information

MDPI

Article

Application of an Ecosystem Methodology Based on Legal Language Processing for the Transformation of Court Decisions and Legal Opinions into Open Data

John Garofalakis [1], Konstantinos Plessas [1], Athanasios Plessas [1,*] and Panoraia Spiliopoulou [2]

[1] Department of Computer Engineering and Informatics, University of Patras, 265 04 Patras, Greece; garofala@ceid.upatras.gr (J.G.); kplessas@ceid.upatras.gr (K.P.)
[2] School of Law, National and Kapodistrian University of Athens, 106 53 Athens, Greece; panoraia@gmail.com
* Correspondence: plessas@ceid.upatras.gr

Received: 24 November 2019; Accepted: 19 December 2019; Published: 22 December 2019

Abstract: Regulation of modern societies requires the generation of large sets of heterogeneous legal documents: bills, acts, decrees, administrative decisions, court decisions, legal opinions, circulars, etc. More and more legal publishing bodies publish these documents online, although usually in formats that are not machine-readable and without following Open Data principles. Until an open by default generation and publication process is employed, ex-post transformation of legal documents into Legal Open Data is required. Since manual transformation is a time-consuming and costly process, automated methods need to be applied. While some research efforts toward the automation of the transformation process exist, the alignment of such approaches with proposed Open Data methodologies in order to promote data exploitation is still an open issue. In this paper, we present a methodology aligned to the Open Data ecosystem approach for the automated transformation of Greek court decisions and legal opinions into Legal Open Data that builds on legal language processing methods and tools. We show that this approach produces Legal Open Data of satisfying quality while highly reducing the need for manual intervention.

Keywords: Akoma Ntoso; legal open data; legal big data; open data ecosystem; natural language processing; domain specific language; legal parsing

1. Introduction

Due to increased complexity of the political, economic, and social environment in which people live and act, modern societies have set up institutions that produce constantly growing sets of legal documents for their regulation. These sets include documents of primary legislation (e.g., the constitution and the laws), documents of secondary legislation (e.g., ministerial decisions), court decisions, legal opinions, and even documents released from public administration bodies regarding the application of the law (e.g., administrative circulars). As technology advances and the web expands, these documents, originally printed on paper, are now available online in most countries. Open access to law is a basis for democratic society and computer technology facilitates public access to legal resources at a low cost [1]. Peruginelli describes law as "the operating system of our society", stating that "the nature of law is so pervasive, it becomes essential for everybody to know about it" [2]. Initiatives such as the "free access to law movement" [3] and public accessibility projects will probably result in governments providing online access to even more legal information resources in the future [4]. The volume of these heterogeneous documents is expected to constantly rise, forming datasets that show many of the characteristics that define Big Data. Even if their volume cannot be compared to the volume of data collected from other sources such as social networks, their

manual analysis and processing is prohibitive and automated approaches are needed to undertake such tasks [5]. It is noteworthy that some researchers explicitly classify legal documents as Big Data since they meet at least two of the major aspects of Big Data: volume and variety [6]. While several governments are taking steps to apply semantic web technologies in the legal publishing domain (e.g., the legislation.gov.uk platform where UK legislation is published in XML and RDF) [7], in most countries these documents are usually made available in unstructured or poorly structured formats; for example, as PDF files, HTML documents, or plain text.

On the other hand, the current trend of Open Data requires that government data (legal documents being a special category of them) are published without technical and legal impediments, in structured and machine-processable formats, and under an open license. Janssen et al. [8] have discussed in detail the expected benefits from the adoption of the Open Data model: transparency, democratic accountability, economic growth, stimulation of innovation, creation of new services, improved decision and policy making, equal access to data, etc. To ensure these benefits of Open Data, researchers propose the "transparency-by-design" model, an organization model for data publication systems, which is expected to contribute to the automatic opening of data [9]. However, this is not yet the case in the legal domain, and the ideal scenario, where legal publishing systems are organized in such a way that legal information is generated in the form of Open Data by default, still seems distant. The availability of data in non-machine-processable formats, and the need for manual processing and conversion are identified as significant barriers for the use of Open Data, while other impediments include data heterogeneity and fragmentation and the lack of metadata for their description [10].

As Agnoloni et al. note in [4], availability of data in structured formats such as XML is a fundamental prerequisite for the interoperability of data originating from heterogeneous sources and the interoperability of applications that exploit them. Moreover, XML markup and semantic annotation facilitates the transition from documents to data [4]. The need for structured and formally described legal documents has led to the development of legal editors, which allow their users to draft new legal documents or to markup existing ones [11]. As a result, legislative bodies may adopt these tools in order to publish structured versions of their newly generated documents. However, the conversion of existing documents to a structured format is a problem that cannot be tackled manually, as the manual markup of such large sets of texts is a laborious and extremely time-consuming process [12]. Since legal language is natural language, and legal documents often have a standard structure and follow predefined common patterns, Natural Language Processing (NLP) tools can be used to automate the conversion process. However, automating tasks related to the processing of legal texts is not as easy as one could expect and sometimes even humans face difficulties in addressing these tasks (for example Dragoni et al. mention the identification of legal rules and conditions as such a task [13]), since legal language combines complex phrases with complex sentence syntax [14]. While NLP is a mature subfield of Artificial Intelligence and several techniques with satisfying performance are available for accomplishing linguistic tasks (e.g., stemming, tokenizing, part-of-speech tagging, sentence splitting, etc.) for some types of text (e.g., narrative or newspaper text), their application to legal text is challenging due to the complexity of legal language [15].

As we show in the Related Work section, some research efforts that focus on automated transformation of legal documents into open formats have already taken place. However, these efforts are not based on the established theoretical frameworks for Open Data exploitation, and in order to fill this gap, we present a methodology aligned with the so-called Open Data ecosystem approach. Our work is based on legal language parsing and processing for the transformation of legal documents available on the web, but residing in heterogeneous sources and in non-machine-readable formats, into Legal Open Data. While our approach can be adapted and applied for any type of legal documents, we focus on Greek court decisions and legal opinions and we show that treating legal language as a Domain Specific Language (DSL) and defining grammar rules for parsing legal texts provides satisfying results, significantly reducing the need for manual intervention, and thus paving the way for fully automating the transformation process in the future.

2. Related Work

Several research efforts take advantage of legal language processing for the automated transformation of unstructured legal documents into Open Data or Linked Open Data. The OpenLaws EU project [5] aimed at aggregating legal resources (legislation and case law) from EU and member states and exposing them as Big Open Legal Data through an innovative platform. NLP techniques were used to semantically analyze the approximately 1,9 million legal documents that were collected and detect legal references. Another relevant EU project was EUcases [16], which adopted Akoma Ntoso as the format for representing law and case law. The objective of the project was the automated transformation of multilingual Legal Open Data into Linked Open Data after semantic and structural analysis with natural language parsing, and in order to achieve it, rule-based parsers that look for specific keywords were implemented. The parsing tools were accomplishing the task of extracting information necessary to form the Akoma Ntoso metadata section and the task of identifying structural units of the legal documents which were used for marking up the texts. ManyLaws [17] is also an ongoing EU project that intends to exploit natural language processing and text mining in order to produce semantically annotated Big Open Legal Data. Furthermore, in [12] Sannier et al. describe their approach and the lessons learned from the large-scale automatic markup of five legislative codes of Luxembourg in Akoma Ntoso. NLP scripts (applying regular expressions for pattern matching) are used in conjunction with a conceptual model to detect and markup the structural elements of the texts and the legal cross-references. The researchers focused exclusively on structural markup and did not proceed to semantic markup of the documents. In another, recent publication [18], the authors describe an automated XML marker transforming legal texts into Akoma Ntoso format, which uses a hybrid approach of both rule-based and machine learning tools for the implementation of the structural marker and the named entity recognizer. Their approach resulted in significant reduction of the required time effort compared to manual markup of the texts.

Interestingly, during recent years, research teams from Greece have put efforts in parsing Greek legislative documents and transforming them into structured data. Chalkidis et al. [19] implemented a rule-based parser that parses PDF documents of the Greek Government Gazette and transforms them into Linked Open Data (RDF triples). In [20] Koniaris et al. describe a DSL based approach for parsing legal documents, identifying their structural elements and metadata, and converting plain text to XML following the LegalDocML schema. Regular expressions are used to detect legal citations and interlink the documents.

