algorithms

MDPI

Article

Solon: A Holistic Approach for Modelling, Managing and Mining Legal Sources

Marios Koniaris [1,*], George Papastefanatos [2] and Ioannis Anagnostopoulos [1]

[1] Department of Computer Science and Biomedical Informatics, School of Sciences, University of Thessaly, 35131 Lamia, Greece; janag@dib.uth.gr

[2] Athena Research Center, 151 25 Marousi, Greece; gpapas@imis.athena-innovation.gr

* Correspondence: mkoniaris@uth.gr

Received: 31 October 2018; Accepted: 30 November 2018; Published: 3 December 2018

Abstract: Recently there has been an exponential growth of the number of publicly available legal resources. Portals allowing users to search legal documents, through keyword queries, are now widespread. However, legal documents are mainly stored and offered in different sources and formats that do not facilitate semantic machine-readable techniques, thus making difficult for legal stakeholders to acquire, modify or interlink legal knowledge. In this paper, we describe Solon, a legal document management platform. It offers advanced modelling, managing and mining functions over legal sources, so as to facilitate access to legal knowledge. It utilizes a novel method for extracting semantic representations of legal sources from unstructured formats, such as PDF and HTML text files, interlinking and enhancing them with classification features. At the same time, utilizing the structure and specific features of legal sources, it provides refined search results. Finally, it allows users to connect and explore legal resources according to their individual needs. To demonstrate the applicability and usefulness of our approach, Solon has been successfully deployed in a public sector production environment, making Greek tax legislation easily accessible to the public. Opening up legislation in this way will help increase transparency and make governments more accountable to citizens.

Keywords: digital libraries; information retrieval; legal Informatics; linked open data; parliament data; open government data

1. Introduction

Public legal information, a part of the common heritage of humanity [1], is a large collection of different normative documents, which keeps growing and changing with time. Nowadays, as a consequence of initiatives for open data and especially open government data, there is a huge increase in the number of portals providing legal documents to interested parties. Legal sources, to which only a qualified audience had access, are now freely available. Although this open access is usually supported by services such as thematic navigation and keyword search, legal sources are primarily stored and disseminated in a semantically poor text representation that does not reflect the structure and semantics of legal data.

Legal sources availability in a structured and standard format, the existence of a common reference vocabulary between the systems involved, is a necessary prerequisite for effective legal document dissemination. Towards this end, national and international initiatives have proposed standards for the semantic representation of legal sources [2–4]. However, as the World e-Parliament survey 2016 [5] reports only 26% of parliaments distribute documents in XML format and 49% operate document management systems. Legal data is mainly offered to the end user in a friendly and human readable manifestation, primary presentation oriented e.g., PDF. In this paper, following functional requirements

for bibliographic records (FRBR) terminology [6], we do consider a manifestation to be the physical embodiment of an expression of a work, any different realisation of an expression i.e., paper format, digital format, etc. The use of such proprietary and unstructured format, makes it impossible to establish an interoperability layer among the different sources of information, to allow reuse and interconnection with repositories in the Semantic Web, overall hampering opportunities for openness and transparency.

At the same time, the recent abundance of Open Legal Data poses extra overhead for both citizens and legal professionals to find useful and relevant legal sources. Consider, for example, a compliance manager researching whether a given company is conducting its business in full compliance with national, and potentially international, laws and regulations specific to a particular industry. He/she has to iteratively browse an enormous number of legal documents, obtained iteratively through keyword queries, most probably in PDF format, selecting, through knowledge and experience, relative documents in order to comprehend conditions and grounds. A legal information system providing refined search results, semantically represented, interlinked and enhanced with classification features is intuitively more informative and helpful that a set of relevant PDF files.

Furthermore, as the legal doctrine increases in volume and complexity, finding a relevant norm may be a challenging task even for experts. To alleviate the data overload problem, modern legal information systems are not only intended to provide advanced search tools for users, but also to organize the legal order, to monitor the impact of the new rules on the regulation, to manage documents life-cycle and different versions chronology [7].

In this paper, we present Solon, a legal document management platform. Solon was an Athenian lawmaker remembered particularly for his efforts to legislate against political, economic, and moral decline in archaic Athens. He is credited with having laid the foundations for Athenian democracy.

Solon's main goal is to assist users reference and retrieve legal and regulatory documents within the exact context of a conceptual reference. It consists of several different components, exposed as Representational State Transfer (REST) services. It is an advanced legal document management platform, operating on legal sources that are automatically discovered and collected from authoritative portals with the help of Web Harvesters/Crawlers on a scheduled basis. It transforms legal sources from presentation oriented, unstructured formats to a suitable for modelling legal sources format, capturing the internal organization of the textual structure and the legal semantics, and classifying them according to a set of rules and interlinking them based on discovered references. Solon utilizes the semantic representation of legal sources, offering, among others fine-grained search results and enabling users to organize legal information according to individual needs.

Our approach, was initially conceived in [8] where we showcased a mechanism allowing for the automatic structuring and semantic indexing of legal documents and primitively was demonstrated in [9], where we showcased its main functionality. Over the years, the platform we initially envisaged has emerged to a mature level. Solon has been successfully deployed in a production environment for the Greek public sector, providing semantic access to Greek tax legislation.

This paper presents in detail the platform functionality and its architectural components, providing a holistic approach to modelling, managing and mining legal sources. We first present the functional requirements that drove the design and implementation of the platform, focusing on the data and meta-data models we utilize, the legal document repository that offers archival of legal documents, the harvesting process as to acquire legal documents from public portals, the Domain Specific Language (DSL) we created as to facilitate the transformation of unstructured raw documents to a structured semantic representation of legal sources, the process utilized for discovering legal citations and interlinking legal sources, the retrieval techniques we adopted for satisfying users information needs and finally the mechanism we employed allowing for a collaborative semantic interlinking of legal sources. Finally, complementing our presentation we provide a real use case example of the platform materialization in a public sector operated production environment, as an evaluation study for the feasibility and applicability of our approach.

2. Architecture

In this Section, we present Solon's main features, the data model utilized for legal sources and afterwards describe the platform's architecture.

2.1. General Characteristics

The main requirements for Solon, are focused on:

- Support for automatic and manual import of unstructured documents from predefined legal sources.
  Legal sources, i.e., regulations, case law and administrative acts, are usually authored by government agencies and disseminated through authoritative government websites. Therefore a reliable mechanism should provide input data to the platform on a scheduled basis.
- Automatic structural analysis and semantic representation of textual data and metadata.
  Input legal sources are usually distributed in unstructured text, often proprietary, formats, thus, it is preferable to transform them to a suitable for modelling legal sources format, capturing the legal semantics.
- Automatic discovery and resolution of legal citations, for each respective structural unit.
  Typically, legal documents refer to authoritative documents and sources, thus forming a network of interconnected legal documents. A precondition for the inter-linkage of legal documents and the Legal Linked Open Data formulation is the discovery and resolution of legal citations.
- Automatic classification of legal sources based on custom rules.
  Since the amount of legal sources available will always outweigh the time one actually has to read them, a popular approach to manage the information overload is to assign legal sources to one or more classes or categories, either manually or algorithmically.
- Support for manual curation of the automatically discovered, structured and semantically enriched content.
  While information extraction processes may achieve high levels of accuracy as to replace manual curation, the later is an importance step towards the completeness of the resulting legal knowledge base. Apart from that, there is also a zero tolerance for errors regarding the provision of legal documents.
- Support for multi criteria and multi faceted search using all metadata identified in documents.
  The aim of any information retrieval system is to deliver content that can precisely match and satisfy the users information need. With the plethora of intended users categories, ranging from novices to legal experts, distinctive information retrieval techniques, utilizing heterogeneous dimensions over the legal information sources, should be supported, so as to precisely match information requests.
- Support for structured content retrieval.
  A Legal Document cannot be processed as a general purpose text document; instead it exhibits strict semantics for each part of it, having a hierarchical structure of nested elements (e.g., articles that contain paragraphs, etc.) and covering multiple legal topics. Thus, a more precise approach than simply indexing and retrieving documents as single units, is needed, in order to provide more accurate and relevant results.
- Support for user-defined collections of legal resources around a topic.
  Integrating end users knowledge into the legal sources, is expected to assist them organize and explore legal sources in various meaningful ways.

2.2. Data Model

Various guidelines for good legislative drafting, both at National and E.U. level [10], have established common formats, which most legal documents abide by. In this work the term "format"

is used as to include topics such as structure, numbering, preferred word use and grammar and template sentences.

In a simplified view, a legislative legal document has the following structure:

- *Introductory part*. The introductory part contains information enabling us to identify the type of document as well as most of its publication and manifestation metadata, e.g., title, number, issuing authority, etc.
- *Text body*. The text body is the main part of the document and it may be structured differently depending on the legal document type. It usually follows a well-defined hierarchical layout of text blocks (*legal blocks*), where high level parts of a document, such as the chapter of a law, contain other blocks, such as articles and paragraphs.
- *End part*. The end part contains closing formulas, the date, and the signatures.

