Navigating the Legal Landscape: Developing Italy's Official Legal Knowledge Graph for Enhanced Legislative and Public Services

Vito Walter Anelli[1,*], Eros Brienza[2], Marco Recupero[2], Francesco Greco[2], Andrea De Maria[2], Tommaso Di Noia[1] and Eugenio Di Sciascio[1]

[1]*Politecnico di Bari, via Orabona n.4, Bari, 70125, Italy*
[2]*Istituto Poligrafico e Zecca dello Stato, via Salaria 691, Roma, 00138, Italy*

Abstract

This paper details the creation of an official graph of Italian law, which is exposed using the Linked Data 5 stars standard and provides a SPARQL endpoint for users to query the data. The graph is constructed using a network analysis of legal documents, with each entity fully dereferenced to provide rich and comprehensive information. The graph also maps official descriptors to facilitate understanding and allows users to navigate the data visually using an intuitive interface. The paper discusses the technical implementation of the graph and the potential benefits for legislators and citizens.

Keywords

Semantic Web, Knowledge Graphs, Linked Data, Law Access, Legal Information

1. Introduction

In recent years, the adoption of Semantic Web technologies has provided a new framework for accessing and sharing information on the web. RDF (Resource Description Framework) triples, ontologies, and Linked Data principles have enabled data integration from different sources, making it possible to create a web of interconnected data. In the legal domain, the adoption of Semantic Web technologies has the potential to revolutionize the way legal information is accessed and processed. Using ontologies and RDF triples can enable the integration of legal data from different sources, making it possible to create a unified and interconnected legal knowledge base. This can facilitate the development of new legal applications and services that can improve the access to legal information for citizens and professionals alike.

The de-facto standard for representing structured information are Knowledge Graphs (KGs), initially conceived to connect documents by means of machine-understandable semantic links between the entities to improve data retrieving and access. The success of knowledge graphs is indicated by the results of the Linking Open Data initiative [1], which linked 1,483 different open-data knowledge graphs in the so-called Linked Open Data Cloud[1]. These knowledge graphs share the same ontology and schema across multiple domains, allowing reasoning, inference, and access to widespread knowledge. Among these graphs, DBpedia [2], Wikidata [3], Yago, FreeBase, Satori[2][3] [4], NELL, Google's Knowledge Graph[4], Facebook's Entities Graph[5], Knowledge Vault, and Bio2RDF are the most appreciated. The adoption of KGs as a source of side-information generated several advancements in the tasks of recommendation [5, 6], knowledge completion [7], preference elicitation [8], user modeling [9], and produced a vast literature.

The legal domain has always been considered challenging due to the overwhelming amount of unstructured text of the national legal corpus that makes it hard to retrieve precise information, similarity, and connections between the various laws. These challenges have sparked the interest in developing specialized versions of large language models [10], such as Chalkidis et al. [11][6], Strubell et al. [12], and Liu et al. [13][7], to enable additional automated tasks that are infeasible for a single human being. However, Large Language models and RDF technologies are not enough to guarantee data is actually "linked".

Ital-IA 2023: 3rd National Conference on Artificial Intelligence, organized by CINI, May 29–31, 2023, Pisa, Italy
*Corresponding author.
vitowalter@poliba.it (V. W. Anelli); e.brienza@ipzs.it (E. Brienza); m.recupero@ipzs.it (M. Recupero); f.greco@ipzs.it (F. Greco); a.demaria@ipzs.it (A. De Maria); tommaso.dinoia@poliba.it (T. Di Noia); eugenio.disciascio@poliba.it (E. Di Sciascio)
0000-0002-5567-4307 (V. W. Anelli); 0000-0002-0939-5462 (T. Di Noia); 0000-0002-5484-9945 (E. Di Sciascio)

[1]https://lod-cloud.net/datasets
[2]https://searchengineland.com/library/bing/bing-satori
[3]https://blogs.bing.com/search/2013/03/21/understand-your-world-with-bing
[4]https://blog.google/products/search/introducing-knowledge-graph-things-not/
[5]https://www.facebook.com/notes/facebookengineering/under-the-hood-the-entitiesgraph/10151490531588920/
[6]https://huggingface.co/nlpaueb
[7]https://github.com/autoliuweijie/K-BERT