3. Open Data Methodology

Several models have been proposed to describe the Open Data lifecycle (i.e., the processes and practices related to data handling from creation to exploitation), each one having its own strengths and weaknesses [21]. Currently, more researchers have been highlighting the need to replace traditional Open Data practices with approaches that focus on the larger Open Data environment, known as the Open Data ecosystem, in order to create value from Open Data. For an extensive literature review on this subject, one may refer to [22]. Following this "ecosystem" approach, the researchers in [23] combined steps from existing models in an extended Open Data lifecycle, which consists of the following stages: create/gather, pre-process, curate, store/obtain, publish, retrieve/acquire, process, use, collaborate with users, and provide feedback. As Open Data ecosystems involve a data provider and a data user level [22], the researchers imagined that the above stages form two interdependent cycles [21]: the inner one referring to the data provider level (create, preprocess, curate, store/obtain, publish) and the outer one to the data user level (retrieve/acquire, process, use, collaborate with users, and provide feedback). Similarly, Lnenicka and Komarkova [24] focusing on Big and Open Linked Data (BOLD) ecosystems, identified several types of involved stakeholders (ecosystem orchestrator, service provider, application provider, data producer, data publisher, data user, and data prosumer) and proposed a BOLD analytics lifecycle consisting of six phases: acquisition and extraction, management and preparation, storage and archiving, processing and analysis, visualization and use, publication, sharing,

and reuse. Each stakeholder participates in different phases of the model with a different role in each phase.

The work presented in this paper is part of a wider project [25] which aims to collect a variety of Greek legal documents from available heterogeneous sources and transform them into Legal Open Data. In order to fully exploit the benefits of Open Data, our methodological approach was designed having the ecosystem approach in mind and is aligned with the models presented above. Previous research efforts do not take into account this proposed methodology and we believe that our work fills in this gap. Table 1 shows the steps of our approach and how they match the stages of the extended Open Data lifecycle (inner cycle for data publishers) and the BOLD analytics lifecycle. As we are not actually the creators of the legal resources, but we are generating structured legal data from the large datasets of unstructured legal documents that we collect, our role is more that of a "transforming" publisher. Since the two models describe different roles for the involved stakeholders of the ecosystem, we set our role to more accurately represent the data provider role from the extended Open Data lifecycle and the data publisher role from the BOLD analytics lifecycle. As a result, as part of our methodology we define steps and actions related to the phases that correspond to these roles.

Table 1. Steps of our methodology and mapping to the stages of the Extended Open Data lifecycle and the BOLD (Big and Open Linked Data) analytics lifecycle.

Extended Open Data Lifecycle [23]	Our Methodology	BOLD Analytics Lifecycle [24]
Create/Gather	1. Identify online sources of legal documents	Acquisition and extraction
	2. Collect documents	
Pre-process	3. Transform to plain text	Management and Preparation
	4. Modelling	
Curate	5. Transform to structured format	
Store	6. Store	Storage and Archiving
		Processing and Analysis
		Visualization and Use
Publish	7. Publish	Publication, Sharing and Reuse

The steps of our methodology for this Legal Open Data project are described below:

1. Identify online sources of legal documents: While publishing of data in fragmented sources is considered to pose severe impediments to their reuse [10], legal documents are usually published through a variety of web platforms and there is not a single access point. Moreover, sometimes the same documents reside in more than one web locations. Consequently, a first step in the process is the identification and evaluation of available online sources of legal documents.
2. Collect documents: During this step, legal documents are gathered from the selected sources of the previous step. This task requires the development of software that takes advantage of available APIs and web services or, in the frequent case that such services are not provided, the implementation of web scrapers. Unfortunately, in the latter case, apart from the required effort for the development of a different script for each source, there is also the disadvantage that even a small change in the source's structure may turn the scraper not functional.
3. Transform to plain text: Legal documents are often published in closed, not machine-processable formats, such as PDF or Microsoft Word. In order to extract the legal content from these files, it is required to convert them in plain text. This step includes also pre-processing tasks for the removal of erroneous or irrelevant elements (e.g., page numbers, footers, headers, etc.) injected from the conversion process.

4. Modelling: Several standards for the modeling of legal resources have been developed as part either of national or international initiatives [26]. Each standard defines a set of metadata and structural and semantic elements. In this step, the appropriate model must be adopted according to the project requirements. In our case, the Akoma Ntoso document model is used and the reasons for this choice are explained in Section 5.
5. Transform to structured format: During this step, NLP techniques are applied in order to identify the metadata of the legal schema (in case they are not available from the original source) and the structural parts of the documents. In addition, semantic information about elements of the documents (e.g., legal references, persons, locations, dates, etc.) is extracted. The legal language processing approach we followed to accomplish the transformation is presented in Section 6.
6. Store: This step involves decisions related to the storage of the generated open datasets. Data can be uploaded to existing repositories (e.g., national data portals) or to newly deployed data managements systems. Moreover, in this step the datasets can get linked to other available open datasets and APIs or web services providing access to them can be developed.
7. Publish: In this final step, legal issues related to the license under which data are published and to intellectual properties rights are covered.

The above methodology was implemented in the framework of our research and the architecture of the system that we designed is shown in Figure 1. In the rest of the paper, we mainly focus on the components that are related to steps 1 and 2 (Section 4—Data collectors), 4 (Section 5—Modelling) and 5 (Section 6—Legal Text Processor) of the methodology. Some information is also provided for the implementation of the other steps, however not in full detail, since they are related more to technical decisions and do not present significant research interest.

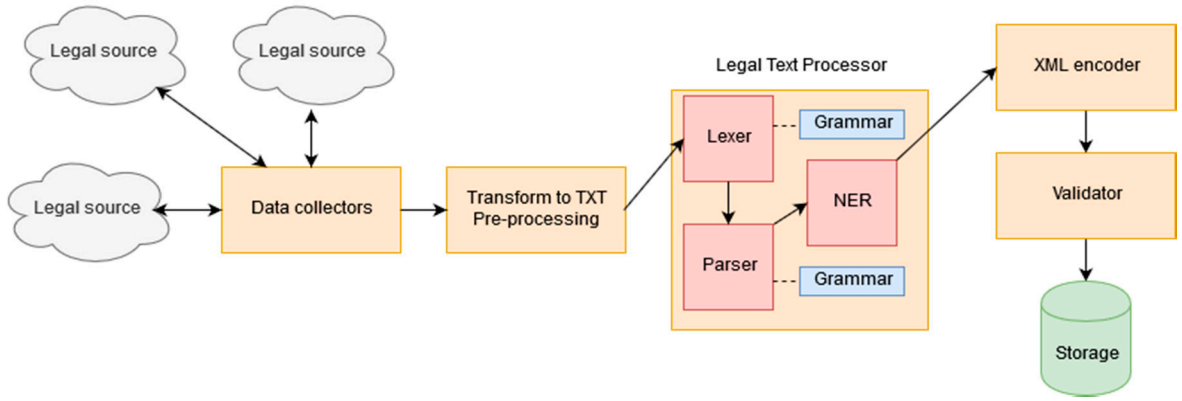


Figure 1. A proposed implementation of our methodology.

4. Sources of Legal Texts

While our project involves several types of Greek legal documents, in this paper we focus on the transformation of two of these types into Open Data: judgments of two of the Supreme Courts of Greece and legal opinions of the Legal Council of State.

In Greece, two of the three Supreme Courts publish their decisions: The Supreme Civil and Criminal Court (SCCC—also known as Arios Pagos, a name originating from Ancient Greece) and the Council of State (COS), which is the Supreme Administrative Court of Greece. Most civil law countries follow a strict policy regarding personal data of applicants and defendants, based on the view that publishing such data does not serve the goals of transparency of the judicial process and spread of knowledge about jurisprudential developments [27]. Consequently, decisions of Greek courts are published online in an anonymized form as far as concerns natural persons. Both courts publish their decisions in poorly structured HTML documents, where the legal text resides as plain text within paragraph or div HTML elements. Since in both cases there is no API or Web Service that would

allow machine-to-machine interoperability, we had to develop web scrappers in order to download the court decisions and the limited set of (non-standardized) metadata elements that describe them. The overhead posed from this scraping process is identified as a barrier to court decisions re-use [27]. The Python crawling framework Scrapy was used for the case of the Supreme Civil and Criminal Court and the Selenium framework for the case of the Council of State, since in the latter case the existing search form adds non-persistent session IDs to the URLs that are difficult to handle and as a result the user click behavior had to be simulated. In total, 40193 judgments of the Supreme Civil and Criminal Court from 1997 to 2017 and 101337 judgments of the Council of State from 1990 to 2016 were retrieved (Figure 2). We should note that not all judgments of these courts are digitized and available to access online.

