Knowledge Graph-based Legal Search over German Court Cases

Ademar Crotti Junior ✉[1], Fabrizio Orlandi[1], Damien Graux[1],
Murhaf Hossari[1], Declan O'Sullivan[1],
Christian Hartz[2], and Christian Dirschl[2]

[1] ADAPT SFI Research Centre, Dublin, Ireland
[2] Wolters Kluwer Deutschland GmbH, Köln & München, Germany
`$firstname.lastname$@{adaptcentre.ie,wolterskluwer.com}`

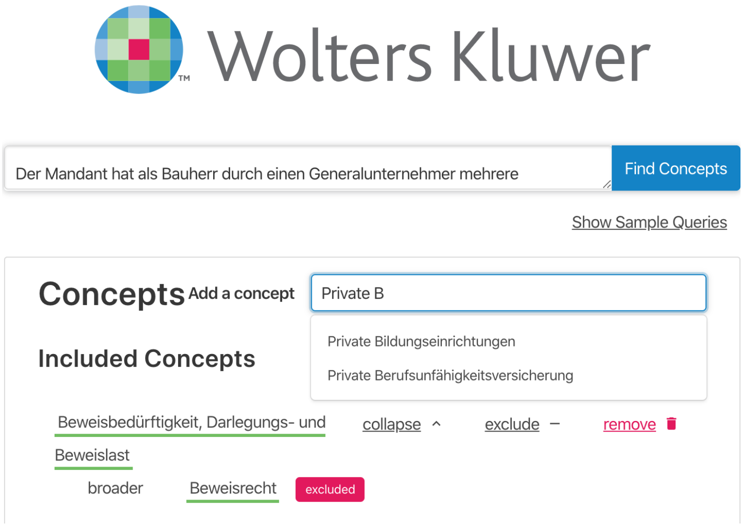
Abstract. The information contained in legal information systems are often accessed through simple keyword interfaces and presented as a simple list of hits. In order to improve search accuracy one may avail of knowledge graphs, where the semantics of the data can be made explicit. This article reports on challenges encountered and achievements made during the development of a knowledge graph-based search engine designed for German court case data at Wolters Kluwer Germany.

1 Introduction

The body of law to which citizens and businesses have to adhere is constantly increasing in volume and complexity [1]. Such information is usually provided by unstructured text within legal documents, for which various solutions have been developed to enable search (based *e.g.* on natural language processing [10] or on structure extraction [8]) and browsing capabilities (using solutions from question-answering systems [3] to multi-facet exploration tool [9]) on large legal corpora. Nonetheless, existing systems usually provide limited keyword-based search interfaces displaying results as a simple list of hits [6]. This makes the process of information retrieval time consuming and inefficient, especially when dealing with large amounts of information [11]. Moreover, the usefulness of such information varies widely and depends on its structure and its representation (see [2] for a classification of 23 legal ontologies). Even though the information is available, users and legal professionals may find the exploration of legal information problematic when interested in specific circumstances or investigating a particular case.

In this context, ADAPT[3] and Wolters Kluwer Germany[4] (WKD) joined forces. Through the years, WKD built a very large dataset containing about a million German language XML-based legal documents. A detailed taxonomy, which is associated with their court case dataset, has also been developed by WKD to

[3] A leading centre developing Linked Data solutions. `https://www.adaptcentre.ie/`
[4] A leading knowledge provider in the legal domain. `https://wolterskluwer.com/`



Der Mandant hat als Bauherr durch einen Generalunternehmer mehrere

Concepts Add a concept: Private B

Private Bildungseinrichtungen

Private Berufsunfähigkeitsversicherung

Included Concepts

Beweisbedürftigkeit, Darlegungs- und Beweislast

broader Beweisrecht excluded

Fig. 1. Search engine user interface.

structure these XML files. The XML format, however, restricts data analysis capabilities by keeping implicit the relationships between concepts and text fragments within legal documents. From these documents, the two partners built together a German-based legal knowledge graph (KG) [4] focusing their efforts on improving the search accuracy and the enrichment opportunities that interlinking features underpinning the Semantic Web approach brings. This paper reports on the search system that was developed leveraging this new KG approach.