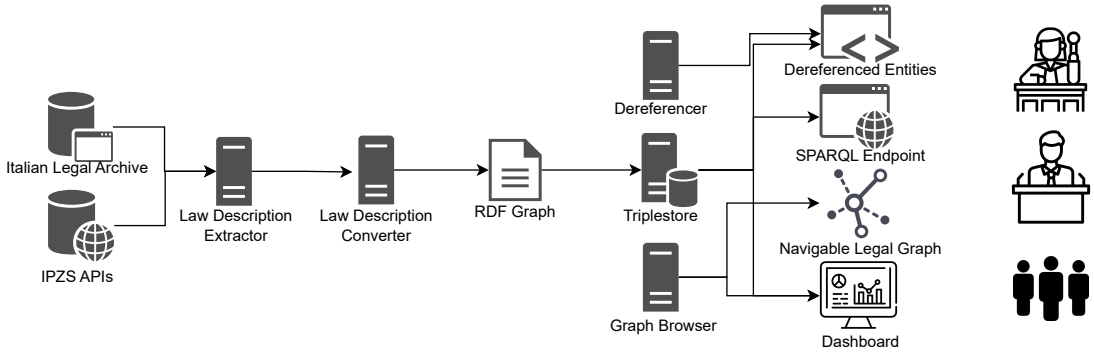

Figure 1: Overview of the System Architecture. The system analyzes the normative corpus and construct an RDF graph then stored in the triplestore. The triplestore is served by exposing a SPARQL endpoint and activates (i) a visual SPARQL endpoint, (ii) a dereferencing tool, (iii) a tool to navigate the normative graph, and (iv) a dashboard to query the system using transparent SPARQL endpoint queries. The human operators, a Judge/lawyer, a Legislator, and the citizens will have four possible interfaces to use, depending on the activity to be performed and their technical preparation.

Indeed, an ontological alignment was still missing.

To this extent, the European Union published the EuroVoc thesaurus[8] to classify legal documents. EuroVoc is a multilingual thesaurus maintained by the Publications Office of the European Union[9], which contains more than 7,000 concepts referring to various activities of the EU and its Member States. The thesaurus is used as a classification schema for the two most well-known legal datasets, *JRC-AcquisV3* [14] and *EURLEX57K* [15], which contain legal documents from the legal information system of the European Union (Eur-Lex). EuroVoc provides all its terms in the official language of the EU member states to enable a multilingual search. The EuroVoc thesaurus has been introduced to harmonize the classification of documents in the communications across EU institutions.

The Official Gazette of the Italian Republic[10] is the official source of knowledge for the laws in force in Italy. It has the mission of disseminating to all citizens the information published in the Gazette in its series, the general and the special ones. The Gazette is published by the Istituto Poligrafico e Zecca dello Stato (IPZS) in collaboration with the Ministry of Justice, which provides for drafting the laws and directing their publication. According to the Law of 13 July 1966, n. 559, IPZS provides for the printing and management, also with IT tools, of the Official Gazette. Legislative information has not been structured in electronic format since the beginning. The management system was conceived to take solely care of the editorial aspect of the publication process of the Official Gazette. This approach caused some limitations:

- the juridical collections are defined on an editorial basis and not on a reliable and detailed classification;
- it is not designed to automatically correlate acts (e.g., in order to easily identify impacts deriving from the repeal, or promulgation, of a law).

To overcome this limitations, Politecnico di Bari and Istituto Poligrafico Zecca dello Stato are cooperating to create an ecosystem capable of representing the graph of Italian regulations. We have realized an experimental pipeline that extracts relevant information from the documents currently produced by IPZS. Currently, this process is performed manually, and the graph is deduced only by the appointed expert. Therefore, the project is part of a set of initiatives aimed at innovating the techniques used by IPZS and improving service delivery.

The project's contributions are significant as they introduce automation to a process that was previously manual, reducing the risk of errors and inconsistencies. Furthermore, the use of containerization technology allows for easy deployment and scaling of the system, making it more efficient and flexible. Moreover, the project uses an approach to document classification based on the extraction of relevant information, including the text, hyperlinks, and connections to other regulations using three main sources: the previously annotated correlations, the textual references to other laws, and the semantic similarity between the laws.

For what regards the framework to expose the graph, it is composed of the three technological pillars, a triplestore, a dereferencer, and a navigator. These tools enable the construction, representation, and navigation of the RDF graph, making it more accessible and user-friendly. The triplestore is used to store the files generated by the processing of the normative corpus, while the dereferencer provides a user-friendly interface to display the information stored about a specific regulation. The navigator allows users to explore the graph visually, enabling them to understand the relationships between different

[8] https://eur-lex.europa.eu/
[9] https://publications.europa.eu/en/web/eu-vocabularies
[10] https://www.gazzettaufficiale.it

regulations.