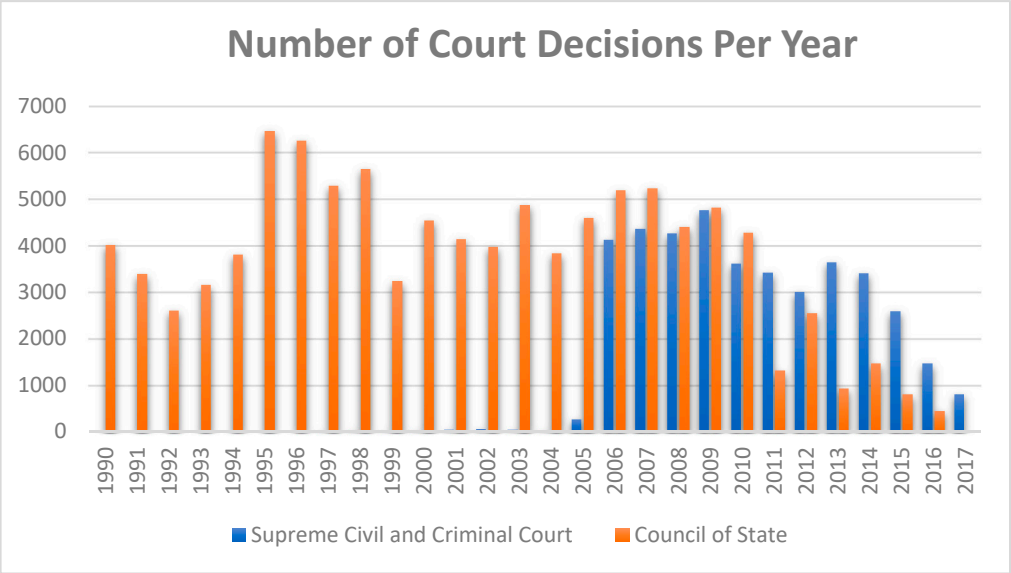


Figure 2. Number of collected court decisions per year.

The Legal Council of State, according to the Constitution of Greece, is assigned with the judicial support and representation of the State. Moreover, it is the responsible body for providing legal opinions to official questions of public administration bodies. These opinions are available from two sources: (a) the website of the Legal Council, where the opinions are available as scanned PDF files (hindering transformation into plain text since the conversion task requires the application of OCR techniques, which are not often able to detect with absolute precision the original text) accompanied by a small set of metadata (e.g., keywords, summary, etc.) and (b) the Diavgeia Portal, where all decisions of Greece's Public Administration are published followed by a set of common (once again non-standardized) metadata elements. Diavgeia was set up in 2010 and as a result only subsequent legal opinions are available, however almost half of the available PDF files can be easily transformed into plain text, since they are not scanned documents, at the cost of losing valuable formatting information (e.g., bold, italic or underlined text). Moreover, the resources are becoming available through a REST API that the Diavgeia platform provides. For these reasons, we implemented a python REST client in order to download the available (non-scanned) legal opinions and their metadata and transform the PDF files into plain text. In addition to this, a python scraper was written and used to collect the available metadata from the website of the Legal Council for the same opinions, since the metadata sets from the two online sources are disjoint. In total, 868 legal opinions from 2010 to 2017 were retrieved (Figure 3).

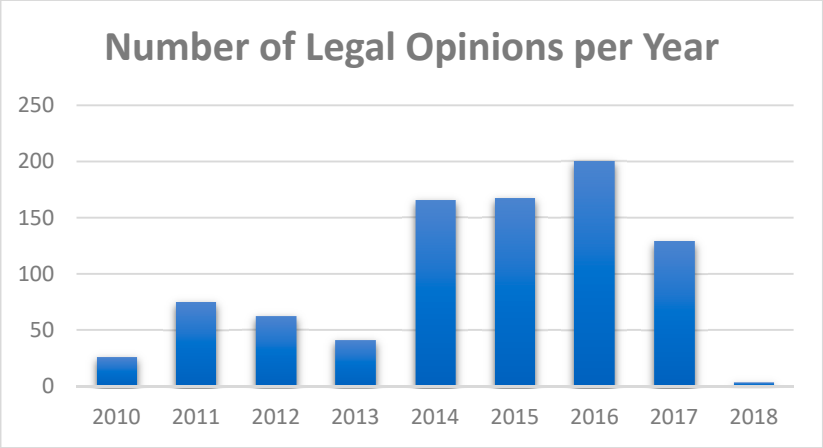


Figure 3. Number of collected legal opinions per year.

5. Modelling of Legal Documents

5.1. The Akoma Ntoso Data Model

As already mentioned, several standards are available for the modelling and representation of legal documents in a machine-readable format. Some of them are adapted to the needs of specific national legal systems (e.g., CHLexML in Switzerland or LexDania in Denmark), while others follow a more generic and extensible design that allows their usage in different national and international legal contexts. According to Pelech-Pilichowski et al. "an existence of shared standard for legal information significantly reduces costs of digitalization of legal information and guaranties higher level of interoperability of different systems" [28]. Following this argument, we decided to investigate the available options for using an international legal standard rather than a custom model adapted to the features of the documents we had collected. Two widely used XML schemas that belong to this category are Akoma Ntoso [29] and CEN Metalex [30].

Metalex is a meta-standard for other standards [30] and does not aim at directly providing a format for the representation of legal documents; its main purpose is instead to enhance interchange of legal documents by allowing the mapping of different compliant standards already used to markup the documents [31]. This is the main reason why we decided to apply the Akoma Ntoso standard as a more appropriate approach in the framework of our project, which requires the markup of legal resources in plain text.

Akoma Ntoso, which was recently accepted as an OASIS standard and conforms to Metalex, provides an XML schema for the machine-readable representation of parliamentary, legislative, and judiciary documents, as well as a naming convention for their unique identification based on the FRBR model [32]. The standard supports both structure and metadata modelling and allows for separation of the different layers of legal documents: text, structure, metadata, ontology, and legal rules. It implements the first three levels and provides hooks to external ontologies and legal knowledge modelling. Apart from structural and metadata elements, the schema provides also semantic elements, which can be used to capture the legal meaning of parts of the text (e.g., recognize legal references, the name of a judge, dates, etc.).

5.2. Legal Metadata

Availability of metadata for published Open Data provides numerous benefits [33]: improved accessibility, discoverability, searchability, storing, preservation etc. The Akoma Ntoso model defines an extensible core set of metadata elements (either required or optional) found within the metadata block (<meta>), which are used to describe the legal document:

- <identification>: a block for the unique identification of the document according to the FRBR model. We are using this block in accordance to the Akoma Ntoso Naming Convention in order to define the International Resource Identifiers (IRIs) for the different levels of the FRBR model (Work, Expression and Manifestation) for the collected court decisions and legal opinions (Figure 4).
- <publication>: this block contains information about the publication source of the document; however, it is optional for our document types and it is not used.
- <classification>: this section is used to assign classification keywords to the document or part of it. As previously mentioned, legal opinions are published on the Legal Council's website followed by a set of descriptive keywords. These keywords are included within this section (an example is shown in Figure 5).
- <lifecycle>: this block lists the events that modify the document.
- <workflow>: this block lists the events that are involved with the legislative, parliamentary or judiciary process. In our case, these are the necessary procedural steps for the delivery of the decision or legal opinion (e.g., public hearing, court conference, decision publication).
- <analysis>: in case of court decisions, it contains the result of the decision and the qualification of the case law citations.
- <references>: a section that models references to other documents or ontology classes. We are using this block and the elements (e.g., TLCLocation, TLCPerson, TLCOrganization, TLCRoles, TLCEvent etc.) of the abstract ontological mechanism of the model (Top Level Classes) to denote references to concepts representing locations, persons, organizations, roles and events. For an overview of the ontological structure of the Akoma Ntoso model, one may refer to [34].
- <proprietary>: this block can be used for capturing local or proprietary metadata elements. For example, in the case of legal opinions, we are using this block to place gathered metadata that do not belong to the previous blocks (e.g., a short summary of the opinion, the unique identifier assigned to the opinion from the Diavgeia platform, information about their acceptance or rejection from the responsible body according to the law etc.)

```
<identification source="#openLawsGR">
    <FRBRWork>
        <FRBRthis value="/akn/gr/judgment/COS/2012/A5515/!main"/>
        <FRBRuri value="/akn/gr/judgment/COS/2012/A5515/"/>
        <FRBRalias value="ECLI:EL:COS:2012:1231A5515.03E3256" name="ECLI"/>
        <FRBRdate date="2012-12-31" name=""/>
        <FRBRauthor href="#councilOfState"/>
        <FRBRcountry value="gr"/>
    </FRBRWork>
    <FRBRExpression>
        <FRBRthis value="/akn/gr/judgment/COS/2012/A5515/ell@/!main"/>
        <FRBRuri value="/akn/gr/judgment/COS/2012/A5515/ell@"/>
        <FRBRdate date="2012-12-31" name=""/>
        <FRBRauthor href="#councilOfState"/>
        <FRBRlanguage language="ell"/>
    </FRBRExpression>
    <FRBRManifestation>
        <FRBRthis value="/akn/gr/judgment/COS/2012/A5515/ell@/!main.xml"/>
        <FRBRuri value="/akn/gr/judgment/COS/2012/A5515/ell@.xml"/>
        <FRBRdate date="2018-10-15" name="XMLConversion"/>
        <FRBRauthor href="#openLawsGR"/>
    </FRBRManifestation>
</identification>
```

Figure 4. The identification metadata block for a court decision from our dataset (A5515/2012 of the Council of State).

```
<classification source="#openLawsGR">
    <keyword eId="keyword_1" value="COMPANIES" showAs="COMPANIES" dictionary="none"/>
    <keyword eId="keyword_2" value="COURT DECISION" showAs="COURT DECISION" dictionary="none"/>
    <keyword eId="keyword_3" value="FINE" showAs="FINE" dictionary="none"/>
    <keyword eId="keyword_4" value="LABOUR INSPECTION" showAs="LABOUR INSPECTION" dictionary="none"/>
</classification>
```

Figure 5. The classification metadata block containing keywords for a legal opinion of our dataset.