A visual illustration of the aforementioned structure, with manually annotated structural parts and metadata values, for a Greek law is shown in Figure 1.

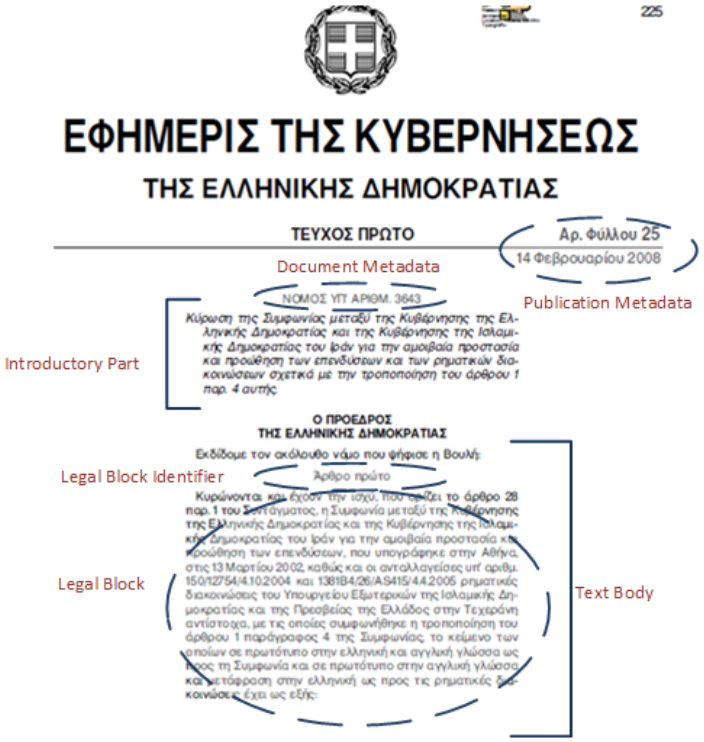

Figure 1. Overview of the structure of a legal document, with manually annotated structural parts [8].

Figure 2 shows the basic elements of a legal act, according to the European Union Interinstitutional Style Guide [10], an ongoing effort to standardize formats and to harmonize the presentation of publications. These guidelines, contain uniform stylistic rules and conventions which must be used by all the institutions, bodies, offices and agencies of the European Union. Their application is obligatory for all those involved in document production (paper or electronic), within institutions and bodies of the European Union. Depending on the complexity of the text, elements such as parts, titles, chapters or sections may be used in the preamble, enacting terms and annexes.

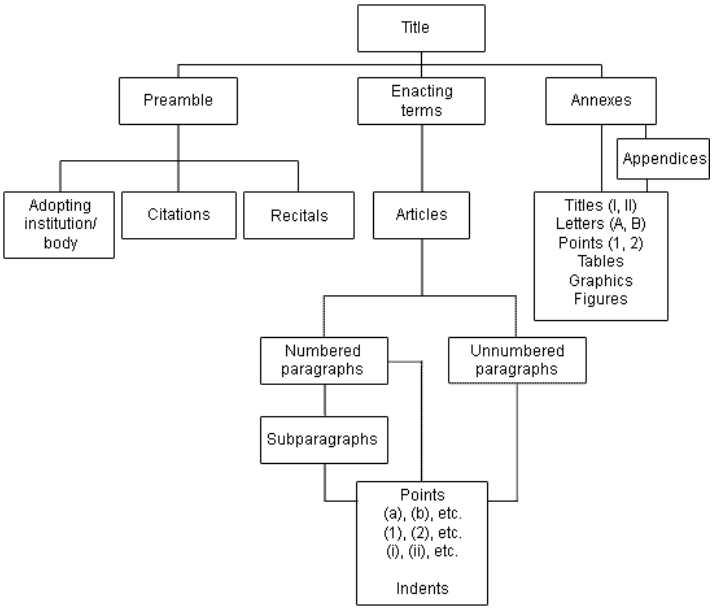

Figure 2. Basic elements of a legal act, according to European Union Interinstitutional Style Guide [10].

Solon models legal documents utilizing the LegalDocML schema, a standardisation of Akoma Ntoso (Architecture for Knowledge-Oriented Management of African Normative Texts using Open Standards and Ontology) schema [4]. LegalDocML is an Organization for the Advancement of Structured Information Standards (OASIS) standard XML schema [11], suitable for modelling legislative, parliamentary and judiciary documents. The LegalDocML schema allows the structural and semantic components of legal sources to be accessible by machine-driven processes, thereby providing support for the creation of high-quality legislative information services and improving accountability and efficiency in legislative, parliamentary and judicial contexts.

To comply with Greek legal documents i.e., ministerial decisions, laws, regulatory acts of the Council of minister, presidential decrees and administrative acts, we initially analyzed Greek legal sources. Afterwards, we created mapping rules between all the legal blocks (e.g., title, chapters, articles, passages, paragraphs) and their corresponding elements in LegalDocML schema. At the metadata level we also created mapping rules between the LegalDocML schema annotating mechanisms i.e., identification, publication, classification, lifecycle, workflow, temporalData, references and presentation metadata categories and the available/potential metadata types found in Greek legal sources. Finally, for local and proprietary metadata of the platfrom we utilize the *proprietary* LegalDocML schema metadata format, based on Dublin Core and FOAF properties vocabularies

Figure 3 provides a visual part of metadata for Greek Law 4172, encodeded in the LegalDocML schema, highlighting the usage of the Functional Requirements for Bibliographic Records (FRBR) and Dublin Core vocabularies among others.

```
        <FRBRdate date="2016-07-15" name="INSERT"/>
        <FRBRauthor as="#editor" href="#HB-bot"/>
        <FRBRlanguage language="gr"/>
    </FRBRExpression>
    <FRBRManifestation>
        <FRBRthis value="/gr/act/2013/4172/4172_2013.xml"/>
        <FRBRuri value="/gr/act/2013/4172/4172_2013.akn"/>
        <FRBRdate date="2016-07-15" name="XMLConversion"/>
        <FRBRauthor as="#editor" href="#HB-bot"/>
        <componentInfo/>
        <FRBRformat value="text/xml"/>
    </FRBRManifestation>
</identification>
<publication date="2013-07-23" number="167 A' 2013" name="et" showAs="ΕΘΝΙΚΟ ΤΥΠΟΓΡΑΦΕΙΟ"/>
<classification source="#main"/>
<classification source="#signer">
    <keyword dictionary="#signer" showAs="ΠΡΟΕΔΡΟΣ ΤΗΣ ΕΛΛΗΝΙΚΗΣ ΔΗΜΟΚΡΑΤΙΑΣ" value="ppd"/>
</classification>
<lifecycle source="#HB-bot">
    <eventRef type="generation" date="2013-07-23"/>
</lifecycle>
<references source="#HB-bot">
    <original href="/gr/act/2013/4172/files/4172_2013.pdf" showAs="Πρωτότυπο"/>
    <TLCOrganization href="#" showAs="ΕΘΝΙΚΟ ΤΥΠΟΓΡΑΦΕΙΟ" eId="et"/>
</references>
<proprietary source="#HB-bot">
    <DCTerms xmlns:ns2="http://docs.oasis-open.org/legaldocml/ns/akn/3.0/CSD13" xmlns:dc="h
        <dc:type>ΝΟΜΟΣ</dc:type>
        <dc:title>Φορολογία εισοδήματος, επείγοντα μέτρα εφαρμογής του ν. 4046/2012, του ν.
        <dc:creator>Ο ΠΡΟΕΔΡΟΣ ΤΗΣ ΕΛΛΗΝΙΚΗΣ ΔΗΜΟΚΡΑΤΙΑΣ</dc:creator>
```

Figure 3. Part of metadata for law 4172, year 2013, annotated following LegalDocML schema.

Overall, in accordance with the five levels of compliance to the LegalDocML schema, our data model conforms to level 3. That is we follow: (a) the document structure defined in the LegalDocML specification (e.g., preface, preamble, body, conclusion, annexes) for the entire document, (b) the naming convention of URI/IRI (FRBR metadata) and IDs defined in the LegalDocML Naming Convention and (c) the basic metadata FRBR, publication, normative reference.

Furthermore, in alignment with the European Legislation Identifier (ELI) standard and OWL ontology, a EU proposed standard providing a method to uniquely identify and access national and European legislation [12], our approach offers the minimum set of metadata required by ELI and assigns a URI at each different legal block modelled in LegalDocML. Hence, the mark-up of each structural unit of the legal sources complies with the ELI standard, facilitating the precise linkage of legal citations for each respective structural unit. As shown in Figure 1, article 1 of the main part of the act with no 3643, published in 2008 by the Greek Parliament, is identified by the'. . . / gr/ act/ 2008/ 3643/ main/ art/ 1/' URI.