Overall, the project's contributions are manifold, as the project introduces an automate approach to the creation of the graph of Italian regulations, which was previously a manual process. The use of containerization technology and the integration of tools to store, represent and navigate the RDF graph makes it more accessible and user-friendly. The system's deployment is improving the quality of the service provided by Istituto Poligrafico Zecca dello Stato, and it is an important step towards the innovation of the techniques in the regulation industry.

2. Italian Legal System Management: Domain Description

The process of publishing a law involves the Italian Ministry of Justice and the Istituto Poligrafico e Zecca dello Stato (IPZS). From the beginning, this process has undergone several variations. The technologies that have followed one another over time have been different, and even today, there are manual activities that introduce the risk of performing rework. Figure 2 shows the process of publishing of a law in the general series.

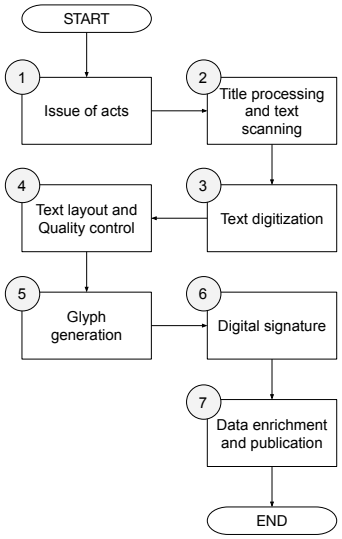

Figure 2: The process of publishing an Italian law: (i) The Ministry of Justice issues the law, (ii) the metadata are assigned to the law, (iii) irrespectively to the original format, a digital version of the law is generated, (iv) the law is laid out and paginated, (v) a security glyph is generated and applied, (vi) the digital law is signed, (vii) the law is combined with metadata and published.

1. The act is issued by the Ministry of Justice in paper or digital format;
2. The composition of the titles of the documents is carried out and a scanned text is attached to each title. A publication of the Official Gazette is assigned;
3. Digitization of text also using OCR techniques;
4. The act is impaginated and undergoes a quality check;
5. A security glyph is generated and applied for each page of the Gazette;
6. Following several checks, the paginated law is digitally signed to create the act certified version.
7. The act is classified, enriched with metadata and relationships with other acts. The act is published.

3. Overview of System's operation

In this study, we present a framework, illustrated in Figure 1, designed to analyze the normative corpus provided by the Istituto Poligrafico Zecca dello Stato and process it to construct a turtle file that can be used to build the RDF graph indices in the triplestore. We treated the graph preparation phase as an independent development phase, as a future replacement of the norm extraction parameter model will require a different data preparation process. The turtle files are then used to build a graph in the triplestore, which is finally served by exposing a query function. This functionality activates a dereferencing tool that allows for a human-friendly and concise visualization of the stored information regarding a specific norm. This tool provides access to another service that is properly configured to navigate the normative graph using transparent SPARQL endpoint queries. Therefore, the human operator will have four possible interfaces to use, depending on the activity to be performed and their technical preparation.

The project involved several design phases, including defining use cases and the actors involved, loading the database and transforming it to match the necessary data model for training, identifying data cleaning activities for the provided data type, researching and identifying methods for integrating the machine learning system with the rest of the software architecture.

4. Architecture of the System

The system for the representation, visualization, and querying of the graph of Italian laws is designed as a tool that not only assists operators in identifying regulatory impacts but also innovates and simplifies a burdensome decision-making process for human operators. In production, this system will be useful in simplifying all other activities involving potential connections between the regulations of the Italian legal system. The system includes the following modules: (i) Law Description Extractor, (ii) Law Description Converter, (iii) Triplestore,

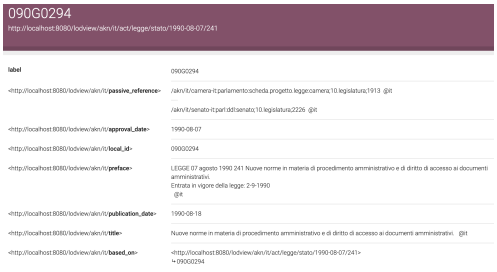

Figure 3.1: An italian law is described as an entity in the Knowledge Graph. The description comprises the editorial code being the label in the preview, the approval and the publication date, the references to the other related laws, and other domain specific descriptors.

