Automatic End-to-End Decomposition and Semantic Annotation of Laws Using High-Performance-Computing and Open Data as a Potential Driver for Digital Transformation

Charalampos Alexopoulos
Department of Information and Communication Systems Engineering
University of the Aegean |
Karlovassi, 83200 Samos, Greece
alexop@aegean.gr

Igor Pihir, Martina Tomičić Furjan
Faculty of Organization and Informatics
University of Zagreb
Pavlinska 2, 42000 Varaždin
{ipihir, mtomicic}@foi.unizg.hr

Abstract. *This paper is dealing with automatic end-to-end law analysis conducted by decomposition and semantic annotation, by using the high-performance-computing and government open data. Legal data and law texts are a category of open data and thus they possess the potential to unlock digital innovation and transformation capacity in governments and businesses, regarding the development of new, better, and more cost-effective services for citizens. For that reason, they can be recognized as a potential digital transformation driver. This research presents a baseline for automation of decomposition and annotation with process and service elements developed for utilization on high-performance-computing infrastructure based on government laws open data and gives insights on how the results of it can initiate digital transformation.*

Keywords. Digital Transformation, Laws Decomposition and Annotation, Automation, Text mining, Hight-Performance-Computing, Open Data.

1 Introduction

The legislation of each state is very important to be open and available online in order for every part of society to have free and unhindered access (Fulton, 2011). Financial and legal services of academia and research institutions are constantly dealing with laws and their interpretation. All new and already existing procedures have been designed or will be designed based on the analysis of legal documents. In addition, significant advancements in the 'legal informatics' research field are observed since governments have started to promote the development of legal information systems (Casanovas et al., 2016). Disruptive technologies like text mining could be used for the development of novel legal information systems and services that able to automatically extract information from unstructured legal data.

This opens possibilities for innovation and digital transformation driven by these newly available legal data. Transformation can thereby be achieved through one or more new business concepts: new service, new product, business model, etc. (Tomičić Furjan, Tomičić-Pupek and Pihir, 2020) enabled by the results of performed laws analysis. "Digital transformation (DT) can be seen as an approach which assumes a radical change of doing business, driven highly by today's necessity to change and adapt to the digital age. This change is mostly made by creating new business models, which define the way how an organization provides value to the customers" (Hrustek, Tomičić Furjan & Pihir, 2019). Data retrieved or created from laws, analysed as proposed in this paper and provided as open data could initiate this change and become a potential DT driver. Open Data Process model for open data paradigm is described in paper (Hrustek, Tomičić Furjan & Pihir, 2020)

The main objective of this paper is to present the service towards the automatic decomposition (creation of legal data) and semantic annotation (based on specific and well-recognized standards) using text mining techniques on laws. It is meant to find a way to build and test a fully automated process towards the creation of big, linked, open legal data.

The Legal documents have specific characteristics which differentiate them from the "normal daily-used" documents due to their length and complexity (Nguyen et al., 2011). The complexity of the legislation is identified in the comprehending of the legal terminology, in other words, legal language, in combination with the existence of correlations as references among laws or other kinds of legal documents. There are different types of legislative documents such as laws, presidential decrees and ministerial decisions. Every type consists of a specific

form including specific structural elements (components), but in this paper we are dealing only with laws. Each reference may amend one or more different laws, and, in many cases, an in-depth search of previous laws is required to detect which component is in force. Such a service of creating and providing big, linked, open legal data (BLOLD) should be supported by a powerful high-performance-computing infrastructure in order to be able to analyse the large amount of data structures and the correlations among them. Based on the results of this service additional and added-value services could be developed. This way, we envision the facilitation of public access for all types of users to open legal data accessed and served via a Legal Web platform in a customizable, structured, intuitive and easy-to-handle way (Lachana et al., 2020). Furthermore, the data will be available for all countries complying with the developed standard and developed text techniques.

The added-value services could be described as follows: i) research through legal corpora, analyzing the alignment of national legislation with EU legislation, ii) comparing national laws which target the same life events, iii) analyzing the references to European legislation by national laws, iv) analyzing related laws within the same Member State, v) timeline analysis for all legal acts, vi) visualization of the progress and current status of a specific national or European piece of legislation and vii) sentiment analysis towards new legislation.