Last in sequence but not least in importance, as a framework for sharing legal data in a way that will enable maximum use and reuse we pursue FAIR data principles into practice. FAIR principles are one of the main pillars of Open science, providing the guidelines for delivering Findable, Accessible, Interoperable and Reusable datasets and services for researchers. FAIR data principles [13] provide a set of criteria, in the form of questions and requirements, that the publication of (research) data must fulfil, such as whether a dataset contains rich metadata and persistent identifiers for the resources, whether it exhibits a standard vocabulary, or follows an Open Access license or records provenance and versioning of resources. In Solon, we follow these principles by employing a standard vocabulary for the description of the legal documents, i.e., LegalDocML, which is an open and well-established schema used in the legal informatics community as well as a standard URI scheme, i.e., ELI, for providing persistent identifiers to resources, legal documents and parts of it, as well as for versioning purposes. We also provide all resources under a CCBY4.0 license. Other aspects of FAIRNess will be also considered in future deployments.

2.3. Architecture

Solon is based on a set of software components integrated through well-defined APIs, that communicate through REST HTTP interfaces and can be utilized not only as parts of the overall architecture, but also as individual services. A high-level view of Solon logical and conceptual architecture is provided in Figure 4.

Storing and managing complex legal sources functionality is offered by the Document Repository. The Crawler module harvests remote information sources as input data, which afterwards the Text Mining module transforms to a semantically rich data structure. Efficient indexing and retrieval of legal information is performed by the Search module. The Collaborative Semantic Interlinking module allows users to connect, organize and explore legal resources according to individual needs. Finally, synergetic to the above modules is the Administration module assisting manual curation of content and general administrative functionality.

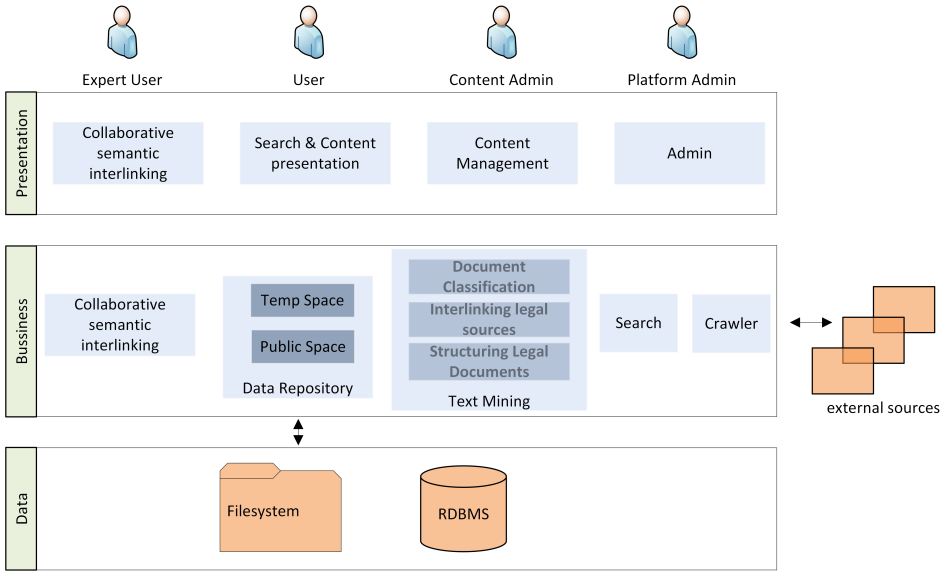

Figure 4. High-level view of Solon logical and conceptual architecture [8].

3. Solon in Depth

In this section, we present the main architectural components and functionalities of Solon, such as the document repository, the harvesting and data mining components, the semantic interlinking as well as the retrieval techniques we adopted for satisfying users information needs. Supplementary technical material is provided at https://github.com/mkoniari/Solon/.

Solon architecture and implementation was performed with the essential prerequisite of employing only open source software methodologies. As such, we do believe that the open-based software approaches, adequately and sufficiently cover Solon's main requirements, as presented in Section 2.1. Furthermore, different approaches/tools will eventually affect system sub-components in terms of implementation details and the platform in terms of system performance and adaptability in different legislation environments. Thus, the modular architecture chosen will allow for changes in various sub-components with the minimal effort/disruption. Regarding the tools chosen, Fedora is one of the main open-source repositories built around an active community and which is widely used in digital libraries. It offers numerous capabilities for metadata management, versioning, support of different data models (RDF, XML, etc.) among others. Apache Solr is also one of the most popular open sourced tools used for keyword indexing and search and it allows tight integration with the Fedora repository.

3.1. Legal Document Repository

The main functionality of the document repository is to offer reliable, trustworthy and persistent archival of digital content, emphasizing on availability and use of the information, in accordance with modern web standards. Solon's legal document repository was build on top of Flexible Extensible Digital Object Repository Architecture (Fedora, http://fedorarepository.org), an open-source, Java based, modular repository framework [14].

Fedora supports flexible and extensible digital objects, which are containers for metadata, one or more representations of the content and relationships to other information resources. Also, it is implemented on the basis of web services, that offer full programmatic management of digital objects as well and search and access to multiple representations of objects.

Figure 5 illustrates the structure of a legal document/object in the repository and the correspondence of REST-based access requests to the object.

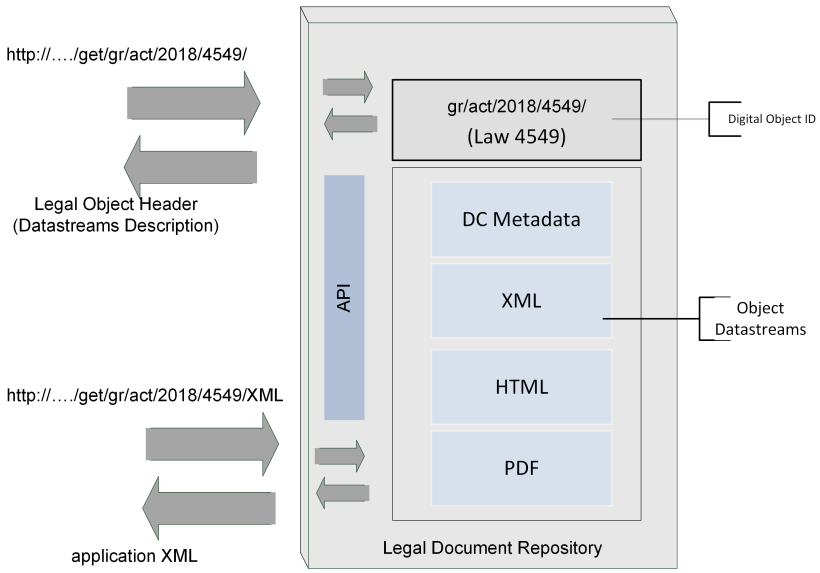

Figure 5. REST based access requests to a Legal document and its components in the Legal Document Repository.

The repository architecture provides the foundation for storing and managing complex digital objects, as well as the relations between them in RDF based relationship model. The repository was build as to assist the semantic layer providing capabilities that make system usage easier, providing powerful facilities for sharing and accessing legal knowledge and a uniform way to classify objects and access the data model.

Digital objects stored in the repository, legal and administrative documents, not only follow a hierarchical data model, but also consist of several documents/files e.g., manifestation format (word, pdf), XML based representation (LegalDocML schema), accompanying material (images), etc. They are stored in a directed acyclic graph of resources where edges represent a parent-child relation. Digital Container management, i.e., create, read, update and delete (CRUD) operations, import and export functionality has been build and exposed through RESTful HTTP API, implementing the W3C Linked Data Platform (LDP) specification [15].

3.2. Crawler—Harvester

Crawling/harvesting legal documents from public portals is a fixed activity, dependent however on the concrete characteristics of the information source to be crawled. The main reason behind this particularity, is the lack of uniformity in terms of interacting with the data provider or the actual content provided. Legal resources are usually provided in unstructured manner, through various

formats. Also accompanying metadata may or not exist, in a variety of formats. Furthermore, a single legal document may consist of several separately offered documents, that should be downloaded together and eventually joined in a bundle e.g., textual representation and images. Last but not least, legal documents may not be directly accessible through an API call, but offered through search results, of varying structure for each information provider, that the crawler has to parse and evaluate in order to find relevant content.

Based upon the aforementioned aspects, aiming also to ensure compatibility and centralized data storage, a distributed architecture was chosen for the design of the harvesting subsystem. It consists of (a) a crawler manager and (b) various crawler implementations, extending a common Crawler interface, interacting with the manager. Figure 6 illustrates, the distributed architecture, which also allows for the easy deployment of new crawlers.