5.3. Modeling Court Decisions

Decisions published from both Supreme Courts (Arios Pagos and Council of State) follow a similar structure and it is possible to take advantage of recurring linguistic patterns to identify the metadata blocks, the structural parts and several semantic elements, such as the names of judges and lawyers, litigant parties, etc. Akoma Ntoso provides a specific document type for the representation of court decisions, the Judgment document type. Figure 6 shows the XSD diagram for the <judgment> element.

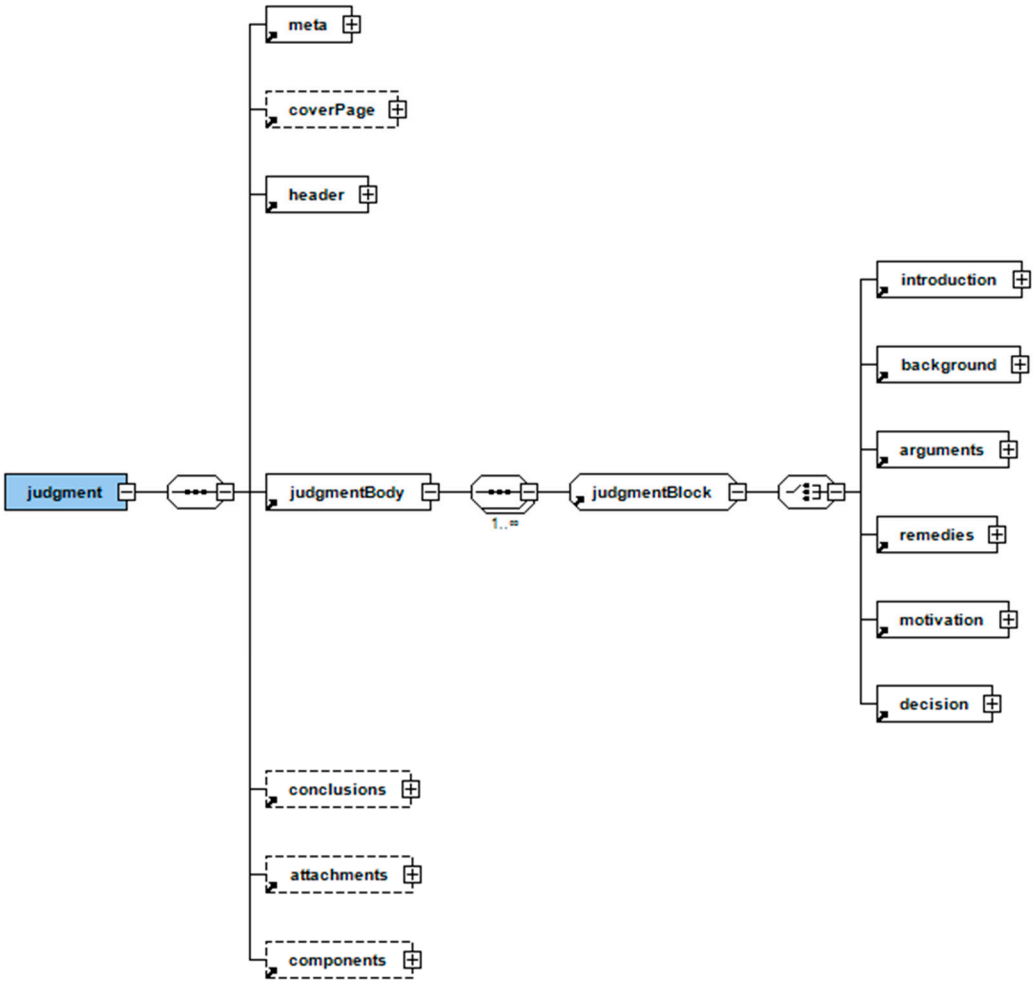


Figure 6. The XSD diagram of the Akoma Ntoso Judgment document type.

At the beginning of a decision of the two Supreme Courts there is usually text referring to the number of the decision and the name of the issuing court, the date of the court hearing, the composition of the court, the litigant parties and their lawyers. This part corresponds to the <header> element. The last part of the decision contains information about the location and date of the court conference and the signatures of the judges. This part is assigned to the <conclusions> section of the <judgment> element. The text between <header and <conclusions> belongs to the <judgmentBody> element, which represents the main body of the decision. The first paragraphs of this section, which usually contain

information about the previous decisions that are appealed and the trial procedure, are assigned to the <introduction> element. A standard phrase (e.g., "the court after studying the relevant documents, considered the law") is followed by a listing of points that explains how the judges reached the decision. This part is assigned to the <motivation> element, since it contains the argumentation of the court. The <background> element is dismissed, since the facts were analyzed in the decisions of lower courts that are appealed, and this analysis is not included in the decisions of the Supreme Courts. Some references to the facts may be found within the motivation list; however, it is impossible to separate them from the argumentation. The final part of the judgment body, usually beginning with the phrase "for these reasons", contains the decision of the court about the case and is included within the <decision> element. Several semantic elements of the Akoma Ntoso can be used to provide semantic information about concepts found in the decision text: <judge>, <party>, <lawyer>, <docDate>, <docProponent>, <role>, <ref> etc.

Compliance with ECLI

European Case Law Identifier (ECLI) [35] is an EU standard that defines a uniform identification scheme for European case law and a minimum set of (Dublin Core) metadata for the description of court decisions. In Greece, ECLI is currently implemented only for the Council of State. As this standard is designed to improve searchability, accessibility, and retrieval of case law, we considered it important to our modeling to be ECLI-compliant. In order to achieve such compliance, we had to define a mapping between Akoma Ntoso identifiers and ECLI, and a mapping between available Akoma Ntoso metadata elements and ECLI required metadata elements.

ECLI consists of five parts separated by colons: a) the word ECLI, b) the EU country code, c) the abbreviation of the court delivering the decision, d) the year of the decision and e) an ordinal number with a maximum of 25 alphanumeric characters and dots following a format decided by each member state. For the case of the Council of State, the fifth part of ECLI is formed by the following sequence of characters: month of the decision (2-digit format), day of the decision (2-digit format), number of the decision, a dot (.) and finally a string consisted of the year (2-digit format) of the notice of appeal to the court and its number. Figure 7 shows how ECLI can be formed from an Akoma Ntoso identifier. We should note that the number and year of the notice of appeal are available as metadata elements at the website of the Council of State. In our case, we take advantage of the <FRBRalias> element, which is a metadata element that can be used to denote other names of the document at the Work FRBR level; therefore, it is appropriate to handle the value of ECLI.

Moreover, Table 2 shows the mapping between the set of required Dublin Core metadata elements of ECLI and the respective metadata elements available in Akoma Ntoso.

Table 2. Mapping between ECLI required metadata elements and the respective Akoma Ntoso elements.

ECLI Metadata Element	Akoma Ntoso Metadata Element or Default Value
dcterms:identifier	The URL from which the text of the decision can be retrieved
dcterms:isVersionOf	The ECLI (<FRBRwork> → <FRBRalias>)
dcterms:creator	<FRBRWork> → <FRBRauthor>
dcterms:coverage	<FRBRWork> → <FRBRcountry>
dcterms:date	<FRBRWork> → <FRBRdate>
dcterms:language	<FRBRExpression> → <FRBRlanguage>
dcterms:publisher	<FRBRWork> → <FRBRauthor>
dcterms:accessRights	"public"
dcterms:type	"judgment"

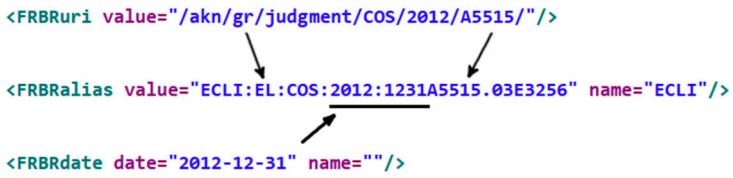


Figure 7. Mapping between Akoma Ntoso identifier and ECLI.

5.4. Modelling Legal Opinions

Akoma Ntoso provides some generic document types that could be used to model legal opinions (e.g., the Statement type can be used to represent formal expressions of opinion or will). However, the structure of legal opinions highly resembles that of court decisions; as a result, we decided to adopt again the Judgment document type as more appropriate for our modelling requirements. In most cases, a legal opinion can be analyzed in the following parts: At the beginning, there is some basic information regarding the opinion itself (opinion's number, session, composition of the Council, etc.). This part is assigned to the <header> element, which is followed by the main body of the legal opinion (<judgmentBody>). The first part of the main body is usually a summary of the question and it is assigned to the <introduction> block, while the second part, which is assigned to the <background> element, since it corresponds to the description of the facts, is usually a detailed background of the case that prompted the Public Administration body to submit the question. The next section usually cites the applicable provisions and the section that follows contains their interpretation relating to the question. Both sections are part of the <motivation> block, since they contain a detailed analysis of the legal arguments that led the members of the Council to express their opinion. The final part of the legal opinion's body cites the concluding opinion that the members of the Council express regarding the question under examination, which is assigned to the <decision> element. The document ends with the signatures (<conclusions> block).