The paper starts with the introduction to the problem, followed by a methodology section and furthermore development of the idea across a literature review in form of a state-of-the-art chapter. The main part of the paper is oriented to the metadata scheme for decomposition and semantic annotation, HPC processing infrastructure, and how the decomposition and semantic annotation automated process is working, and which services are used/developed to support it. It ends with concluding remarks, acknowledgment, and bibliography.

2 Methodology

Research presented in this paper is part of larger project initiative streaming to automate the decomposition and semantic annotation of laws in such a way, so that the purpose and meaning of laws could be reused, compared, and automatically processed in process and semantic way. Research uses literature review, process and data modelling, metamodeling, datamining and high-performance computation to extract data and automation metadata-based annotation.

This research is oriented to problem description, literature review to position the problem, and architecture development for automation of decomposition and semantic annotation process. Research is conducted on results of previously ended hundreds of processes analysis and by finding and classifying the paths to automated processing of legal documents which in the end can trigger digital transformation.

3 The state of the art in legal text mining

One of the most challenging, but also most potential developments, comes with the web of data (Avila-Garzon, 2020) and the inherent mass of freely-available information, i.e., open data (Zeleti, Ojo, & Curry, 2016). Legal data and law texts is a category of open government data and thus they possess the potential to unlock innovation in governments and businesses, regarding the development of new, better, and more cost-effective services for citizens (Zuiderwijk & Janssen, 2014). In case of public administrations and governments, the distribution, availability and access towards legal information is crucial. Yet, there exist some severe issues at the moment regarding this access. One of them is found in form of available APIs, which are not always up and running on a 24/7 basis, paired with slow systems and often non-compliant data towards standard or even self-issued schemata. This in turn makes the use of automated crawling and analysis more than difficult (Charalabidis, et al., 2018).

Text mining, also known as text data mining, intelligent text analysis (Gupta & Lehal, 2009; Erhardt, Schneider, & Blaschke, 2006) or knowledge discovery from textual (structured) databases (Justicia De La Torre et al., 2018; Delen & Crossland, 2008), has been defined as “the discovery by computer of new, previously unknown, information by automatically extracting information from different written resources” (Hassani et al., 2020). Generally, text mining refers to the process of extracting interesting and non-trivial patterns or knowledge from unstructured text documents (Justicia De La Torre et al., 2018; Gupta & Lehal, 2009). Legal text mining analyses legal texts in order to extract useful legal information such as an overview of text’s content (Merkl & Schweighofer, 1997). Legal text documents are being unstructured stocked except in cases that online legal databases provide easy access to citizens, businesses etc. Furthermore, as pointed out by (Hassani et al., 2020) legal text documents are stored using natural language, so text mining can be suitably used for efficient analysis of such documents.

4 Metadata schema for law decomposition and semantic annotation based on Open Data

The Manylaws metadata schema is based on three standard ontologies, each one describing specific

elements of a document. The Data Catalog Vocabulary (DCAT)[1] is used to transform legal data in open legal data, the European Legislation Identifier (ELI)[2] to describe correlations between laws and EU Directives and Akoma Ntoso (AKN)[3] to describe parliamentary, legislative documents and the document lifecycle. Manylaws specific elements are added to cover the need of searching across borders for a specific term.

The scope of this complicated schema is to support the services of the Manylaws portal and to convert the legal documents into big, linked, open legal data. The basic elements that are connected with the Manylaws Services are included in Table 1. Specifically, the main scope of using DCAT-AP[4] is to convert the legal document to open data and the opportunity to being harvested by the European Data Portal (EDP) according to its guidelines, using all the mandatory fields that are required for a dataset in order to be harvested from the EDP[5] (e.g the title, the keywords, the theme of the legal document, the location, the description etc). ELI is used in order to imprint all the references that are included to a legal document because more and more European countries are using this ontology to describe legal documents. AKN is used in order to cover all the parliamentary supporting documents and steps that are required for the publication and adoption of a law. Furthermore, AKN is using a more flexible way to present the body of a legal document (e.g. lifecycle, workflow and body). Finally, the Manylaws Platform Extra (MLEXTRA) elements are being used to cover all the data that are not covered from the previous ontologies/ vocabularies. Particularly, the elements contain information for the similarity between two laws, between the articles of two laws, the translated body of the law, the nouns that are included in the body of the law and the ngrams that are a combination between adjectives and nouns, verbs and nouns, etc.