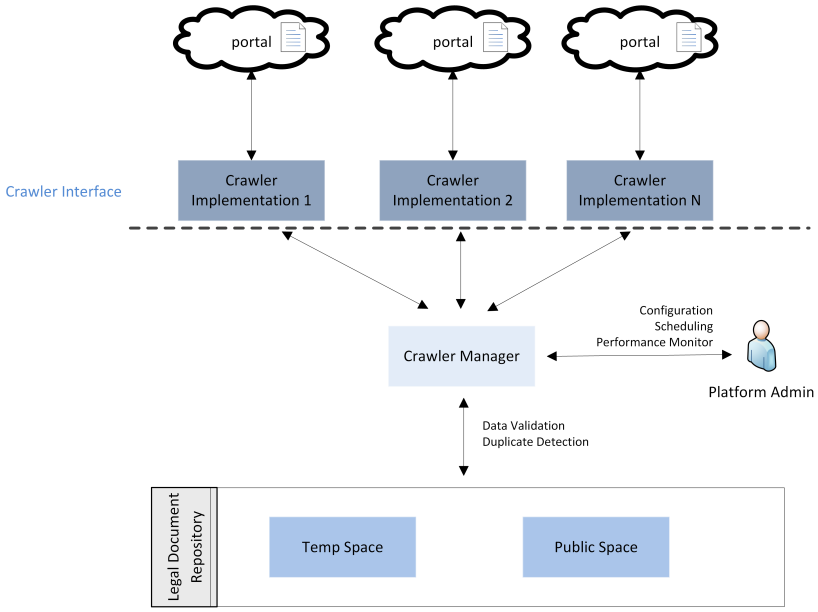

Figure 6. Crawler Architecture. A Crawler manager, interacting with the Legal Document Repository, coordinates discrete crawler implementations, extending a common Crawler interface, that harvest specific information providers.

The crawler manager is mainly responsible for interacting with the Legal Document Repository, ensuring consistent data validation and storage of the acquired data, avoiding also duplication of content. Furthermore, functionality for periodic scheduling and performance monitoring of all crawling activities within the system is offered by the crawler manager, thus, ensuring smooth performance of the system.

Each concrete crawler implementation, build upon the unique characteristics of the information provider to be utilized, is responsible to find and download new or updated data and to deliver the relevant data to the crawler manager. Since crawlers interact with specific information providers, independently of each other, they may be located on the same or in different VMs, providing for load balancing capabilities.

Upon the successful validation, by the crawler manager, of data available to the system, data is then stored in a temporary workspace of the repository, initiating the text mining procedures.

3.3. Text Mining

A pipeline strategy, coordinated by a controller, invokes a list of transformers in sequence, passing the output of a transformer as input to the next transformer, enabling the chaining of multiple transformers to perform more complex tasks. Major tasks performed involve acquiring

a structured semantic representation of legal sources, interlinking legal sources and performing advanced classification based upon tailor made rules.

Figure 7 provides an overview of our Text Mining strategy, consisting of (a) Extracting structural and semantic info from legal documents, (b) interlinking legal sources and (c) document classification.

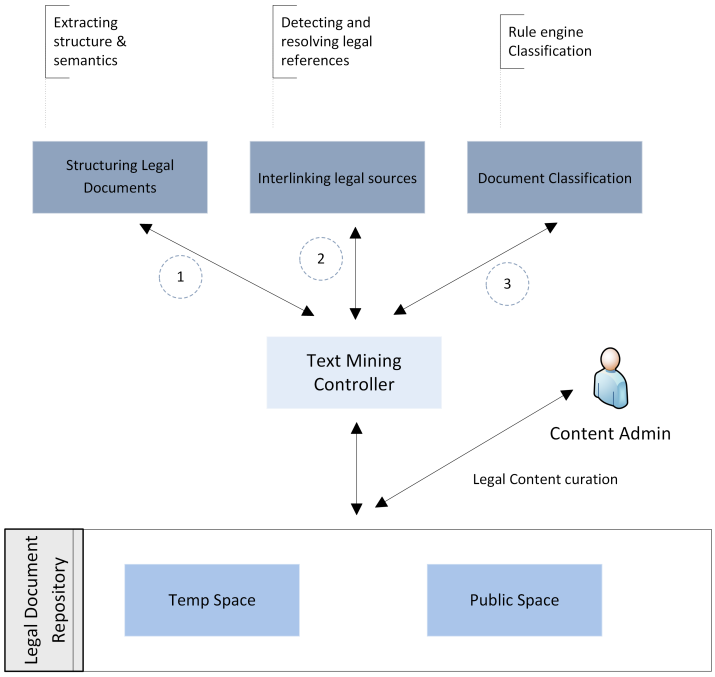

Figure 7. Text Mining Process. Unstructured documents are firstly transformed to a suitable for modelling legal sources format, capturing the legal semantics. Afterwards legal citations are discovered and resolved and finally documents are assigned to one or more classes or categories.

3.3.1. Structuring Legal Documents

As legal documents are usually distributed, through various technical formats, in a unstructured, presentation oriented manifestation, a process is needed for extracting a machine-readable, semantic representation of them. In this Section, we provide an overview of our approach, on Automatic Structuring and Semantic Indexing of Legal Documents. Our method identifies, with high accuracy, textual and metadata elements of unstructured documents and provides a semantic translation of them to the utilized data model, as presented in Section 2.2.

A DSL Languange for Legal Documents

Domain-specific modeling [16], a software engineering methodology for designing and developing systems directly from the domain-specific models, offers tailor-made solutions to problems in a particular domain. The structure of a legal document, as described in Section 2.2, can be expressed by a domain-specific model and thus extracted by employing syntax rules (DSL) [17,18]. Hence, document structure analysis methods provide us with the proper framework in order to describe syntactic structure of documents with an abstract document model. When combined with a method to implement a document structure parser by a combination of syntactic parsers, it equips us with a parser that has high generality and extensibility.

A context-free grammar (GFG), described in Extended Backus–Naur Form (EBNF is a formal way of representing context free grammars), G is a tuple $G = (V, T, P, S)$ where

- V is the (finite) set of variables (or nonterminals or syntactic categories). Each variable represents a language, i.e., a set of strings
- T is a finite set of terminals, i.e., the symbols that form the strings of the language being defined

- P is a set of production rules that represent the recursive definition of the language.
- S is the start symbol that represents the language being defined. Other variables represent auxiliary classes of strings that are used to help define the language of the start symbol.

Initially, we expressed legal documents structure in the form of a set of syntactic rules, i.e., a domain-specific language (DSL) for legal documents, used for generating a syntactic document parser [17,18].

In this way, a syntax rule for the top level document structure of the legal document presented in Figure 1 can be described as shown in Figure 8. Preface and preamble compose the introductory part and body the text part. Nonterminals such as body and conclusions are defined separately.

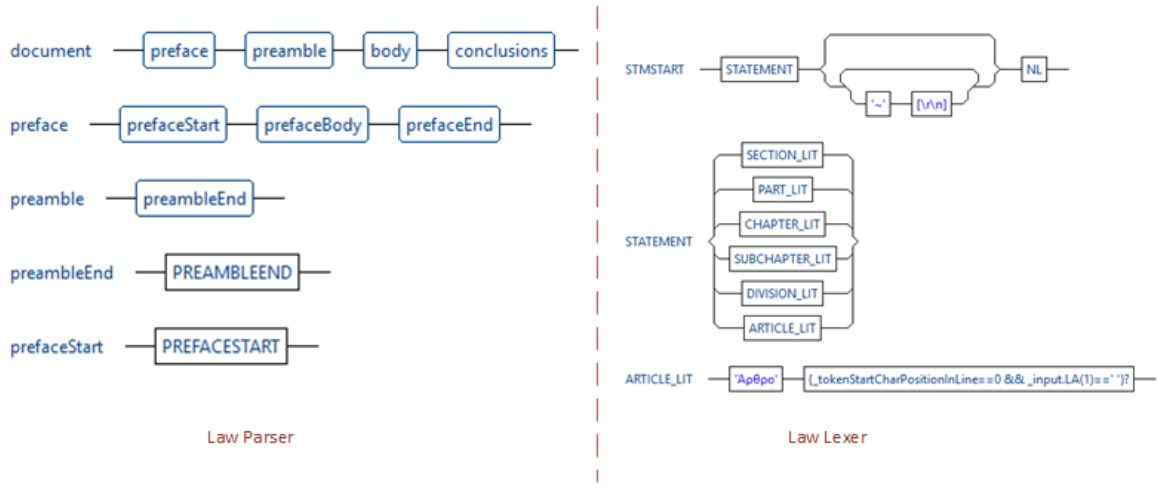

Figure 8. Overview of Legal Sources Structure Parser [8].

Figure 8 provides a segmented railway overview of parser rules (left pane) and lexer rules (right pane) for laws. As a rule of thumb legal block definition is kept in the parser (with lower-case symbols) and string declaration in the lexer (with upper-case symbols).

Next, we employ parser generators for implementing the parsing mechanism, based on the defined GFG and the set of syntax rules defined. ANother Tool for Language Recognition (ANTLR), a language framework for constructing recognizers, interpreters, compilers and translators from grammatical descriptions containing actions in a variety of target languages, is among the most powerful and popular parser generators. ANTLR accepts as input any context-free grammar that does not contain indirect or hidden left-recursion and generates a lexer and a recursive-descent parser that uses an ALL(*) production prediction function [19]. Experiments presented in [19] have shown that ALL(*) outperforms general (Java) parsers by orders of magnitude, exhibiting linear time and space behavior for various languages.

The lexer, parser, and tree walker are generated by ANTLR from the corresponding ANTLR grammars. In this way, our method has high generality and extensibility. Among others the advantages of our methodology are:

- we use a powerful abstraction that seperates programming from legal domain knowledge
- we can easily extend our grammars to provide for more legal documents e.g., judgements
- it is easier to maintain/evolve the procedure
- our layered implementation allows to easily adopt to new schemas/standards.