6. Legal Language Processing

The availability of big datasets of legal documents facilitated the development of the research area of Legal Analytics, a term referring to the extraction of structured knowledge from unstructured legal texts. Moreno and Redondo note that "this task is very demanding on resources (especially manpower with enough expertise to train the systems) and it is also highly knowledge-intensive" [36]. The automation of such processes results in big sets of semantically annotated legal data, allowing their computational exploitation for more advanced tasks such as legal reasoning. In the Introduction section, we mentioned that legal documents are written in natural language and consequently legal language processing is based on traditional NLP tools and methods and we additionally noted that the language used in legal texts is more complex than in other domains. As Venturi shows in [37], legal language differs syntactically from ordinary language. The researcher states that " ... beyond the general NLP difficulties, the specificities of domain–specific features make the automatic processing of these kind of corpora a challenging task and demand specific solutions ... ".

In this section, we present our approach and the methods and techniques employed for the processing of our legal corpus in order to transform the available legal documents (after converting them to plain text where necessary and applying pre-processing to remove erroneous or irrelevant elements injected from this conversion process) into Open Data. This processing consists of three tasks: (a) legal structure identification, (b) legal references extraction/resolution and (c) named entity recognition.

6.1. Legal Structure Identification

In this work, we treat legal language as a Domain Specific Language (DSL), an approach also followed by Koniaris et al. for other legal documents types in [20]. DSLs are languages specifically tailored for the needs of a particular problem or application domain [38], in our case that being the

legal domain. As presented in the previous section about modeling of legal documents, court decisions and legal opinions usually follow a specific structure and the distinction of structural elements is based on recurrent linguistic patterns. In this sense, we consider the syntax rules that model the structural elements of these legal texts as the grammar of the DSL used in the legal sub-domain of judgments and legal opinions. The grammar specifies the sequence of tokens and words that make up a structurally valid text, and consequently it is possible to generate from that grammar a parser able to recognize valid texts and create an abstract syntax tree or graph. In order to define the grammar of the DSL describing the legal texts of our dataset, we created a base set of 75 randomly selected documents (25 legal opinions, 25 decisions of the Supreme Civil and Criminal Court and 25 decisions of the Council of State as shown in Table 3), which were used to extract the recurrent linguistic patterns and transform them into grammar rules.

Table 3. Word count analysis of the base set of documents used to extract the grammar rules of the DSL.

	Number of Documents	Number of Words	Mean Number of Words Per Document
Supreme Civil and Criminal Court	25	54278	2171
Council of State	25	50112	2004
Legal Council	25	86374	3454
Total	75	190764	2543

To achieve our goal, we used ANTLR v4 (ANother Tool for Language Recognition) [39], a popular lexer and parser generator. ANTLR is already used for parsing tasks in the legal domain, e.g., by the legal analytics company Lex Machina or by the Solon platform [20] for extracting structural and semantic information from legal texts. When the appropriate lexer and parser grammars are defined, ANTLR is able to generate:

- The lexer, which performs the lexical analysis (also known as tokenization) process on the input text. During this process a sequence of characters are converted into a sequence of tokens, which are then passed to the parser.
- The parser, which performs syntactic analysis of the input text, based on the tokens sequence of the lexical analysis and the parser grammar, transforming it to a structured representation, such as a parse tree.

The grammar rules of the lexer define the sequences of characters that are treated as the symbols of the DSL modelling judgments and legal opinions. Such symbols could be considered the numbers, sequences of characters of the Greek alphabet, special characters, special phrases found in these legal texts, etc. Table 4 shows some representative rules defined within the lexer grammar.

On the other hand, the parser's grammar rules define how the sequence of tokens of a legal text will be transformed into a parse tree. The rules respect our modelling of the legal documents that was presented earlier and are defined in such a way that by traversing the resulting parse tree, it is possible to identify the structural parts of the legal text.

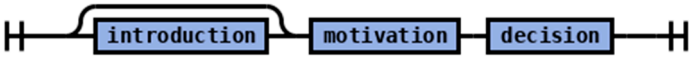
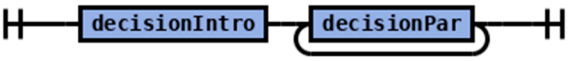
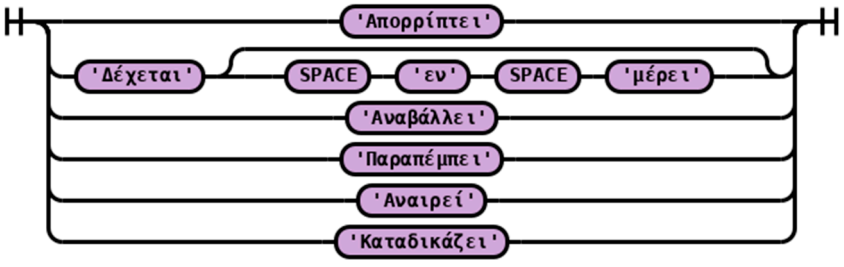
In Table 5 we include sample rules (since it is impractical to cite all of them) and the respective syntax diagrams (also known as railroad diagrams) for the DSL that describes court decisions. According to the grammar, parsing a court decision expects to match the rules that identify the following sections: header, judgment body, and conclusions. An analysis of the rule for identifying the body of the judgment shows that a decision can optionally contain the introduction section and it is required to match the rules for the identification of the motivation and decision sections. Furthermore, the decision part should consist of an introductory paragraph (decisionIntro rule that matches paragraphs starting with the special phrase "For these reasons" that courts use before the announcement of their decision and after the section that contains their argumentation, the respective rule is omitted) followed by

one or more paragraphs (decisionPar). These paragraphs optionally start with a verb or a phrase indicating the outcome of the decision followed by any sequence of tokens and one or more line breaks. The lexer rule matching the token representing the outcome of the decision is already explained in Table 4. Figure 8 shows a simplified model of a court decision translated in English and how textual patterns usually found in specific positions of the text are selected as templates that form the base for the creation of the grammar rules, which extract elements such as: number of decision, court name, motivation blocklist items, decision and outcomes, conclusions and signatures, etc.

Table 4. Examples of defined lexer grammar rules (in extended Backus-Naur form notation).

Lexer Rule	English Translation	Description
NEXT_LINE: '\n' | '\r'		Line break
NUM: [0–9]+		Numerical sequence
DOT: '.'		Dot
COMMA: ','		Comma
OUTCOME: Απορρίπτει' | 'Δέχεται' (SPACE 'εν' SPACE 'μέρει')? | 'Αναβάλλει' | 'Παραπέμπει' | 'Αναιρεί' | 'Καταδικάζει'	OUTCOME: 'Rejects' | 'Accepts' (SPACE 'partially')? | 'Postpones' | 'Refers' | 'Overturns' | 'Condemns'	Token representing the outcome of the court's decision (e.g., dismiss, approve, remit etc.)
HISTORY_HEADER: 'Σύντομο' SPACE 'Ιστορικό' | 'Ιστορικό' SPACE 'Ερωτήματος' | 'Σύντομο Ιστορικό' | 'Ιστορικό' | 'Ανάλυση' | 'Ιστορικό' SPACE 'της' SPACE 'υπόθεσης'	HISTORY_HEADER: 'Brief' SPACE 'History' | 'Question' SPACE 'History' | 'History' | 'Analysis' | 'Background of' SPACE 'the' SPACE 'case'	Phrases used to indicate the header of the legal opinions' section containing the history/background of the case that led the administrative body to submit an official question (e.g., History, Short History, History of the case, Analysis etc.).

Table 5. Sample defined rules and syntax diagrams for court decisions.

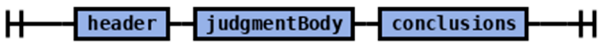
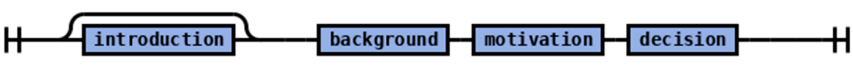
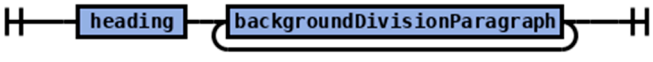
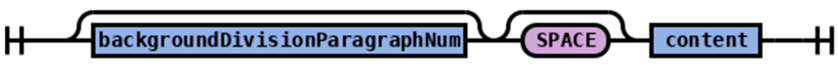
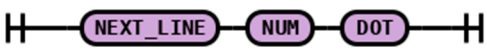
Rule	Syntax Diagram
judgment	
judgmentBody	
decision	
decisionPar	
outcome (English translation in Table 4)	

```
Decision Number 1294/2013

The Council of State

Consisted of the following judges: .....

The Court met at xx/xx/xxxx to judge the appeal of .... against the decision of ....

......

The Court thought in accordance with the law

1. Because ...

2. ......

3. ......

For these reasons

Rejects the appeal ....

Convicts the accused ...

Decision published in Athens at xx/xx/xxxx ....

Signed by .....
```

Figure 8. Simplified model of a court decision translated in English.