These three basic ontological approaches have been aligned and matched to each other in order to avoid duplication of metadata and data fields. This way, the final metadata schema includes the mapped fields and increases the flexibility of the whole approach. As the final step of the process, all data is stored into a virtuoso endpoint, so they are offered as linked data enabling also direct acquisition from the European Data Portal.

5 The process using text mining and HPC

Figure 1 describes the automated process of decomposition and semantic annotation. All the components and they are not triggered by user actions. They are related to processes that either retrieve information when available or produce new information from the retrieved data which will be stored in XML/RDF files when ready. These components operate periodically, at fixed intervals and their content is often renewed. These micro-services could be connected to any source available retrieving the available data, in most of the cases in .pdf format. For example, in the Greek case, a web crawler is used for retrieving the Greek legislation from the Greek National Printing Office (Greek NPO). The retrieved data (PDFs) are being stored in a pre-processing repository.

Whenever a new file is being stored, the RapidMiner is triggered and initiates Hadoop. The Hadoop Cluster is used for parallelized processing of massive data in a reasonable time. Specifically, for the Greek data Hadoop is used for extracting all the necessary metadata from the legal documents and for annotating these metadata with semantics in rdf/xml (based on ManyLaws metadata schema).

As soon as the semantic annotation of the metadata is finished, Hadoop transmits all the data to be translated to the eTranslation DSI which updates the Virtuoso triple store after the translation is finished with the updated rdf/xml.

[1] https://dcat.org/

[2] https://eur-lex.europa.eu/eli-register/about.html

[3] http://www.akomantoso.org/

[4] https://joinup.ec.europa.eu/collection/semantic-interoperability-community-semic/solution/dcat-application-profile-data-portals-europe

[5] https://data.europa.eu/en

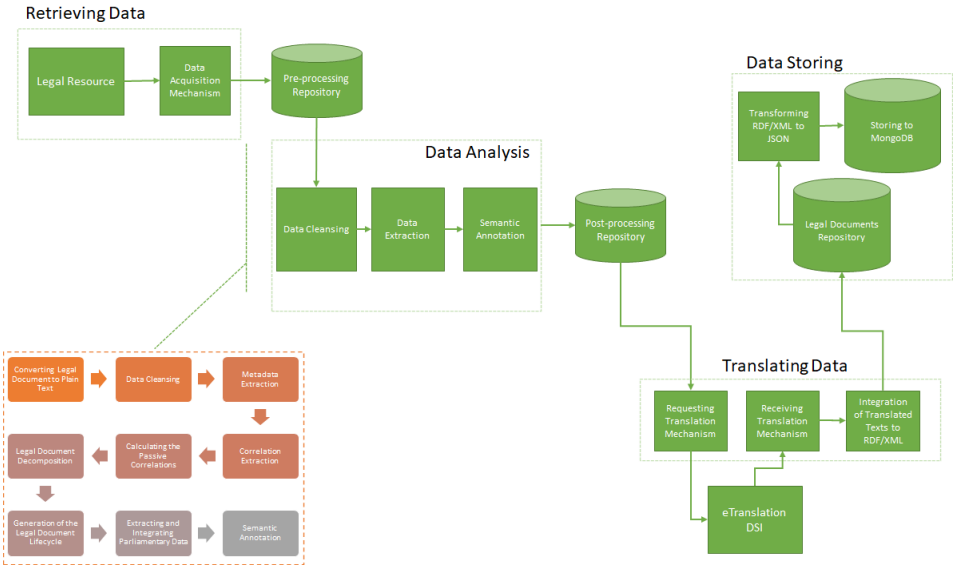

Figure 1: The decomposition and semantic annotation automated process

Given the vast amount of legal data published in the participating EU Member States, as well as their respective languages, analyzing it requires powers beyond that of a simple desktop computer, and thus High-Performance Computing is necessary.

6 Utilization of high-performance-computing infrastructure

This section presents the tools and algorithms of the designed process and the results of the high-performance computing (HPC) infrastructure usage in order to process and annotate big linked open legal data. The necessary tools and algorithms capable of handling large volumes and high velocity of data in the HPC environment have been developed and tested in the Greek and European case. The Greek National HPC Infrastructure "ARIS"[6] has been utilized in order to analyse a big number of legal documents.