For the identification of the syntax rules, we heavily rely on domain knowledge from the legal experts. They provided us with feedback on the structural parts and their relationships (nesting, succession, etc.) within legal sources. Combining different syntax rules together enables to identify the type of the document and the different text block elements it contains. Thus, document structure

analysis methods provide us with the proper framework in order to describe syntactic structure of documents with an abstract document model. Our method exhibits high generality and extensibility, by utilizing powerful abstraction that separates programming from legal domain knowledge, enabling us to easily extend our grammar addressing specialized cases of legal or administrative documents. Despite the encountered difficulties and shortcomings in the identification process, results of our approach are very positive. Overall, our implementation is highly extensible and achieves high accuracy for a variety of legal and administrative documents, as noted in the accuracy and scalability analysis presented in [8].

Parsing Process

The main steps of our approach are: (a) identify the structure of the legal documents, (b) identify legal documents metadata, and (c) validate produced files against the selected schema.

As a pre-processing step, in our method, all legal documents are converted into plain text files. Notably, this step does have the downside of discarding valuable style and layout information found in the presentation document format (pdf/word). Several open source content analysis frameworks can efficiently and effectively perform this task, e.g., Apache Tika. Also, scanned text stored as an image in the supplied document can be adequately handled by an OCR process. While an optical character recognition or OCR system transforms a scanned image into text, it cannot offer 100% accuracy. Although the OCR process accuracy has improved notably over the past decades, as a result of more elaborated algorithms, it may still lose valuable information.

As such, we do believe that manual content curation process is an importance step towards the completeness of the resulting legal knowledge base, especially in cases where there is limited tolerance for errors. Solon utilizes the Apache Tika framework for converting PDF files into plain text. Tesseract OCR can be seemingly integrated with Apache Tika as to handle scanned text stored in PDF images. Solon architecture and implementation provides two models of operation regarding scanned images: (a) having them stored as images and accompany the resulting XML file and (b) enriching the text file with content acquired from the OCR process. However, in the particular use case presented in this paper, Section 4, it was requested by the program owner that Solon operates without the OCR sub-process.

Figure 9 provides a visual overview of our parsing process. The input document is firstly transformed into a token stream by the lexer. In the stream, tokens are associated with a regular expression that define the set of possible character sequences used to form the token. Afterwards, the parser, parses syntactic structure of the token stream creating an Abstract Syntax Tree (AST). AST is a tree data structure representing the abstract syntactic structure of the input stream, which is processed by a tree walker that generates the in-memory document model. Afterwards, a transformer formats the in-memory document model to the chosen legal schema (Section 2.2). Finally, upon successful completion of the semantic checker/validator, the output document is serialized.

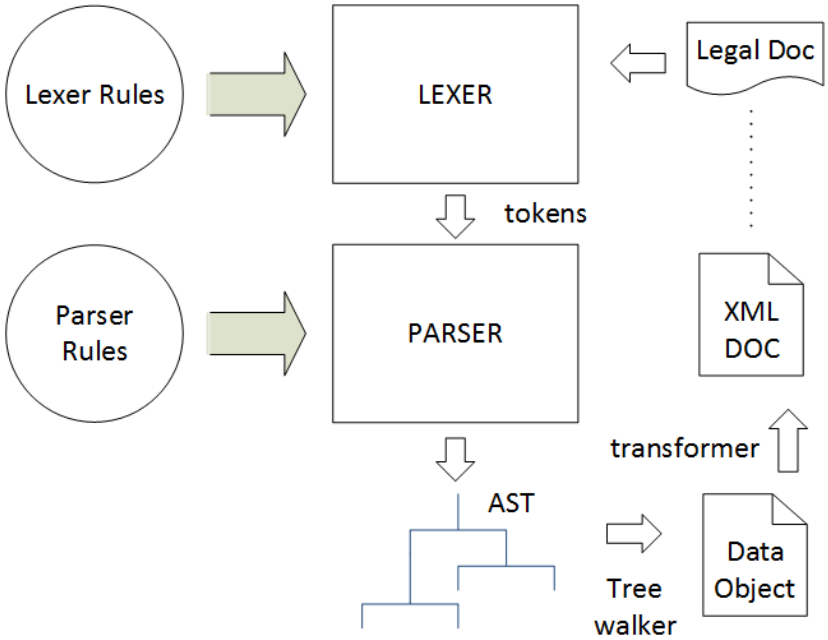

Figure 9. Overview of Legal Document Structure Parser [8].

Figure 10 provides a visual overview of our parsing methodology. It follows a pipeline strategy, utilizing a top-down approach, that can be summarized into the following five steps:

(a) *Document Type Identification*. A first step in our methodology is to properly identify the type of legal document, since the type of the document defines both the document structure, available metadata and the internal semantic organization of the resulting document. Thus, identifying the type of legal document is a fundamental task to effectively model legal documents. This is achieved by scanning the beginning of the text for several predefined keywords that effectively distinguish document types.

(b) *Structural Analysis*. Having identified the type of legal document, we distinguish legal blocks in the document (e.g., front matter, body, conclusions, annexes). Structural Analysis differs between various document types as predefined in the corresponding grammar and implemented in the appropriate parser.

(c) *Legal Blocks Isolation*. Legal blocks, previously identified in the structural analysis step, are iteratively broken down, into distinct elements, according to the corresponding grammar. Within this step, we try to identify and describe the structure of the legal documents; this includes identification and mark-up of each structural unit of the document (title, articles, sections, chapters, paragraphs, annexes, etc.). In that way, the structural units can be later precisely referred by linking tools where there is a legal citation quoting the respective structural unit. In this step, we also identify document metadata values. Our parser follows an eager approach as to identify as many metadata values as possible, as we assume by design we assume that no metadata is available to the parser. If content metadata is available to our parser, it is parametrically defined whether supplied metadata should prevail over discovered metadata, in cases of conflict, or as an override mechanism.

(d) *Legal Modelling*. In this step, an in-memory model of the document is iteratively constructed as new elements are identified in the text source. Also, this step assigns permanent URI to legal resources based on the technical specifications of the chosen data model and schema (Section 2.2).

(e) *Semantic check and validation*. As a final step semantic check and validation detects any inconsistencies the text may contain from the legal point of view or any discrepancies with

the chosen legal meta model. Our data model follows the LegalDocML schema, a schema where certain elements form a hierarchy of containment. A hierarchy is a set of arbitrarily deep nested sections with title and numbering. Each level of the nesting can contain either more nested sections or blocks and no text is allowed directly inside the hierarchy, but only within the appropriate block element. As a first process in our semantic check and validation step, we perform a schema validation i.e., validate the resulting XML file against the LegalDocML schema. Furthermore, LegalDocML uses only one hierarchy, with predefined names and no constraints on their order or systematic layering. However, legal documents follow constraints on the order or sequence of such hierarchical nested elements e.g., a sub-chapter should be nested inside a chapter or paragraph numbered '4' should precede paragraph numbered '5'. Thus, a second process is needed, verifying that the resulting XML files adhere to the constraints of legal documents. Potential errors identified are logged so that the content administrator can access whether they are based on inconsistencies of input text or a system malfunction.

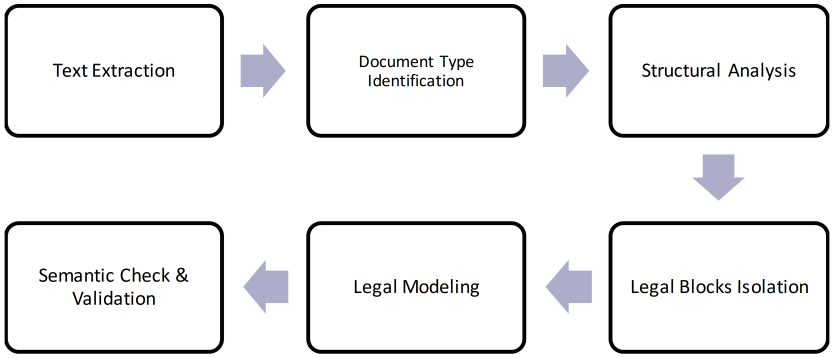

Figure 10. Parsing follows a pipeline strategy utilizing a top-down approach.

3.3.2. Interlinking Legal Sources

The electronic availability of legal sources in a structured and standard format and the interlinking of legal sources establishing interconnections within the same or similar repositories are necessary preconditions for effective legal document dissemination. Given the abundant use of citations between legal sources, human tagging, is a rather expensive operation, both in terms of time and effort, making it a a less attractive option. Methods for automated extraction of references provide a viable alternative. Although legal citations should in theory follow a predictable structure, as official guidelines for legislative drafting establish rules for legal references, deviations from these formal guidelines are the rule rather than the exception. This is attributed to many different factors, e.g., changes in citation norms over time, human factor etc.