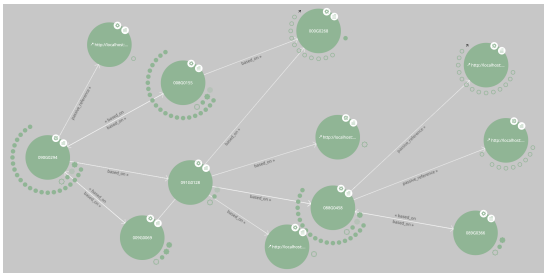

Figure 3.2: A subgraph of the Italian legal system generated while browsing the Knowledge Graph exploiting the Lodlive technology. The initial entity is expanded enabling the visualization of the related laws.

(iv) Dereferencer, (v) Graph Browser, (vi) SPARQL Endpoint, (vii) User-friendly Dashboard.

Law Description Extractor Module. The Extractor loads the entire regulatory database available to the Istituto Poligrafico Zecca dello Stato. Despite the current storage of regulations in the form of files, the module allows to retrieve such information remotely through appropriately prepared APIs. The module is currently able to read and load the text of each regulation and all metadata provided by IPZS. The module includes the following features: (i) Read regulation and its descriptors; (ii) Collection of regulations into an interoperable data structure suitable for subsequent conversion into a graph; (iii) Recognition of currently in-force regulations.

Law Description Converter Module. The Converter transforms the data structure of regulations into a file suitable for loading into a triplestore. Currently, the used format is turtle. However, as the graph representation system used may vary, so may the data processing activities. Therefore, the module provides a modular structure, where each transformation phase can be activated and deactivated if necessary. The module includes the following features: (i) Construction of the regulation skeleton in the form of a graph; (ii) Identification and use of prefixes and predicates currently in use in Linked Open Data; (iii) Composition of the descriptor triples of a regulation; (iv) Saving the database in an interoperable format for graph representation.

Triplestore Module. The Triplestore loads the triples previously produced into a system capable of constructing the RDF graph and its respective indices. The graph can be divided into multiple subgraphs, and the system must allow individual subgraphs to be loaded and removed. The module includes the following features: (i) Loading files representative of the graph; (ii) Identification of individual subgraphs; (iii) Querying of subgraphs; (iv) Updating of subgraphs.

Dereferencer. The Dereferencer allow for publishing data in RDF format following the Linked Open Data standards through a SPARQL endpoint. The tool guarantees a good user experience by providing access to RDF resources through web pages. Additionally, the module is fully customizable, enabling the modification of the web interface, page style, and all the content that needs to be displayed. In addition to dereferencing resources, the module must allow for publishing data in RDF, guaranteeing different solutions for serialization. For example, it is possible to access the same resource by requesting a response in HTML (web page), JSON, turtle, n-triples. The module includes the following features: (i) Access to individual entities (norms) of the graph; (ii) Reporting all information related to the entity; (iii) Saving information related to an entity.

Graph Browser. This module allow for navigating data exposed in the form of Linked Open Data, requiring only a SPARQL endpoint to be contacted. Starting from a node within the graph, it lets navigate all the relations starting from it, thus reaching all connected entities. Proceeding recursively, it allows for analyzing and navigating the knowledge graph without any limitation. The module includes the following features: (i) Visualization of the entity and its related predicates; (ii) Visualization of the objects of triples (information about the norm or other norms); (iii) Possibility to continue navigation through new encountered norms.

SPARQL Endpoint. This module exposes the query functionalities on the graph using the standard SPARQL language. Query results could be obtainable also by using APIs and can be returned using different interoperable formats. Since the methods for using the model could vary depending on the system of classification used, the module provides a unified interface for the use of such functions. The module includes the following features: (i) APIs for accessing the query system; (ii) Graphical interface for writing queries; (iii) Possibility to choose the response format, also through content negotiation.

User-friendly Dashboard. The Dashboard makes available several ready-made queries for the user who does not have familiarity with the SPARQL query language. Since future queries may differ from those currently available, the module is divided into modular tasks that can be easily maintained and extended. The module includes the following features: (i) Graphical interface that masks the actual complexity of queries; (ii) Possibility to choose the response format; (iii) Modular encapsulation of queries.

5. A platform at the service of Legislators and Citizens

In order to identify the user-type of the services implemented following this study, the "personas model" was employed. The model describes the behaviors, difficulties and needs of the two main typical users identified by the study: the legislator and the lawyer/judge.