Similarly, following the presented modelling of legal opinions, we defined the grammar rules of the DSL that can be used to describe this type of legal documents. In Table 6, we show a short example of the rules used to identify the background section. As already explained, legal opinions have a similar structure to court decisions and consequently the Akoma Ntoso judgment type is used for their representation. The first rule is used to identify the header, the body, and the conclusions of the opinion. Parsing the body of the opinion, we expect to find the introduction (optionally), the background, the motivation, and the decision sections. The background section consists of a heading (a string consisting either of an upper-case Greek letter or a Latin number, followed by the token HISTORY_HEADER, which was described in Table 4, and a dot) and one or more paragraphs (backgroundDivisionParagraph). Each paragraph optionally starts with the numbering (a number followed by a dot—backgroundDivisionParagraphNum) and continues with the main content of the paragraph.

Table 6. Sample defined rules and syntax diagrams for legal opinions.

Rule	Syntax Diagram
judgment	
judgmentBody	
background	
Background Division Paragraph	
Background Division ParagraphNum	

6.2. Legal References Extraction

A significant feature of legal documents is that they often refer to each other through legal references, forming complex networks. Extraction and machine-readable annotation of legal references are essential tasks for enhancing the navigation of legal documents, ensure interoperability, facilitate information retrieval, and allow for graph representations of the legal corpus [40]. Agnoloni and Venturi elaborate on legal citations and note that "legal citations . . . should be processed separately through specifically trained parsers able to identify and extract the significant components of a textual legal citation (issuing authority, date, number, etc.) and transform them into a formalized link that can be resolved to the referred text" [41]. Moreover, the researchers underline that while there exist drafting recommendations for writing legal references, they are seldom followed and thus the variability of reference styles makes the automatic legal reference extraction a challenging task. Similar remarks can also be found in [40].

Following the above remarks, we decided to implement a parser for the detection of legal references in legal documents. Since, according to the existing literature (e.g., [42]), rule-based approaches seem able to undertake the task of legal references extraction with high precision and recall, we decided to adopt the same DSL approach presented in the previous section. Confirming our findings from the literature review that we conducted, Agnoloni and Venturi argue that reliable and accurate reference parsers typically rely on manually crafted rules based on regular expressions and grammars [41]. In order to create the set of grammar rules of the DSL, we again used the base set of the 75 randomly selected legal documents and we asked a legal expert to identify and markup legal references using the BRAT annotation tool [43]. We decided to work with references to court decisions, laws, legal opinions, and EU legislation, excluding references to administrative circulars and ministerial decisions since there is a much wider variation in the drafting styles of these citations. We adopt the distinction of references found in [44], according to which references can be classified as simple (comprised of a label, a number and/or a publication date or indirect anaphors to an earlier reference) or complex (multi-valued and multi-layered references) and as complete (when the reference includes the necessary information to identify the referred text) or incomplete (when the necessary information to identify the referred text needs to be inferred from the context). We provide some examples of references classification in Table 7, while an analysis of the extracted references from our base set of randomly selected legal documents can be found in Table 8.

Table 7. Examples of references classification.

Reference Classification			Example
Complete	Complex	Multi-valued	articles 12 and 13 of law 4386/2017
		Multi-layered	paragraph 2 of article 1 of law 4469/2017
	Simple		according to law 4554/2018
Incomplete	Complex	Multi-valued	the two previous paragraphs were modified
		Multi-layered	taking into account the second paragraph of the same article
	Simple		the same article states

Next, after inspecting the syntax and structure of the extracted references, we created the grammar rules that match textual legal citations. These rules match both simple and complex, but only complete references. At this phase, we decided not to markup incomplete references, since resolving them and identifying the value of the href attribute requires further processing and this task was out of the scope of our research. Figures 9 and 10 depict the resulting parse trees when the grammar rules match a complex reference to a structural unit of a law and a simple reference to a court decision. It is obvious

that by traversing the parse tree it is possible to identify the type of the reference and find the necessary information for resolving the referred text, taking advantage of the node labels.

Table 8. Analysis of manually extracted references from the base set of 75 randomly selected legal documents.

Type of Documents	# of Refs	# of Simple Refs	# of Complex Refs	# of Complete Refs	# of Incomplete Refs
Decisions of the Council of State	815	447	368	486	329
Decisions of the Supreme Civil and Criminal Court	718	319	399	492	226
Legal Opinions	1631	837	794	1051	580
Total	3164	1603	1561	2029	1135

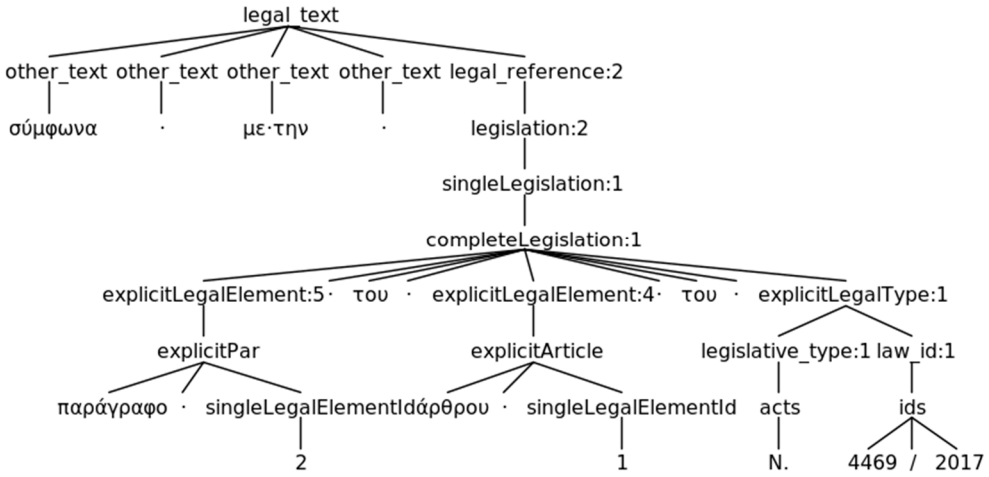


Figure 9. The resulting parse tree when a grammar rule matches the complex (multi-layered) and complete reference "paragraph 2 of article 1 of law 4469/2017".

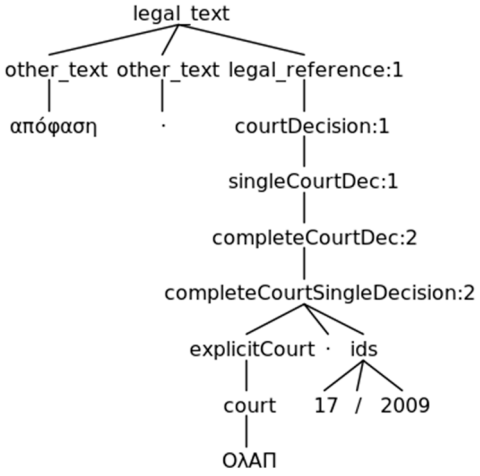


Figure 10. The resulting parse tree when a grammar rule matches a simple and complete reference to decision 17/2009 of the plenary of the Supreme Civil and Criminal Court.

Table 9 includes the English translated parser and lexer rules that match the legal reference "paragraph 2 of article 1 of law 4469/2017" for which the parse tree is shown in Figure 9. Actually, the rules are much more complex; however, for reasons of clarity and simplicity we include a simplified version based on the specific legal reference that is used as an example. According to the rules, a reference falling into the category of complete legislation is expected to be structured as "explicitLegalElement of explicitLegalElement of explicitLegalType". The explicitLegalElement rule captures text related to

structural units of the referred legislative document (e.g., chapters, articles, paragraphs, alineas etc.), while explicitLegalType matches the type of the document (e.g., act, presidential decree etc.). The string "paragraph 2" matches the rule used to detect references to paragraphs (explicitPar), while "article 1" the rule detecting references to articles (explicitArticle) and "law 4469/2017" the rule used to detect references to acts (legislative_type SPACE? (OF SPACE)? law_id, where legislative_type matches "law" and law_id matches "4469/2017"). In this example, the legal citation refers to single structural units (singleLegalElementId rule is matched) of the document, however other references could cite multiple structural units or even a range of them. Similarly, other grammar rules are used to detect the remaining identified textual patterns that are followed by the authors of court decisions and legal opinions when referring to other legal documents.

Table 9. Parser and lexer grammar rules (translated to English) that match the legal reference "paragraph 2 of article 1 of law 4469/2017".