The task of detecting and resolving references constitutes of several steps. Firstly, following terminology defined in [20], in cooperation with legal experts we identified and categorized the types of references, e.g., simple, complex, complete and incomplete and produced sets of specific string patterns. Building upon that outcome, we created sets of regular expressions and developed several algorithms for the identification of legal citations. Our research was initially focused in references linking Greek legal sources but afterwards we extended our field of study to cope for references between Greek and EU legal sources. Upon reference discovery, reference resolution follows, identifying the URI's of the legal sources cited. As already mentioned, our parser creates work level identifiers for each structural unit identified, following ELI standard. Our link resolver, follows the same principles and constructs ELI compatible URI's for each reference discovered. To cope with aliases, that can not be automatically processed, we have compiled a database of relevant info, functioning as a resolution mechanism, between well known aliases in the field of law and work level identifiers in terms of URI.

3.3.3. Document Classification

Our classification mechanism, follows a deterministic approach, based on a custom developed rule engine. Rules, defined through the administration UI module, are executed against the legal sources using priorities, as to avoid any potential conflicts. Rules can be simple or combined forming complex chains of operation, acting upon the textual data or metadata of the legal sources. For example 'if the issuing department of the administrative act is x, the signer of the document is y and the date of issue is within the z range then classify the legal source as w'.

As such, there is the possibility that a certain document will fail to classify. In this particular case, during the manual content curation phase, operators may choose to manually enrich the document with desired categories or publish as is. In the former case, they will also investigate the reason behind this lack of classification, potentially adding new rules or altering existing. In the latter case, the document would still be discoverable based on the metadata discovered in the parsing process.

Additionally, we are currently in the process of utilizing machine learning techniques as to automatically classify legal sources with descriptors from EuroVoc (http://eurovoc.europa.eu), a multilingual, multidisciplinary thesaurus covering the activities of the EU, which is also used to index EU legislation sources.

3.4. Collaborative Semantic Interlinking

Annotations are a popular mechanism to integrate end users' knowledge into the digital curation processes and augment digital items with additional information. In a previous work [21], the infrastructure allowing users to reuse established ontologies as well as semantics created on-the-fly in order to annotate and share resources over dynamically defined usage contexts was presented. The Collaborative semantic interlinking module aims at enhancing a collaborative environment for legal resources management with domain knowledge. Users are able to leverage linked data technologies for creating, exploiting and organizing collaborative information spaces containing diverse legal sources hosted in the platform. Doing so we provide a layer of semantics, inherently shared between end users, assisting them to manage, connect, organize and explore legal resources in various meaningful ways. Furthermore, exploring legal resources via a faceted browsing functionality offers an intuitive way of searching over them.

3.5. Search

The aim of any information retrieval system is to deliver content that can precisely match and satisfy the users information need. Since our intended user group includes people of various levels of domain and technical expertise, ranging form legal experts to novices, we employed various techniques, utilizing various dimensions over the legal information sources, as to precisely match information requests. Our information retrieval component, has been build on top of Solr (http://lucene.apache.org/solr/) , a popular open source enterprise search engine, offering full-text search, hit highlighting, faceted search, real-time indexing, supporting advanced customization through a plugin architecture and that can be integrated with our repository component.

In traditional IR techniques [22] documents are indexed and retrieved as single atomic units. However, the nature of legal sources, e.g., legal documents tend to cover multiple topics, be quite long, have a hierarchical structure of nested element, implies that a more refined approach is needed. As such, we follow structured retrieval techniques [23], allowing for the combination of textual criteria, natural-language sentences, with structural criteria, restrictions in terms of the units in which to search the query. By considering legal documents as aggregates of interrelated structural elements that need to be indexed and retrieved, both as a whole and separately according to user requirements, we can perform complex queries which combine metadata and full-text predicates in a single query. Consequently our approach offers more accurate and relevant results, however it leads to a redundant index, since text occurring at depth i of the document tree is indexed i times.

Additionally, as a means of improving user satisfaction by increasing the variety of information shown to user, we also employ search result diversification methods. In our previous works [24–26] we originally introduced the concept of diversifying legal information search results, we analysed the impact of various legal sources features in computing the query-document relevance and document-document similarity scores, we introduced legal domain-specific diversification criteria and performed an exhaustive evaluation of several state of the art methods from the web search, network analysis and text summarization domains. In Solon's information retrieval component, alongside the default ranking process, we offer alternative ranking models, based on diversification techniques, demonstrating notable improvements in terms of enriching search results with otherwise hidden aspects of the legal query space.

4. Digital Library System

Solon architecture has been recently materialized, operating in a production environment, http://elib.aade.gr/elib, under the supervision of the Independent Authority for Public Revenue (tax authority in Greece), http://www.aade.gr/, aiming to provide semantic access to Greek tax legislation.

The project purpose, from the Independent Authority for Public Revenue point of view, is to 'provide citizens and civil servants, with a user-friendly, electronic tool for locating and retrieving legal and regulatory documents in an easy, accurate and up-to-date manner. In addition, through the Electronic Library, information will be provided in the exact context of a conceptual reference, without unnecessary information' [27].

In this particular use case, we are concerned about the authoritativeness of information providers. With primary legal sources, such as regulations, case law and administrative acts, the government is the author, and government websites are authoritative publishers. We currently, utilize three external sources of legal information. The first one is the website of the *Greek government gazette* (http://www.et.gr), the Official Journal of the Hellenic Republic, which provides, presidential decrees, laws and regulatory acts of the Council of minister. In this web portal, access to legal documents is done thought the use of a search form where the user has to specify keywords of the legal document he/she is interested in, the publication year and the type of legal document. Relevant documents are then presented to the end user, through a search result page, in pdf file format, without any accompanying metadata. The second deployed Crawler harvests the *Transparency Portal* (https://diavgeia.gov.gr/en), where all Greek government institutions are obliged to upload their acts and decisions. Furthermore, administrative acts and decisions are not valid unless published in the Transparency Portal. A REST-like data API provides access to all decisions and supplementary information, including a taxonomic structure of its content, in various formats i.e., XML JSON, atom, RSS. Finally, a third crawler is deployed, as an alternate information feed, covering special content upload requests. Authorized users through the web interface, can upload legal sources to the platform, thus providing an alternate information feed.

Overall the platform supports laws, presidential decrees, ministerial decisions, regulatory acts of the Council of minister, acts and decisions from government institutions as well as Regulations and Directives from the European Parliament and the Council. It currently hosts more than 39,700 legal and regulatory documents.

Figure 11 illustrates the platforms basic navigational facets accessible at http://elib.aade.gr/elib/navigation. Note that Κανονισμός and Οδηγία refer to EU Regulations and Directives respectively.

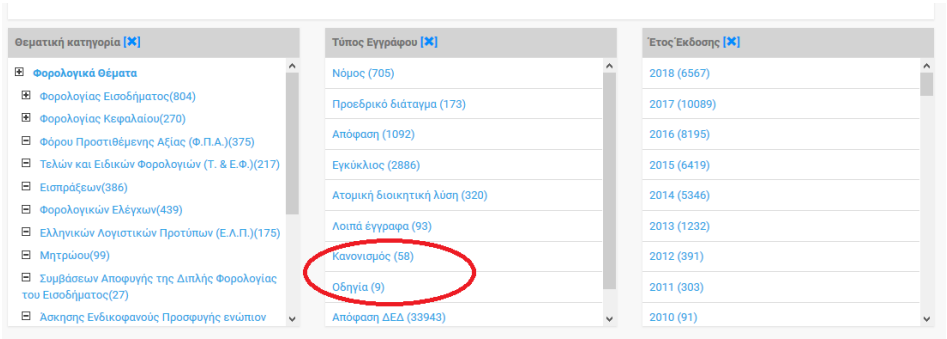

Figure 11. Basic navigation facets. Numbers in parenthesis denote the number of legal documents having that feature. The (red) circle highlights Regulations and Directives from the European Parliament and the Council (In Greek).

Unstructured legal sources are then modelled by the parser into a machine-readable, semantic representation, which is afterwards enriched by the process of interlinking legal sources and document classification. Manual curation process, as performed by civil servants, through the administrative interface fine tunes the automatic procedure. Advanced search facilities, offered by the search component, provide accurate legal sources, that can be further customized utilizing functionality exposed by the collaborative semantic interlinking module, thus enabling users to organize legal information according to individual needs and explore legal resources in various meaningful ways.

Figure 12 illustrates the platforms advanced search form, accessible at http://elib.aade.gr/elib/search. Users can utilize a combination of multiple metadata/classification criteria, automatically discovered from legal documents, as to satisfy their information needs.