2 Enabling semantic search over German court cases

The starting point of this project was a dataset of about one million documents, in German, containing information about legal court cases in Germany that has been built by WKD over the past decades. A taxonomy covering legal concepts was also developed by WKD's experts to provide structure to the information contained in their dataset. The Simple Knowledge Organization System (SKOS)[5] vocabulary is used to describe legal concepts and their relations in this taxonomy. In WKD's taxonomy, each legal concept is represented as an instance of `skos:Concept`, with their relationships being expressed through the properties `skos:narrower` and `skos:broader`.

[5] SKOS is a W3C Recommendation designed for representation of thesauri, classification schemes, taxonomies, subject-heading systems, or any other type of structured controlled vocabulary. `https://www.w3.org/TR/skos-reference/`

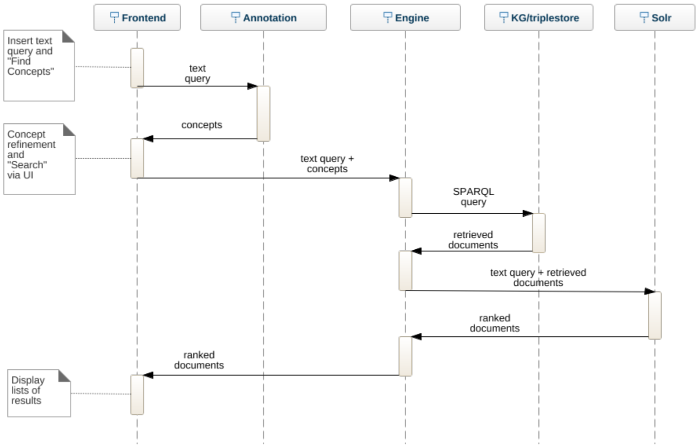


Fig. 2. Sequence diagram for the proposed search engine.

2.1 Challenges

The first challenge has been to annotate and take into account the preexisting structure of such a large, domain-specific, corpus of legal text, in order to build the knowledge graph. Secondly, extensive effort has been invested into capturing the requirements that needed to be featured in the developed engine, working in close collaboration with legal experts. Finally, close collaboration with German end-users has been undertaken in order to evaluate the accuracy of the results achieved and the value created for the business. Keeping in mind that the new search system was bound not to revolutionise the preexisting search experience of the legal experts, typically based on traditional keyword based search.

2.2 Data processing pipeline

The main input of our data pipeline, and the search engine itself, is the WKD legal text corpus. The legal documents were originally stored as XML files with complex schema. The first step of our data pipeline aimed at segmenting these legal documents into smaller and logically coherent fragments, following a specifically developed ontology. Further, an automatic annotation tool was developed to leverage WKD's taxonomy of legal concepts, and applied to the generated fragments. The annotation of taxonomy concepts expresses the connection between legal documents and textual pieces of supporting evidence - the fragments - within and across different documents. These documents are then used to generate two artefacts: the knowledge graph, which is later stored into an RDF triple store; and a Solr[6] index which is later used to support the ranking of documents to be retrieved by the search engine (see Section 2.3). RML [5] mappings

[6] https://lucene.apache.org/solr/

were used for the semantic uplift phase. We chose to use mapping languages for allowing mapping definitions to be expressed separately from the implementations that execute them. The use of declarative mappings facilitate the reuse and maintenance of mappings, where changes in the semantic model or in the input data only require mapping definitions to be updated accordingly, without the need for adjustments in the engine responsible for the generation of the KG. The semantic model used to represent the legal documents from WKD's dataset, as well as the semantic uplift process, have been described in details in [4].

