Introducing Solon: A Semantic Platform for Managing Legal Sources

Marios Koniaris[1](✉), George Papastefanatos[2], Marios Meimaris[2], and Giorgos Alexiou[2]

[1] KDBS Lab, School of ECE,
National Technical University of Athens, Athens, Greece
mkoniari@dblab.ece.ntua.gr
[2] Athena Research Center, Athens, Greece

Abstract. In this paper we introduce *Solon*, a legal document management platform aiming to improve access to legal sources by offering advanced modelling, managing and mining functions. It utilizes a novel method for extracting semantic representations of legal sources from unstructured formats, interlinking and enriching them with advanced classification features. Also, it provides refined search results utilizing the structure and specific features of legal sources, allowing users to connect and explore legal resources according to individual needs.

Keywords: Digital libraries · Information retrieval · Legal informatics

1 Introduction

As a consequence of many open data initiatives, a plethora of publicly available portals and datasets provide legal resources to citizens and legislation stakeholders. However, legal resources are mostly disseminated in a semantically poor, human-readable textual representation [1], mainly PDF, which can not capture the structure and the legal semantics of the data, making it impossible to reuse and establish an interoperability layer among repositories in the Semantic Web.

To address these issues, we introduce *Solon*[1], an advanced system architecture, aiming to assist users locate and retrieve legal and regulatory documents within the exact context of a conceptual reference. It consists of several different components, exposed as REST services. It operates on unstructured legal sources, capturing the internal organisation of the textual structure and the legal semantics, interlinking them based on discovered references and classifying them according to a set of rules. It exploits the semantic representation of legal sources, offering, among others fine-grained search results and enabling users to organize legal information according to individual needs. Recent efforts have also addressed the need for semantic representation of greek legal information [2],

[1] Solon was an Athenian lawmaker, credited with having laid the foundations for Athenian democracy.

J. Kamps et al. (Eds.): TPDL 2017, LNCS 10450, pp. 603–607, 2017.
DOI: 10.1007/978-3-319-67008-9_53

whereas *Solon* has been successfully deployed in a public sector production environment[2]. In this paper, we initially demonstrate the main features offered, we present the main architectural components and discuss future work aspects.

2 Architecture

Requirements and General Characteristics. The main requirements for *Solon*, are focused on (i) support for automatic and manual import of unstructured legal sources from predefined repositories, (ii) automatic structural analysis and semantic representation of legal sources, (iii) automatic discovery and resolution of legal citations, (iv) automatic classification of legal sources based on custom rules, (v) support for manual content curation, (vi) multi criteria and multi faceted search using all metadata identified in documents, and (vii) support for user-defined collections of legal resources around a topic.

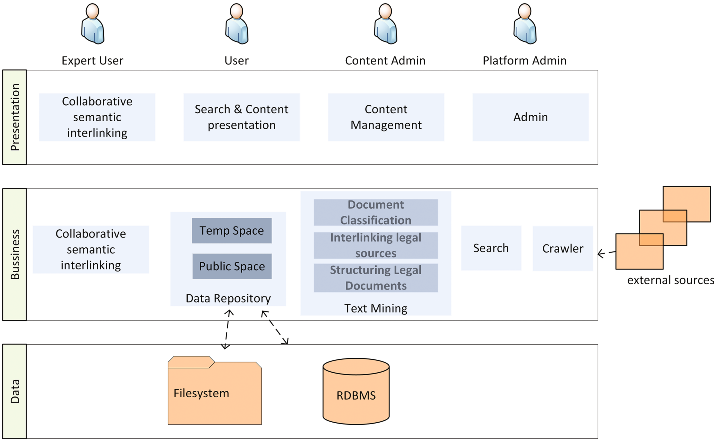

Fig. 1. High-level view of *Solon* logical and conceptual architecture

Architecture. The system's architecture, Fig. 1, is composed of different components exposed to the remaining platform as REST services. The core business layer consists of (i) the Document Repository, which provides functionality for storing and managing complex legal sources, (ii) the Crawler module, which harvests remote information sources as input data, (iii) the Text Mining module, which transforms the Crawler's input into a semantically rich data structure, (iv) the Search module, which is responsible for the efficient indexing and retrieval of legal information, and (v) the Collaborative Semantic Interlinking module. Complementary to the aforementioned modules is the presentation layer, which offers UI and Administration functionality.

[2] http://www.publicrevenue.gr/elib/.

Data Model. *Solon* utilizes the Akoma Ntoso (AKN) schema [3] to model legal documents, an OASIS standard, XML schema for modelling parliamentary, legislative, and judiciary documents. To accommodate for the structure and metadata of Greek legal and administrative documents we provide extensions to the schema in terms of Dublin Core and FRBR vocabularies. In short, our model maps all sections of a legal document, such as articles and paragraphs, into semantically rich legal resources, identified by a URI.

Legal Document Repository. Our legal document repository was build on top of Flexible Extensible Digital Object Repository Architecture, Fedora[3]. Since a legal resource may be accompanied by several files e.g., manifestation format (word, pdf), XML based representation (AKN schema), accompanying material (images), objects are stored in digital containers utilizing a directed acyclic graph of resources where edges represent a parent-child relation. Digital Container management (CRUD) operations are exposed through a RESTful HTTP API, implementing the W3C Linked Data Platform specification.