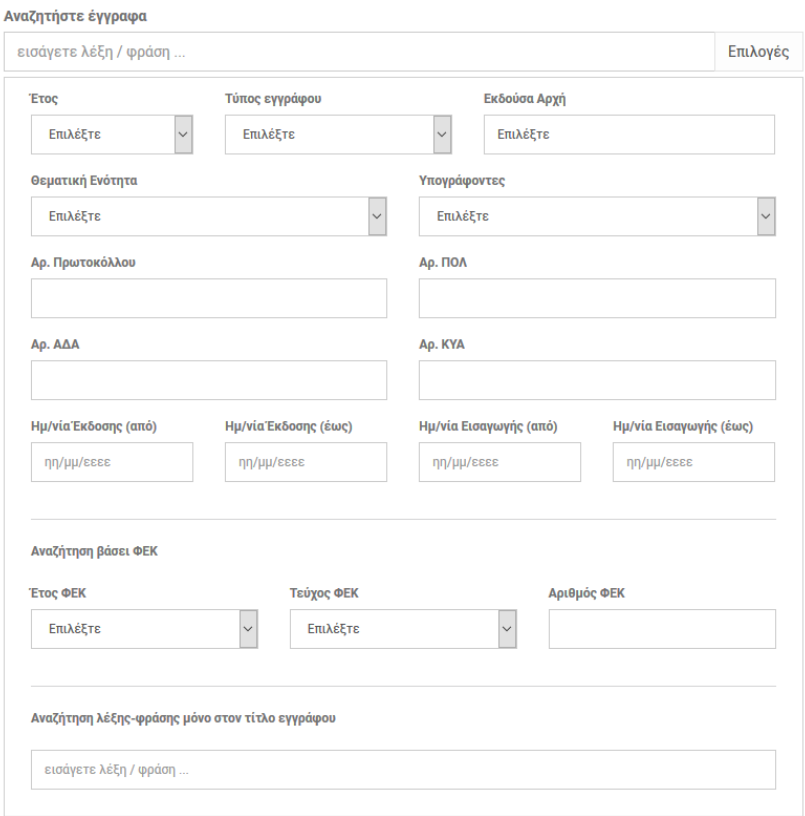

Figure 12. Advanced Search Form. Users can combine multiple metadata/classification criteria automatically discovered from legal documents (In Greek).

Solon covers a wide range of user groups. Several civil servants of the Independent Authority for Public Revenue have been assigned the *content manager* role, thus, they are responsible for curating digital content, releasing the content for public use, and fine-tuning the text mining process. *Administrators* are responsible for managing access permissions/rights to user groups and roles and monitor overall system performance. General public, *unregistered users* can browse, read and search published content in the platform. Finally, *registered end users* can additionally augment digital items with additional information and organize legal sources into collaborative information spaces.

Annotations are a popular mechanism to integrate end users' knowledge into the digital curation processes and augment digital items with additional information. In a previous work [21], the infrastructure allowing users to reuse established ontologies as well as semantics created on-the-fly in order to annotate and share resources over dynamically defined usage contexts was presented. The Collaborative semantic interlinking module aims at enhancing a collaborative environment for legal resources management with domain knowledge. Users are able to leverage linked data technologies for creating, exploiting and organizing collaborative information spaces containing diverse legal sources hosted in the platform. Doing so we provide a layer of semantics, inherently shared between end users, assisting them to manage, connect, organize and explore legal resources in various meaningful ways. Furthermore, exploring legal resources via a faceted browsing functionality offers an intuitive way of searching over them.

Users, both internal (Independent Authority for Public Revenue personel) and external (general public) were positively impressed with the use of the platform. We performed several interviews with representatives from the aforementioned categories of end users, collecting their feedback and received several suggestions, mainly focused upon improving the usability of the application. Based upon the initial feedback, we are preparing an on line survey as to qualitative and quantitative assess provided services, as perceived by end users. Individual methods and components of our platform have been extensively evaluated seperetaly [8,25,26], but we also plan on a more refined technical evaluation of the platform.

5. Related Work

Several lines of work are related to the present paper. In this section, we first present related work on Legal Resources Meta Models, afterwards on Legal Document Management Systems, then on Legal Document Structuring techniques, followed by Legal Citations and finally on Legal Text Retrieval techniques.

5.1. Legal Resources Meta Models

A legal XML schema is required to structure documents, represent metadata and (cross-)links, and format legislative documents. Several legal document standards were introduced in the recent years. Among the more wide spread ones are MetaLex [2], NormeInRete [3], AkomaNtoso [4], United States Legislative Markup (https://github.com/usgpo/uslm), United Kingdom Legislation (http://www.legislation.gov.uk/developer/formats/xml) and LexML Brasil [28]. A comparative analysis of the main standards for legislative documents, underlining specific strengths and weaknesses of each standard, at European and extra-European levels is presented in [29]. In this work we utilize the Akoma Ntoso schema providing extensions to accommodate for custom Greek legal sources structure and metadata.

5.2. Legal Document Management Systems

Early optimism with the advent of the first legal expert systems has now, for various reasons, disappeared [30] and research is currently more focussed in the area of legal document management systems. Eunomos [31] is a legal knowledge management system employing ontologies to classify norms. Compared with Solon it does not conform with Linked Data Platform specification or offers advance search facilities. The MetaLex Document Server [32] offers Legal Documents as versioned

Linked Data. It applies a generic conversion mechanism from legacy legal XML syntaxes to CEN MetaLex, while in contrast Solon automatically structures legal sources in XML and identifies/resolves legal citations. In a similar approach [33] aims to publish Finnish law as a Linked Open Data service, modelling statutes and court decisions in RDF, focusing more on re-use of the data in various mash-up services, rather than providing a Legal Document Management System.

5.3. Legal Document Structuring

A machine-readable representation of legislation may be created at the same time that the text is edited, asking the user to provide information about the text, through specialized editing software, as presented in [34,35], or afterwards. In this work, we focus on the latter approach of creating a model after the text is created, also suited for legacy legal sources. The work presented in [36], where the authors present a method for document structure analysis with syntactic model and parsers for Japanese legal judgments, is closer to ours. In contrast with [36], our method does not rely on PEG [37] rules, but on CFG parser which allows for more flexibility in regards to ambiguity and context-dependent grammar by means of predicates [19].

5.4. Legal Citations

Various research methods have been proposed for detecting and resolving references between sources of law [20,38,39]. Network theory has also been applied in the field of law to construct citation networks [40], evaluating the relevance of court decisions [41], assisting summarizing legal cases [42] and proposing a network-based approach to model the law [43]. In this work, we employ automatic methods for detecting and resolving references between sources of law and also utilize the network of citations as to assist user navigation and diversify search results.

5.5. Legal Text Retrieval

Various dimensions of relevance in legal information retrieval based on specific features of legal information are examined in [44]. Legal text retrieval traditionally relies on external knowledge sources, such as thesauri and classification schemes, utlizing several supervised learning methods [45]. In this work, we do not rely on external knowledge sources but utilize the structure and specific features of legal sources as to provide more refined and diversified search results.

6. Conclusions and Future Work

In this paper, we presented Solon, a platform suitable for modelling, managing and mining legal sources. It utilizes a novel method for extracting semantic representations of legal sources from unstructured formats, interlinking and adding classification features. Also, it provides more refined and diversified search results utilizing the structure and specific features of legal sources, while allowing users to connect, organize and explore legal resources according to individual needs.

Currently, Solon deals with ministerial decisions, laws, regulatory acts of the Council of minister, presidential decrees and administrative acts; however, we are currently extending its applicability to cover judicial legal sources as well. Several other extensions of this research are currently under implementation or investigation. These include the application of natural language processing for the identification of named entities, the temporal management of legal resources (e.g., the time validity of a legal block), the automatic soft encoding of legislative sources (e.g., generation of amending documents based on the editing of an existing law or proposal) and, finally, the automatic consolidation of (proposed) legislation based on the original and its amending documents and finally the application of machine learning methods for advanced classification of legal documents. Furthermore, as a means to enrich our classification module, we are currently experimenting with machine learning techniques. Specifically, we wish to automatically classify legal sources with descriptors from the multilingual and multidisciplinary thesaurus EuroVoc, which is also used to index E.U. legislation sources.

Author Contributions: M.K. conceived the idea, designed the platform, drafted the initial manuscript and revised the manuscript; G.P. and I.A. designed the platform, helped to draft the initial manuscript and revised the final version.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflict of interest.

Dataset: https://github.com/mkoniari/Solon.