Persona 1 - Violet (lawyer): Profile description and opportunities of the system in relation to her needs. Violet is a lawyer who consults online legislative collections daily to search for normative references to use for her hearings. The consultation activities bring good results but usually take several hours of careful work.

The difficulties in carrying out her work are due to the fact that the laws generally refer to many other laws, and often the connections between them are challenging to find. The legislative database is wide, and a satisfactory search often requires many hours of work. Moreover, consultation systems require precise text inputs, not being able to understand natural language.

Violet needs to reduce his research times without compromising the quality of his work. She would like to carry out smart queries, which accept more flexible inputs and return all the material she needs. The output should not be fragmented way and easily navigable.

The system under study provides a knowledge graph of the entire legislative database and recognizes active and passive impacts between the laws. Following a query, the system isolates and returns a subgraph. This representation greatly simplifies Violet's work; now, she can navigate within the subgraph and expand it in the depth that she prefers.

Persona 2 - Paul (legislator): Profile description and opportunities of the system in relation to his needs. Typical operations in charge of Paul are the introduction, the modification, the repeal of a law, and the transposition of a European directive. He deals with the gold-plating processes of European directives and with simplifying and merging laws when they are too fragmented.

These and other operations are never risk-free and could cause unwanted impacts on other laws in force. The legislative corpus is wide and not devoid of entropy. Sometimes, the only solution to make a set of laws more understandable is to group them into a consolidated text.

Paul needs an assisted system that can make the activities of identifying and verifying legislative impacts. The system under study includes the existing impacts between laws and is able to receive a draft of a new law as input, understand its natural language, and carry out an assisted preliminary verification of the impacts that this would have on other laws in force after its promulgation.

6. Conclusion

In conclusion, Politecnico di Bari and Istituto Poligrafico Zecca dello Stato developed an experimental pipeline that automates the creation of the graph of Italian regulations. The project reduces the risk of errors and inconsistencies by introducing automation to a previously manual process. Furthermore, the use of containerization technology and the integration of tools to store, represent, and navigate the RDF graph makes it more accessible and user-friendly. The project is an important step towards the innovation of the techniques used in the regulation industry, and it improves the quality of the service provided by Istituto Poligrafico Zecca dello Stato. The framework enables the construction, representation, and navigation of the RDF graph, making it easier for users to explore the relationships between different regulations. Overall, the project represents a significant achievement in the field of regulation management.

Acknowledgements

The project IPZS-PRJ4_IA_NORMATIVO supported this research.

References

[1] T. Heath, C. Bizer, Linked Data: Evolving the Web into a Global Data Space, Synthesis Lectures on the Semantic Web, Morgan & Claypool Publishers, 2011. URL: https://doi.org/10.2200/S00334ED1V01Y201102WBE001. doi:10.2200/S00334ED1V01Y201102WBE001.

[2] J. Lehmann, R. Isele, M. Jakob, A. Jentzsch, D. Kontokostas, P. N. Mendes, S. Hellmann, M. Morsey, P. van Kleef, S. Auer, C. Bizer, Dbpedia - A large-scale, multilingual knowledge base extracted from wikipedia, Semantic Web 6 (2015) 167–195. URL: https://doi.org/10.3233/SW-140134. doi:10.3233/SW-140134.

[3] D. Vrandecic, M. Krötzsch, Wikidata: a free collaborative knowledgebase, Commun. ACM 57

(2014) 78–85. URL: https://doi.org/10.1145/2629489. doi:10.1145/2629489.

[4] D. Liu, T. Bai, J. Lian, X. Zhao, G. Sun, J. Wen, X. Xie, News graph: An enhanced knowledge graph for news recommendation, in: V. W. Anelli, T. D. Noia (Eds.), Proceedings of the Second Workshop on Knowledge-aware and Conversational Recommender Systems, co-located with 28th ACM International Conference on Information and Knowledge Management, KaRS@CIKM 2019, Beijing, China, November 7, 2019, volume 2601 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2019, pp. 1–7. URL: http://ceur-ws.org/Vol-2601/kars2019_paper_01.pdf.

[5] A. Ferrara, V. Anelli, A. Mancino, T. Di Noia, E. Di Sciascio, Kgflex: Efficient recommendation with sparse feature factorization and knowledge graphs, ACM Transaction on Recommender Systems 1 (2023). URL: https://doi.org/10.1145/3588901. doi:10.1145/3588901.