The services and applications ran for 188 hours in the HPC system. The resources consumed are as follows: 8 nodes each one with 40 processing cores (totalling 320 processing cores); the total utilization of computing power amounted to 60,246 core hours; during the HPC operation 11,363,596 datasets were generated including the testing periods. It worth mentioning that the results achieved through the HPC operation in about 20 days, would require a high-end personal computer to work for almost 3,5 years.

Table 1 describes the micro-services that are necessary for the decomposition and semantic annotation. These micro-services are designed to work in parallel, so the utilization of an HPC infrastructure is possible. Thus, the development of the Message Passing Interface (MPI) was necessary.

MPI is an open library standard for distributed memory parallelization. The library API (Application Programme Interface) specification is available for C and Fortran. There exist unofficial language bindings for many other programming languages, e.g. Python a, b or Java1, 2, 3. The first standard document was released in 1994. MPI has become the de-facto standard to program HPC cluster systems and is often the only way available. There exist many implementations, Open Source and proprietary. The latest version of the standard is MPI 3.1 (released in 2015). The MPI interface is meant to provide essential virtual topology, synchronization and communication functionality between a set of processes (that have been mapped to nodes/servers/computer instances) in a language-independent way, with language-specific syntax (bindings), plus a few language-specific features. MPI programs always work with processes, but programmers commonly refer to the processes as processors. Typically, for maximum performance to be achieved, each CPU (or core in a multi-core machine) will be assigned a single process. This assignment happens at runtime through the agent that starts the MPI program, normally called mpirun or mpiexec.

The developed MPI program used to support our process divides the main process into forty (40) micro-services. For this purpose, we developed a script to divide the bulk data into forty parts. Each micro-service runs a specific part of the process on the bulk legal data that requires processing (data mining and semantic annotation) to be aligned with the developed metadata schema.

[6] https://hpc.grnet.gr/en/

Table 1. Micro-services Produced in the ARIS HPC

Service Name	Description
Title Extraction	This micro-service is used to extract the title of the legal document.
Number Extraction	This micro-service is used to extract the number and the year of the legal document.
Text Decomposition	This micro-service is used to decompose the body of the legal document in articles, paragraphs etc.
Extracting Ngrams	This micro-service is used to remove the common words in the legal document and produce the Ngrams.
Extracting Keywords / calculation	This micro-service is used to compare the extracted keywords with the EUROVOC 3rd and 4th level and finally produce the keywords of the legal document.
Correlations Extraction / calculation	This micro-service is used to extract the correlations of the legal document.
Theme extraction / calculation	This micro-service is used to compare the ngrams with the 1st and 2nd level of EUROVOC and find the most common themes.
Publication Date Extraction	This component is used to extract the publication date of the legal document.
Passive Correlations	This component is used to read all produced xmls for Austrian, Greek and European legal documents and update the necessary xmls with the passive correlations.

Even if the research is in pilot phase, the extraction text mining part and passive correlations of law in several pilot countries is working.

7 Conclusion

Automatic end-to-end law analysis conducted by decomposition and semantic annotation, by using the high-performance-computing and government open data can be presented through the processing model and micro-services already developed. and have potential to become a driver of the digital transformation. Even if results could be seen as pilot research on law data from Greece law texts, as open data, they possess the potential to unlock digital innovation and transformation capacity.

This paper presents research baseline for automation of decomposition and annotation with process and service elements developed for utilization on high-performance-computing infrastructure based on government laws open data.

Possible added-value services could be described as follows: i) research through legal corpora, analyzing the alignment of national legislation with EU legislation, ii) comparing national laws iii) analyzing the references to European legislation by national laws, iv) analyzing related laws within the same Member State, v) timeline analysis for all legal acts, vi) visualization of the progress and current status of a specific national or European piece of legislation and vii) sentiment analysis towards new legislation.

Some limitations exist regarding the added-value services that could be offered in the future. There might be a new conceptualised added-value service in the future that could not be supported by the developed and aligned data model. To solve this, we might need to create new data fields or incorporate more fields from the current ontologies. Finally, this solution needs to be tested to more countries to prove its efficiency and effectiveness to different legal environments.