References

1. World Legal Information Institute. Declaration on Free Access to Law. 2012. Available online: http://www.worldlii.org/worldlii/declaration/ (accessed on 2 December 2018).
2. Boer, A.; Winkels, R.; Vitali, F. Metalex xml and the legal knowledge interchange format. In *Computable Models of the Law*; Springer: Berlin, Germany, 2008; pp. 21–41.
3. Marchetti, A.; Megale, F.; Seta, E.; Vitali, F. Using XML as a means to access legislative documents: Italian and foreign experiences. *ACM SIGAPP Appl. Comput. Rev.* **2002**, *10*, 54–62. [CrossRef]
4. Barabucci, G.; Cervone, L.; Palmirani, M.; Peroni, S.; Vitali, F. Multi-layer markup and ontological structures in Akoma Ntoso. In *AI Approaches to the Complexity of Legal Systems. Complex Systems, the Semantic Web, Ontologies, Argumentation, and Dialogue*; Springer: Berlin, Germany, 2010; pp. 133–149.
5. Inter-Parliamentary Union. World e-Parliament Report 2016. Available online: http://www.ipu.org/pdf/publications/eparl16-en.pdf (accessed on 2 December 2018).
6. Tillett, B. A Conceptual Model for the Bibliographic Universe. *Technicalities* **2003**, *25*, 5. [CrossRef]
7. Francesconi, E. On the Future of Legal Publishing Services in the Semantic Web. *Future Internet* **2018**, *10*, 48. [CrossRef]
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, Patras, Greece, 10–12 November 2016.
9. Koniaris, M.; Papastefanatos, G.; Meimaris, M.; Alexiou, G. Introducing Solon: A Semantic Platform for Managing Legal Sources. In *International Conference on Theory and Practice of Digital Libraries*; Springer: Cham, Switzerland, 2017; pp. 603–607.
10. Publications Office of the European Union. *Interinstitutional Style Guide: 2012*; EU Publications: Luxembourg, 2012. [CrossRef].
11. Organization for the Advancement of Structured Information Standards (OASIS). Advancing Worldwide Best Practices for the Use of XML in Legal Documents, OASIS LegalDocumentML (LegalDocML) TC, 2016. Available online: https://www.oasis-open.org/committees/legaldocml/ (accessed on 2 December 2018).
12. Official Journal of the European Union. Council Conclusions of 6 November 2017 on the European Legislation Identifier *(2017/C 441/05)*, OJ C 441, 22.12.2017, 2017; pp. 8–12. Available online: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52017XG1222%2802%29 (accessed on 2 December 2018).
13. The future of Research Communications and e-Scholarship, FORCE11. The FAIR Data Principles, 2018. Available online: https://www.force11.org/group/fairgroup/fairprinciples (accessed on 2 December 2018).
14. Lagoze, C.; Payette, S.; Shin, E.; Wilper, C. Fedora: An architecture for complex objects and their relationships. *Int. J. Digit. Libr.* **2006**, *6*, 124–138. [CrossRef]
15. World Wide Web Consortium (W3C). Linked Data Platform 1.0. Available online: http://www.w3.org/TR/ldp/ (accessed on 2 December 2018).
16. Evans, E. *Domain-Driven Design: Tackling Complexity in the Heart of Software*; Addison-Wesley: Boston, MA, USA, 2004.
17. Parr, T. *Language Implementation Patterns: Create Your Own Domain-Specific and General Programming Languages*, 1st ed.; The Pragmatic Programmers: Raleigh, NC, USA, 2009.
18. Fowler, M. *Domain Specific Languages*; Addison-Wesley Professional: Boston, MA, USA, 2010.
19. Parr, T.; Harwell, S.; Fisher, K. Adaptive LL (*) Parsing: The Power of Dynamic Analysis. *ACM SIGPLAN Notices* **2014**, *49*, 579–598. [CrossRef]
20. De Maat, E.; Winkels, R.; van Engers, T. Automated Detection of Reference Structures in Law. In *Proceedings of JURIX 2006*; IOS Press: Amsterdam, The Netherlands, 2006; pp. 41–50.

21. Meimaris, M.; Alexiou, G.; Papastefanatos, G. LinkZoo: A linked data platform for collaborative management of heterogeneous resources. In *European Semantic Web Conference*; Springer: Cham, Switzerland, 2014; pp. 407–412.
22. Hand, D.J.; Mannila, H.; Smyth, P. *Principles of Data Mining*; MIT Press: Cambridge, MA, USA, 2001.
23. Amer-Yahia, S.; Lalmas, M. XML Search: Languages, INEX and Scoring. *SIGMOD Rec.* **2006**, *35*, 16–23. [CrossRef]
24. Koniaris, M.; Anagnostopoulos, I.; Vassiliou, Y. Diversifying the Legal Order. In *IFIP International Conference on Artificial Intelligence Applications and Innovations*; Springer: Cham, Switzerland, 2016; pp. 499–509.
25. Koniaris, M.; Anagnostopoulos, I.; Vassiliou, Y. Multi-dimension Diversification in Legal Information Retrieval. In *International Conference on Web Information Systems Engineering*; Springer: Cham, Switzerland, 2016; pp. 174–189.
26. Koniaris, M.; Anagnostopoulos, I.; Vassiliou, Y. Evaluation of Diversification Techniques for Legal Information Retrieval. *Algorithms* **2017**, *10*, 22. [CrossRef]
27. Independent Authority for Public Revenue. Business Plan, 2016. Available online: https://www.aade.gr/sites/default/files/2016-12/epixirisiako_sxedio_ggde_2016_v5.pdf (accessed on 2 December 2018). (In Greek)
28. Lima Jao, C.F. LexML Brasil, Parte 3—LexML XML Schema, Version 1.0, 2016. Available online: http://projeto.lexml.gov.br/documentacao/Parte-3-XML-Schema.pdf (accessed on 2 December 2018).
29. Lupo, C.; Vitali, F.; Francesconi, E.; Palmirani, M.; Winkels, R.; de Maat, E.; Boer, A.; Mascellani, P. ESTRELLA Project, Deliverable D3.1—General XML Format(s) for Legal Sources, Version 1.0. Available online: https://pdfs.semanticscholar.org/a5ee/a8dfc5bad0e9d368cd60fffe1e885c237fe8.pdf (accessed on 2 December 2018).
30. Leith, P. The rise and fall of the legal expert system. *Eur. J. Law Technol.* **2010**, *1*, 1. [CrossRef]
31. Boella, G.; Humphreys, L.; Martin, M.; Rossi, P.; van der Torre, L. Eunomos, a legal document and knowledge management system to build legal services. In *International Workshop on AI Approaches to the Complexity of Legal Systems*; Springer: Cham, Switzerland, 2011; pp. 131–146.
32. Hoekstra, R. The MetaLex Document Server. In *Proceedings of the 10th International Semantic Web Conference*; Springer: Berlin/Heidelberg, Germany, 2011; pp. 128–143.
33. Frosterus, M.; Tuominen, J.; Hyvönen, E. Facilitating Re-use of Legal Data in Applications—Finnish Law as a Linked Open Data Service. In *Legal Knowledge and Information Systems*; IOS Press: Amsterdam, The Netherlands, 2014; pp. 115–124.
34. Van De Ven, S.; Hoekstra, R.; Winkels, R.; de Maat, E.; Kollár, Á. MetaVex: Regulation drafting meets the semantic web. In *Computable Models of the Law*; Springer: Berlin, Germany, 2008; pp. 42–55.
35. Agnoloni, T.; Francesconi, E.; Spinosa, P. xmLegesEditor: An Opensource Visual XML Editor for Supporting Legal National Standards. Available online: http://www.xmleges.org/ita/images/articoli/art17.pdf (accessed on 2 December 2018).
36. Igari, H.; Shimazu, A.; Ochimizu, K. Document structure analysis with syntactic model and parsers: Application to legal judgments. *JSAI Int. Symp. Artif. Intell.* **2011**, *7258*, 126–140.
37. Ford, B. Parsing expression grammars: A recognition-based syntactic foundation. *ACM SIGPLAN Notices* **2004**, *39*, 111–122. [CrossRef]
38. Opijnen, M.V.; Verwer, N.; Meijer, J. Beyond the Experiment: The eXtendable Legal Link eXtractor. Workshop on Automated Detection, Extraction and Analysis of Semantic Information in Legal Texts, 2015. Available online: https://ssrn.com/abstract=2626521 (accessed on 2 December 2018).
39. Agnoloni, T.; Bacci, L.; Peruginelli, G.; van Opijnen, M.; van den Oever, J.; Palmirani, M.; Cervone, L.; Bujor, O.; Lecuona, A.A.; García, A.B.; et al. Linking European Case Law: BO-ECLI Parser, an Open Framework for the Automatic Extraction of Legal Links. *Legal Knowl. Inf. Syst.* **2017**, doi:10.3233/978-1-61499-838-9-113. [CrossRef]
40. Marx, S.M. Citation networks in the law. *Jurimetrics J.* **1970**, *10*, 121–137.
41. Fowler, J.H.; Johnson, T.R.; Spriggs, J.F.; Jeon, S.; Wahlbeck, P.J. Network Analysis and the Law: Measuring the Legal Importance of Precedents at the U.S. Supreme Court. *Political Anal.* **2006**, *15*, 324–346. [CrossRef]
42. Galgani, F.; Compton, P.; Hoffmann, A. Citation based summarisation of legal texts. In *PRICAI 2012: Trends in Artificial Intelligence*; Springer: Berlin/Heidelberg, Germany, 2012; pp. 40–52.
43. Koniaris, M.; Anagnostopoulos, I.; Vassiliou, Y. Network analysis in the legal domain: A complex model for European Union legal sources. *J. Complex Netw.* **2018**, *6*, 243–268. [CrossRef]

44. Van Opijnen, M.; Santos, C. On the concept of relevance in legal information retrieval. *Artif. Intell. Law* **2017**, *25*, 65–87. [CrossRef]
45. Moens, M. Innovative techniques for legal text retrieval. *Artif. Intell. Law* **2001**, *9*, 29–57. [CrossRef]