khvalchik, M., Blaschke, C., and Revenko., A. 2019. "Question Formulation and Question Answering for Knowledge Graph Completion". In Anderst-Kotsis G. et al. (eds), *Database and Expert Systems Applications.* DEXA 2019, Communications in Computer and Information Science, 1062, Cham: Springer, pp. 166-171.
8. Kroetsch, M. and Weikum, G. 2016. "Special Issue on Knowledge Graphs". *Journal of Web Semantics*, 37 (38): 53-54.
9. Martín-Chozas, P., Montiel-Ponsoda, E., and Rodríguez-Doncel, V. 2019. "Language resources as linked data for the legal domain". In G. Peruginelli and S. Faro (eds.), *Knowledge of the Law in the Big Data Age*. Amsterdam: IOS Press, pp. 170-180.
10. Montiel-Ponsoda, E., and Rodríguez-Doncel, V. 2018. "Lynx: Building the legal knowledge graph for smart compliance services in multilingual Europe". In G. Rehm, V. Rodríguez-Doncel, and J. Moreno-Schneider (eds.) *Proceedings of the 1st workshop on LREC (language resources and technologies for the legal knowledge graph) workshop*, Miyazaki, Japan, pp. 19-22. http://lrec-conf.org/workshops/lrec2018/W22/pdf/book_of_proceedings.pdf#page=26
11. Neubert, J. 2009. "Bringing the "Thesaurus for Economics" on to the Web of Linked Data". *Linked Data On the Web*, LDOW, http://ceur-ws.org/Vol-538/ldow2009_paper7.pdf
12. Pandit, H., Fatema, K., O'Sullivan, D., and Lewis, D. 2018. "GDPRtEXT-GDPR as a linked data resource." In *European Semantic Web Conference* (ESWC). Cham: Springer, pp. 481-495.
13. Rodriguez-Doncel, V., Santos, C., Casanovas, P., Gómez-Pérez, A., and Gracia J. 2018. "A Linked Data Terminology for Copyright Based on Ontolex-Lemon". In U. Pagallo, M. Palmirani, P. Casanovas, G. Sartor, and S. Villata (eds.) *AI Approaches to the Complexity of Legal Systems*. AICOL, LNCS 10791, Springer, Cham, pp. 410-423.
14. Schmidt, D., Dal Bosco, A., Trojahn, C., Vieira, R., and Quaresma, P. 2020. "Aligning IATE Criminal Terminology to SUMO". In *Int. Conf. on Computational Processing of the Portuguese Language*. LNCS 12037, Cham: Springer, pp. 98-108.
15. Schneider, J.M. and Rehm, G. 2018. "Towards a Workflow Manager for Curation Technologies in the Legal Domain". In G. Rehm, V. Rodríguez-Doncel, and J. Moreno-Schneider (eds.) *Proceedings of the 1st workshop on LREC (language resources and technologies for the legal knowledge graph) workshop*, Miyazaki, Japan, pp. 30-35.
16. Schneider, J. M., Rehm, G., Montiel-Ponsoda, E., Rodriguez-Doncel, V., Revenko, A., Karampatakis, S., and Maganza, F. 2020. "Orchestrating NLP Services for the Legal Domain". In *Proceedings of the 12th Language Resources and Evaluation Conference,* pp. 2332-2340. Available at: https://arxiv.org/abs/2003.12900