Rule Name	Rule Content
completeLegislation	explicitLegalElement OF SPACE explicitLegalElement OF SPACE explicitLegalType
explicitLegalElement	explicitPart|explicitChapter|explicitArticle|explicitPar| explicitSubPar|explicitCase|explicitAlinea|explicitPoint
explicitPar	PAR_TEXT SPACE? (multipleLegalElementIds|singleLegalElementId|range_id)
explicitArticle	ARTICLE_TEXT SPACE? (multipleLegalElementIds|singleLegalElementId|range_id)
singleLegalElementId	NUM|GREEK_NUM|TEXTUAL_NUM
explicitLegalType	legislative_type SPACE? (OF SPACE)? law_id|law_id SPACE legislative_type
legal_id	ids | ALL_CHARS SLASH NUM
ids	NUM SPACE? SLASH SPACE? NUM
legislative_type	acts | presidential_decree | compulsory_law | decree_law | decree | royal_decree
PAR_TEXT	'paragraph' | 'paragraphs' | 'par.' | '§' | '§§'
ARTICLE_TEXT	'article' | 'articles' | 'art.' | 'ar.'
NUM	[0–9]+

6.3. Named Entity Recognition

Named Entity Recognition (NER) refers to the identification of words or phrases denoting entities belonging to certain categories, such as locations, persons or organizations. Applying NER to legal texts is an important task, since it is a process able to assign semantic meaning to legal entities (e.g., juries, courts, lawyers) that hold a significant role in the legal process. In this work, a multi-domain Named Entity Recognizer [45] developed by the Institute for Language and Speech Processing of the "Athena" Research Center was employed. The Named Entity Recognizer follows a machine learning approach (based on single-level maximum entropy), it can be adapted to multiple domains and it is able to recognize entities belonging to the classes PERSONS, LOCATIONS, and ORGANIZATIONS. The researchers of the ILSP trained their model using a set of 35 tagged legal texts that we provided. Then, they used the Named Entity Recognizer to extract the recognized entities from all the files of our legal dataset of court decisions and legal opinions and provided us with the tagged entities within the legal texts. However, further training of the model seems to be needed to improve its accuracy. Some other entities with semantic meaning, such as the number of the decision or opinion, the name of the legal body, the issue date, the date of the court hearing, etc. can be found at fixed positions within the legal texts and rule-based processing (use of regular expressions in our case) is able to identify them.

Semantic information is attached to these entities using the available Akoma Ntoso elements, such as <docProponent>, <docType>, <docNumber>, <date> etc. Table 10 sums up the semantic elements that are detected during the transformation process and the method for their detection. Figure 11 shows an example of marked-up entities tagged with the Akoma Ntoso elements <person> and <date> and the mechanism that the standard provides, through the <references> section of the <meta> element and the Top-Level Classes, for their semantic connection to external (fictional in our case) ontologies.

Table 10. Semantic elements of the Akoma Ntoso standard used in the transformed documents.

Element	Recognition Method	Remarks
<doctype>	ANTLR Grammar	The type of document (e.g,. legal opinion)
<docProponent>	ANTLR Grammar	The issuing authority (e.g., Council of State)
<docNumber>	ANTLR Grammar	The number of the decision or opinion
<person>	ILSP NER component	Elements connected to the abstract Top Level Classes (TLCPerson, TLCOrganization, TLCLocation) with the refersTo attribute
<organization>	ILSP NER component	
<location>	ILSP NER component	
<date>	Regular expressions	Court conference date, public hearing date and decision publication date. Connected with the TLCEvent classes and the steps of the judiciary process of the <workflow> metadata element
<outcome>	ANTLR Grammar	The outcome of the decision, found within the <decision> element (e.g., dismiss, approve, remit etc.)

```
...
<meta>
...
    <references source="#openLawsGR">
        <TLCPerson eId="Metaksia_Androbitsanea" href="/akn/ontology/person/gr/Metaksia_Androbitsanea" showAs="Μεταξία Ανδροβιτσανέα"/>
        <TLCEvent eId="opinionSignatureDate" href="/akn/ontology/event/gr/opinionSignatureDate" showAs="Ημερομηνία θεώρησης γνωμοδότησης"/>
    </references>
</meta>

...

<conclusions>
    <p>Θεωρήθηκε</p>
    <p>Αθήνα, <date refersTo="opinionSignatureDate" date="2016-06-20">20-6-2016</date></p>
    <p>Η Πρόεδρος του Τμήματος Ο Εισηγητής</p>
    <p><person refersTo="#Metaksia_Androbitsanea">Μεταξία Ανδροβιτσανέα</person> Γεώργιος Γρυλωνάκης</p>
    <p>Αντιπρόεδρος Ν.Σ.Κ. Πάρεδρος Ν.Σ.Κ.</p>
</conclusions>

...
```

Figure 11. Example of extracted semantic entities (person and date) and their connection to external ontological classes.

7. Components' Integration and Documents Transformation

Following the architecture shown in Figure 1, all distinct components were integrated into a functional system developed using the python programming language. The system was setup in a server equipped with a 4-core Intel Xeon E3-1220v6 CPU @3,00GHz and 8GB RAM, which can be considered a low-end solution in terms of computational power. After collecting the available documents from the selected legal sources, transforming them into plain text and applying necessary pre-processing, the resulting set of 142398 text files was used to feed the legal language processor pipeline.

Using parallel processing (with the GNU Parallel tool [46]) we were able to take advantage of all CPU cores at the same time for the processing of the legal texts. The process is as follows: each file initially passes through the Lexer and the resulting tokens are given as input to the Parser. The resulting parse tree is enriched with NER tags for the extracted entities, and the output is provided as input to

the XML encoder. This module traverses the parse tree and takes advantage of the nodes' labels in order to identify the parts that form the elements of the XML file. Especially for metadata elements, metadata collected from the legal sources are combined with the necessary nodes of the parse tree in order to create the content of the <meta> element. Finally, the XML file is validated against the Akoma Ntoso Schema.

The transformation process lasted for almost 15 d. The average time for the transformation of each document was 29.91 s. However, there is room for even better performance by using Java instead of Python, since the Java target of ANTLR can be up to 20 times faster than the Python target and parsing occupies almost 95% of each document processing time. Finally, 127,061 XML files were found to be valid Akoma Ntoso documents, a number that accounts for a percentage of 89.23% of the total number of documents (91.47% of judgments of the Council of State, 83.85% of judgments of the Supreme Civil and Criminal Court and 77.30% of legal opinions). Most validation failures are found for documents published in specific years (e.g., 1995–1998 and 2009 for the Council of State and 2014–2017 for the Supreme Civil and Criminal Court), for which there are no documents in our training set or in the opposite case, the selected documents were not representative of the drafting style followed during these periods. Consequently, a more careful selection of the training set, instead of the random selection that we adopted, could lead to even better results.

The source code of our system along with a sample collection of automatically generated XML files are available in a Github repository: https://github.com/OpenLawsGR/judgments2AKN. Moreover, in Table 11 we provide a short representative example of a court decision (part of the text was omitted due to space limitations) marked up in the Akoma Ntoso standard.

Table 11. Example of a court decision (some parts are omitted due to space limitations) automatically marked up in Akoma Ntoso. Different colors are used to denote the distinct structural and semantic elements.

Sample Decision's Akoma Ntoso Markup
<?xml version = '1.0' encoding = 'UTF-8'?>
<akomaNtoso xmlns:xsi = "http://www.w3.org/2001/XMLSchema-instance" xmlns =
"http://docs.oasis-open.org/legaldocml/ns/akn/3.0" xsi:schemaLocation =
"http://docs.oasis-open.org/legaldocml/akn-core/v1.0/os/part2-specs/schemas/akomantoso30.xsd">
<judgment name = "decision">
<meta>
<identification source = "#openLawsGR">
<FRBRWork>
<FRBRthis value = "/akn/gr/judgment/SCCC/2013/1294/!main"/>
<FRBRuri value = "/akn/gr/judgment/SCCC/2013/1294/"/>
<FRBRdate name = "" date = "2013-10-31"/>
<FRBRauthor href="#SCCC"/>
<FRBRcountry value = "gr"/>
</FRBRWork>
<FRBRExpression>
<FRBRthis value = "/akn/gr/judgment/SCCC/2013/1294/ell@/!main"/>
<FRBRuri value = "/akn/gr/judgment/SCCC/2013/1294/ell@"/>
<FRBRdate name = "" date = "2013-10-31"/>
<FRBRauthor href = "#SCCC"/>
<FRBRlanguage language = "ell"/>
</FRBRExpression>
<FRBRManifestation>
<FRBRthis value = "/akn/gr/judgment/SCCC/2013/1294/ell@/!main.xml"/>
<FRBRuri value = "/akn/gr/judgment/SCCC/2013/1294/ell@.xml"/>
<FRBRdate date = "2019-11-05" name = "XMLConversion"/>
<FRBRauthor href = "#openLawsGR"/>
</FRBRManifestation>