Crawler. The crawler is based on a distributed architecture, and consists of (a) a Crawler manager and (b) various crawler implementations, extending a common Crawler interface, delivering relevant data to the Crawler manager. The Crawler manager interacts with the Legal Document Repository, ensuring consistent data validation and storage, avoiding also duplication of content. It is also responsible for the periodic scheduling and performance monitoring of all crawling activities within the system.

Text Mining. A pipeline strategy invokes a list of transformers in sequence, that acquire a structured semantic representation of legal sources, interlink legal sources and perform advanced classification based upon tailor made rules that map text to semantic components.

- **Parser.** *Solon* employs the automatic structuring and semantic indexing approach for legal documents, presented in [4].

- **Interlinking legal sources.** Given the abundant use of citations between legal sources, *Solon* employs automated methods for identification and resolution of legal citations, between Greek and EU legal sources, following the methodology utilized in [5]. This component first discovers citations, and then resolves them by creating their respective URIs, following ELI, an EU proposed standard for a European Legislation Identifier. The legal corpus can then be modelled using a graph model as presented in [6].

- **Document Classification.** The Document Classification mechanism is based on a custom developed rule engine following a deterministic approach. Rules are defined through the administration UI module and executed against the legal sources using priorities. Rules can be simple or combined, forming complex chains of operation, acting upon the textual data or metadata of the legal sources.

Collection Management. *Solon* employs a linked data enabled collaborative semantic interlinking mechanism, that allows users to create legal

[3] http://fedorarepository.org.

collections, provide custom semantic annotations to legal resources, and collaborate on shared resources, utilizing the model presented in [7].

Search. *Solon*'s information retrieval component, has been build on top of Solr[4], integrated with our repository component. Since legal documents tend to be quite long, covering multiple topics, we follow structured retrieval techniques, utilizing the legal sources hierarchical structure of nested elements.

3 Evaluation and Demonstrator

Solon has been successfully deployed in a public sector production environment (see footnote 2), under the supervision of the Independent Authority for Public Revenue[5], aiming to provide semantic access to Greek tax legislation. It currently hosts more than 4000 legal and regulatory documents. In the demo, we will showcase the main functionality, addressing the needs of both: (a) the public users e.g., browse the legal knowledge base, search, cite legal resources and (b) the authenticated users e.g., select a legal document and upload it to the repository, curate the structure/metadata, publish it, create legal collections.

4 Conclusion and Future Work

In this paper, we presented Solon a platform suitable for modelling, managing and mining legal sources. As future work, we are investigating the adoption of the recently proposed ELI extension as an OWL ontology and the temporal management of legal sources. Additionally, we plan to employ search result diversification methods, as a means of improving user satisfaction by increasing the variety of information shown to user, based on our previous work where we performed an exhaustive evaluation of several state of the art methods [8].

References

1. Inter-Parliamentary Union: World e-Parliament Report (2016). http://www.ipu.org/pdf/publications/eparl16-en.pdf
2. Chalkidis, I., Nikolaou, C., Soursos, P., Koubarakis, M.: Modeling and querying greek legislation using semantic web technologies. In: Blomqvist, E., Maynard, D., Gangemi, A., Hoekstra, R., Hitzler, P., Hartig, O. (eds.) ESWC 2017. LNCS, vol. 10249, pp. 591–606. Springer, Cham (2017). doi:10.1007/978-3-319-58068-5_36
3. Barabucci, G., Cervone, L., Palmirani, M., Peroni, S., Vitali, F.: Multi-layer markup and ontological structures in Akoma Ntoso. In: Casanovas, P., Pagallo, U., Sartor, G., Ajani, G. (eds.) AICOL -2009. LNCS, vol. 6237, pp. 133–149. Springer, Heidelberg (2010). doi:10.1007/978-3-642-16524-5_9
4. 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, PCI 2016. ACM (2016)

[4] http://lucene.apache.org/solr/.

[5] http://www.aade.gr/, formerly known as General Secretariat of Public Revenue.

5. Opijnen, M.v., Verwer, N., Meijer, J.: Beyond the experiment: the extendable legal link extractor (2015). https://ssrn.com/abstract=2626521
6. Koniaris, M., Anagnostopoulos, I., Vassiliou, Y.: Network analysis in the legal domain: A complex model for european union legal sources. In: Physics and Society, Cornell University Library, arxiv (2015). http://arXiv.org/abs/1501.05237
7. Meimaris, M., Alexiou, G., Papastefanatos, G.: LinkZoo: a linked data platform for collaborative management of heterogeneous resources. In: Presutti, V., Blomqvist, E., Troncy, R., Sack, H., Papadakis, I., Tordai, A. (eds.) ESWC 2014. LNCS, vol. 8798, pp. 407–412. Springer, Cham (2014). doi:10.1007/978-3-319-11955-7_57
8. Koniaris, M., Anagnostopoulos, I., Vassiliou, Y.: Evaluation of diversification techniques for legal information retrieval. Algorithms **10**(1), 22 (2017)