Acknowledgments

Research in this paper is part of larger projects activities in one way connected to use and reuse of Open Data. This research paper is partially supported by Twinning Open Data Operational project (grant agreement No. 857592) and the ManyLaws project (grant agreement. No. INEA/CEF/ICT/A2017/1567 047) and do not necessarily reflect the opinion of the European Union. Both projects received funding from the European Union.

References

Avila-Garzon, C. (2020). Applications, methodologies, and technologies for linked open

data: a systematic literature review. International Journal on Semantic Web and Information Systems (IJSWIS), 16(3), 53-69.

Casanovas P., Palmirani M., Peroni S., van Engers T., & Vitali F. (2016). Semantic web for the legal domain: The next step.Semantic Web, 7(3), 213–227. 10.3233/SW-160224

Charalabidis, Y., Zuiderwijk, A., Alexopoulos, C., Janssen, M., Lampoltshammer, T., & Ferro, E. (2018). Open Data Interoperability. In: The World of Open Data. Public Administration and Information Technology, vol 28. Springer, Cham. https://doi.org/10.1007/978-3-319-90850-2_5

Delen, D., & Crossland, M. D. (2008). Seeding the survey and analysis of research literature with text mining. 34(3), pp.1707-1720.

Erhardt, R. A., Schneider, R., & Blaschke, C. (2006). Status of text-mining techniques applied to biomedical text. Drug discovery today, 11(7-8), pp. 315-325

Fulton C. (2011). Web accessibility, libraries, and the law.Information Technology and Libraries, 30(1), 34–43. 10.6017/ital.v30i1.3043

Gupta, V., & Lehal, G. S. (2009). A survery of text mining techniques and applications. Journal of emerging technologies in web intelligence, 1(1), pp. 60-76.

Hassani, H., Beneki, C., Unger, S., Mazinani, M. T., & Yeganegi, M. R. (2020). Text mining in big data analytics. Big Data and Cognitive Computing, 4(1), 1.

Hrustek, L., Furjan, M. T., & Pihir, I. (2020). ENABLING OPEN DATA PARADIGM FOR BUSINESS IMPROVEMENT. Economic and Social Development: Book of Proceedings, 174-183.

Hrustek, L., Tomičić Furjan, M. & Pihir, I. (2019). Influence of Digital Transformation Drivers on Business Model Creation. In Proceedings of the 42nd International Convention on Information and Communication Technology, Electronics and Microelectronics; Croatian Society for Information and Communication Technology, Electronics and Microelectronics, Opatija, Croatia, 20–24 May 2019; pp. 1509–1513.

Justicia De La Torre, C., Sánchez, D., Blanco, I., & Martín-Bautista, M. J. (2018). Text mining: techniques, applications, and challenges. International journal of uncertainty, fuzziness and knowledge-based systems, 26(04), 553-582.

Lachana, Z., Loutsaris, M. A., Alexopoulos, C., & Charalabidis, Y. (2020). Automated Analysis and Interrelation of Legal Elements Based on Text Mining. International Journal of E-Services and Mobile Applications (IJESMA), 12(2), 79-96. http://doi.org/10.4018/IJESMA.2020040105

Merkl, D., & Schweighofer, E. (1997). En route to data mining in legal text corpora: Clustering, neural computation, and interational treaties. In Database and Expert Systems Applications, 1997. Proceedings., Eighth International Workshop on, IEEE, pp. 465-470.

Nguyen L. M., Bach N. X., & Shimazu A. (2011). Supervised and semi-supervised sequence learning for recognition of requisite part and effectuation part in law sentences.Proceedings of the 9th International Workshop on Finite State Methods and Natural Language Processing, Association for Computational Linguistics, pp. 21-29.

Tomičić Furjan, M., Tomičić-Pupek, K. & Pihir, I. (2020). "Understanding Digital Transformation Initiatives: Case Studies Analysis", *Business Systems Research*11(1), 125-141.

Zeleti, F. A., Ojo, A., & Curry, E. (2016). Exploring the economic value of open government data. Government Information Quarterly, 33(3), 535–551.

Zuiderwijk, A., & Janssen, M. (2014a). Open data policies, their implementation and impact: A framework for comparison. Government Information Quarterly, 31(1), 17–29 https://doi.org/10.1016/j.giq.2013.04.003