2.3 Search engine

Once the knowledge graph is generated, the search engine operates by transforming a query written in legal German (typically describing court case facts) into taxonomy concepts, before matching them against the structured annotated documents in the KG. A user interface (Figure 1) renders the automatically matched concepts and allows users to manually add, or remove, relevant concepts to the query. The UI supports users in navigating the hierarchy of the legal taxonomy concepts and refining their search query. The identified concepts are then used to query the KG, directly using SPARQL [7]. The SPARQL query matches annotated documents and fragments in the KG with the query concepts. It also ranks the retrieved documents by assigning more weight to those annotated with more specific concepts (*i.e.* narrower concepts in the taxonomy). In order to improve the ranking of the documents, the documents retrieved using SPARQL are then re-ranked using Solr. Fig. 2 illustrates a sequence diagram for the KG based search engine, while the UI is in Fig. 1. As a result, the Semantic Web powered architecture allows experts to explore further a legal knowledge base, offering an interactive and transparent concept-based search as an alternative to the conventional "black-box" approach which relies on pure text-search engines.

3 Conclusions

This paper presents ongoing efforts towards the development of a knowledge graph-based search system for the legal domain. Despite the challenges, mainly due to the complexity of the domain, this novel design provides WKD's end-users with a transparent search experience occurring at the legal concept level. Moreover, the Semantic Web standards and technologies unlock new possibilities for future developments, allowing cutting edge features such as data interlinking and logical inferring of knowledge. Furthermore, the engine could benefit from the latest findings in the area of statistical relational learning, paving the way for new applications. Finally, we believe our approach, not only could save end-users' time, but above all offers their companies new ranges of information and ways of exploration.

Acknowledgments

This paper was supported by Wolters Kluwer Germany, Science Foundation Ireland (Grant 13/RC/2106) and EDGE Marie Skłodowska-Curie grant agreement No. 713567, as part of the ADAPT Centre for Digital Content Technology (`http://www.adaptcentre.ie/`) at Trinity College Dublin.

References

1. Boella, G., Caro, L.D., Humphreys, L., Robaldo, L., Rossi, P., van der Torre, L.: Eunomos, a legal document and knowledge management system for the web to provide relevant, reliable and up-to-date information on the law. Artificial Intelligence and Law **24**(3) (2016)
2. Breuker, J., Casanovas, P., Klein, M.C., Francesconi, E.: The flood, the channels, and the dykes: managing legal information a globalized and digital world. Law, ontologies and the semantic web channeling the legal informational flood pp. 0199–220 (2009)
3. Collarana, D., Heuss, T., Lehmann, J., Lytra, I., Maheshwari, G., Nedelchev, R., Schmidt, T., Trivedi, P.: A question answering system on regulatory documents. In: JURIX. pp. 41–50 (2018)
4. Crotti Junior, A., Orlandi, F., O'Sullivan, D., Dirschl, C., Reul, Q.: Using mapping languages for building legal knowledge graphs from XML files. In: Workshop on Contextualized Knowledge Graphs co-located with the 18th International Semantic Web Conference (ISWC) (2019)
5. Dimou, A., Vander Sande, M., Colpaert, P., Verborgh, R., Mannens, E., Van de Walle, R.: RML: A Generic Language for Integrated RDF Mappings of Heterogeneous Data. In: Proceedings of the 7th LDoW Workshop at WWW (2014)
6. Filtz, E.: Building and processing a knowledge-graph for legal data. In: The Semantic Web. Springer (2017)
7. Harris, S., Seaborne, A., Prud'hommeaux, E.: SPARQL 1.1 query language. W3C recommendation **21**(10) (2013)
8. Koniaris, M., Papastefanatos, G., Vassiliou, Y.: Towards automatic structuring and semantic indexing of legal documents. In: Proceedings of the 20th Pan-Hellenic Conference on Informatics. pp. 1–6 (2016)
9. Lee, S., Kim, P., Seo, D., Kim, J., Lee, J., Jung, H., Dirschl, C.: Multi-faceted navigation of legal documents. In: 2011 International Conference on Internet of Things and 4th International Conference on Cyber, Physical and Social Computing. pp. 537–540. IEEE (2011)
10. Nejad, N.M., Jabat, P., Nedelchev, R., Scerri, S., Graux, D.: Establishing a strong baseline for privacy policy classification. IFIP International Conference on ICT Systems Security and Privacy Protection (2020)
11. Schweighofer, E.: Semantic indexing of legal documents. In: Semantic Processing of Legal Texts: Where the Language of Law Meets the Law of Language, pp. 157–169. Springer (2010)