[6] V. W. Anelli, T. D. Noia, E. D. Sciascio, A. Ragone, J. Trotta, Semantic interpretation of top-n recommendations, IEEE Trans. Knowl. Data Eng. 34 (2022) 2416–2428. URL: https://doi.org/10.1109/TKDE.2020.3010215. doi:10.1109/TKDE.2020.3010215.

[7] G. He, J. Li, W. X. Zhao, P. Liu, J. Wen, Mining implicit entity preference from user-item interaction data for knowledge graph completion via adversarial learning, in: Y. Huang, I. King, T. Liu, M. van Steen (Eds.), WWW '20: The Web Conference 2020, Taipei, Taiwan, April 20-24, 2020, ACM / IW3C2, 2020, pp. 740–751. URL: https://doi.org/10.1145/3366423.3380155. doi:10.1145/3366423.3380155.

[8] V. W. Anelli, R. D. Leone, T. D. Noia, T. Lukasiewicz, J. Rosati, Combining RDF and SPARQL with cp-theories to reason about preferences in a linked data setting, Semantic Web 11 (2020) 391–419. URL: https://doi.org/10.3233/SW-180339. doi:10.3233/SW-180339.

[9] H. Wang, F. Zhang, J. Wang, M. Zhao, W. Li, X. Xie, M. Guo, Exploring high-order user preference on the knowledge graph for recommender systems, ACM Trans. Inf. Syst. 37 (2019) 32:1–32:26. URL: https://doi.org/10.1145/3312738. doi:10.1145/3312738.

[10] D. Licari, G. Comandé, ITALIAN-LEGAL-BERT: A pre-trained transformer language model for italian law, in: D. Symeonidou, R. Yu, D. Ceolin, M. Poveda-Villalón, D. Audrito, L. D. Caro, F. Grasso, R. Nai, E. Sulis, F. J. Ekaputra, O. Kutz, N. Troquard (Eds.), Companion Proceedings of the 23rd International Conference on Knowledge Engineering and Knowledge Management, Bozen-Bolzano, Italy, September 26-29, 2022, volume 3256 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2022. URL: https://ceur-ws.org/Vol-3256/km4law3.pdf.

[11] I. Chalkidis, M. Fergadiotis, P. Malakasiotis, N. Aletras, I. Androutsopoulos, LEGAL-BERT: "preparing the muppets for court'", in: T. Cohn, Y. He, Y. Liu (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2020, Online Event, 16-20 November 2020, volume EMNLP 2020 of *Findings of ACL*, Association for Computational Linguistics, 2020, pp. 2898–2904. URL: https://doi.org/10.18653/v1/2020.findings-emnlp.261. doi:10.18653/v1/2020.findings-emnlp.261.

[12] E. Strubell, A. Ganesh, A. McCallum, Energy and policy considerations for deep learning in NLP, in: A. Korhonen, D. R. Traum, L. Màrquez (Eds.), Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers, Association for Computational Linguistics, 2019, pp. 3645–3650. URL: https://doi.org/10.18653/v1/p19-1355. doi:10.18653/v1/p19-1355.

[13] W. Liu, P. Zhou, Z. Zhao, Z. Wang, Q. Ju, H. Deng, P. Wang, K-BERT: enabling language representation with knowledge graph, in: The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020, AAAI Press, 2020, pp. 2901–2908. URL: https://aaai.org/ojs/index.php/AAAI/article/view/5681.

[14] E. L. Mencía, J. Fürnkranz, Efficient multilabel classification algorithms for large-scale problems in the legal domain, in: E. Francesconi, S. Montemagni, W. Peters, D. Tiscornia (Eds.), Semantic Processing of Legal Texts: Where the Language of Law Meets the Law of Language, volume 6036 of *Lecture Notes in Computer Science*, Springer, 2010, pp. 192–215. URL: https://doi.org/10.1007/978-3-642-12837-0_11. doi:10.1007/978-3-642-12837-0_11.

[15] I. Chalkidis, M. Fergadiotis, P. Malakasiotis, I. Androutsopoulos, Large-scale multi-label text classification on EU legislation, in: A. Korhonen, D. R. Traum, L. Màrquez (Eds.), Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers, Association for Computational Linguistics, 2019, pp. 6314–6322. URL: https://doi.org/10.18653/v1/p19-1636. doi:10.18653/v1/p19-1636.