Table 11. *Cont.*

Sample Decision's Akoma Ntoso Markup

```
</identification>
<lifecycle source = "#openLawsGR">
<eventRef date = "2019-11-05" source = "#original" type = "generation"/>
</lifecycle>
<workflow source = "#openLawsGR">
<step by = "#SCCC" date = "2013-10-31" refersTo = "#courtConferenceDate"/>
<step by = "#SCCC" date = "2013-10-31" refersTo = "#decisionPublicationDate"/>
<step by="#SCCC" date = "2013-10-08" refersTo = "#publicHearingDate"/>
</workflow>
<references source = "#openLawsGR">
<original eId = "original" href = "/akn/gr/judgment/SCCC/2013/1294/ell@" showAs = "Original"/>
<TLCOrganization eId = "dikastirio_areios_pagos" href =
"/akn/ontology/organization/gr/dikastirio_areios_pagos" showAs = "ΤΟ ΔΙΚΑΣΤΗΡΙΟ ΤΟΥ ΑΡΕΙΟΥ
ΠΑΓΟΥ"/>
<TLCPerson eId = "Grigorios_Koytsopoylo" href = "/akn/ontology/person/gr/Grigorios_Koytsopoylo" showAs
= "Γρηγόριο Κουτσόπουλο"/>
<TLCPerson eId = "Xaralampos_Athanasios" href="/akn/ontology/person/gr/Xaralampos_Athanasios"
showAs="Χαράλαμπου Αθανασίου"/>
<TLCOrganization eId = "trimelis_plimmeleiodikeio_Athina" href =
"/akn/ontology/organization/gr/trimelis_plimmeleiodikeio_Athina" showAs="Τριμελούς Πλημμελειοδικείου
Αθηνών"/>
<TLCEvent eId = "publicHearingDate" href = "/akn/ontology/event/gr/publicHearingDate" showAs =
"Ημερομηνία δημόσιας συνεδρίασης"/>
<TLCEvent eId = "decisionPublicationDate" href = "/akn/ontology/event/gr/decisionPublicationDate" showAs
= "Ημερομηνία δημοσίευσης απόφασης"/>
<TLCEvent eId = "courtConferenceDate" href = "/akn/ontology/event/gr/courtConferenceDate" showAs =
"Ημερομηνία διάσκεψης"/>
</references>
</meta>
<header>
<p>Αριθμός<docNumber>1294/2013</docNumber></p>
<p>
<docProponent>
<organization refersTo = "#dikastirio_areios_pagos">ΤΟ ΔΙΚΑΣΤΗΡΙΟ ΤΟΥ ΑΡΕΙΟΥ
ΠΑΓΟΥ</organization>
</docProponent>
</p>
<p>ΣΤ' Ποινικό Τμήμα</p>
<p>Συγκροτήθηκε από τους Δικαστές: <person refersTo = "#Grigorios_Koytsopoylo">Γρηγόριο
Κουτσόπουλο</person>, Αντιπρόεδρο Αρείου Πάγου, ... </p>
<p>Συνήλθε σε δημόσια συνεδρίαση στο Κατάστημά του στις <date refers To = "publicHearingDate"
date = "2013-10-08">8 Οκτωβρίου 2013</date> με την παρουσία του Αντεισαγγελέα του Αρείου Πάγου
Νικολάου Παντελή (γιατί κωλύεται η Εισαγγελέας) και του Γραμματέως <person refersTo =
"#Xaralampos_Athanasios">Χαράλαμπου Αθανασίου</person>, για να δικάσει την αίτηση του
αναιρεσείοντος - κατηγορουμένου, Σ. Ν. του Ι., κατοίκου ... , που δεν παραστάθηκε στο ακροατήριο,
περί αναιρέσεως της <ref href =
"/akn/gr/judgment/MagistrateCourtAthens/2012/60989/!main">60989/2012 αποφάσεως του <organization
refers To = "#trimelis_plimmeleiodikeio_Athina">Τριμελούς Πλημμελειοδικείου
Αθηνών</organization></ref>.</p>
</header>
<judgmentBody>
<introduction>
<p>Το Τριμελές Πλημμελειοδικείου Αθηνών, με την ως άνω απόφασή του διέταξε όσα λεπτομερώς
αναφέρονται σ' αυτή και ο αναιρεσείων - κατηγορούμενος ζητεί την αναίρεση αυτής, για τους λόγους
που αναφέρονται στην από 1 Απριλίου 2013 αίτησή του αναιρέσεως, η οποία καταχωρίστηκε στο οικείο
πινάκιο με τον αριθμό 620/13.</p>
</introduction>
```

Table 11. *Cont.*

Sample Decision's Akoma Ntoso Markup
<motivation> Α φ ο ύ ά κ ο υ σ ε Τον Αντεισαγγελέα που πρότεινε να απορριφθεί ως ανυποστήρικτη η προκείμενη αίτηση.   ΣΚΕΦΤΗΚΕ ΣΥΜΦΩΝΑ ΜΕ ΤΟ ΝΟΜΟ **<blockList eId = "motivation_list_1">** **<item eId = "motivation_list_1__item_1">** <num>1.</num> … **</item>** **<item eId = "motivation_list_1__item_2">** <num>2.</num> Στην προκειμένη περίπτωση, όπως προκύπτει από το υπό ημερομηνία 18 Ιουνίου 2013 αποδεικτικό επίδοσης της επιμελήτριας Δικαστηρίων Εισαγγελίας του Αρείου Πάγου .ο αναιρεσείων κλητεύθηκε από τον Εισαγγελέα του Αρείου Πάγου νόμιμα και εμπρόθεσμα, για να εμφανισθεί στη συνεδρίαση που αναφέρεται στην αρχή της απόφασης αυτής, πλην όμως δεν εμφανίσθηκε κατ" αυτήν και την εκφώνηση της υπόθεσης ενώπιον του Δικαστηρίου τούτου. Κατά συνέπεια, η υπό κρίση αίτηση αναίρεσης πρέπει να απορριφθεί και να επιβληθούν στον αναιρεσείοντα τα δικαστικά έξοδα (**<ref href = "/akn/gr/act/presidentialDecree/1986/258/!main#art_583__par_1">**άρθρο 583 παρ. 1 Κ.Ποιν.Δ**</ref>**.) **</item>** **</blockList>** </motivation> <decision> ΓΙΑ ΤΟΥΣ ΛΟΓΟΥΣ ΑΥΤΟΥΣ   **<outcome>**Απορρίπτει**</outcome>** την από 1-4-2013 αίτηση του Σ. Ν. του Ι … ..   **<outcome>**Καταδικάζει**</outcome>** τον αναιρεσείοντα στα δικαστικά έξοδα που ανέρχονται σε διακόσια πενήντα (250) ευρώ. </decision> </judgmentBody> <conclusions> Κρίθηκε και αποφασίσθηκε στην Αθήνα στις **<date refersTo = "courtConferenceDate" date = "2013-10-31">**31 Οκτωβρίου 2013**</date>** . Και   Δημοσιεύθηκε στην Αθήνα, σε δημόσια συνεδρίαση στο ακροατήριό του, στις **<date refers To = "decisionPublicationDate" date = "2013-10-31">**31 Οκτωβρίου 2013**</date>**.   Ο ΑΝΤΙΠΡΟΕΔΡΟΣ Ο ΓΡΑΜΜΑΤΕΑΣ </conclusions> </judgment> </akomaNtoso>

8. Evaluation Results

In order to evaluate the performance of our approach, we created a second set of 75 randomly selected documents, 25 from each legal body, as shown in Table 12. These documents were automatically transformed into Akoma Ntoso XML files following our approach, and a legal expert was asked to manually assess the performance of our method for legal structure identification and legal citations extraction and resolution.

To evaluate the quality of the structural markup, we followed the same approach as Sannier et al. in [12]: we classified the main structural elements (header, introduction, background, motivation, decision, conclusions) as fully correct (FC) if no manual corrections were needed, partially correct (PC) if the elements were present but corrections were needed and missed (M) if they were absent. The results of the evaluation of the basic structural elements' annotation are shown in Table 13. In our case, where the legal texts follow the Judgment type structure, there is a relatively small number of main structural elements in each document, in contrast to legal acts, where there exist usually dozens of chapters, articles, paragraphs, etc. The value of the metric Q that Sannier, et al. proposed (Q = FC/(FC+PC+M)), which denotes the proportion of fully correct elements over the total number

of elements is 94.24%, meaning that the additional manual effort required to fix problems of the automated process related to the structure of the generated files is considerably low, taking also into account that partially correct elements (4.15%) require less effort to fix than missed elements (1,61%). Consequently, our approach performs extremely well regarding legal structure identification, even for documents like legal opinions, where drafting rules are much more loose and each author follows their own style.

Table 12. Word count analysis of the set of documents used for evaluation purposes.

	Number of Documents	Number of Words	Mean Number of Words Per Document
Supreme Civil and Criminal Court	25	51914	2076
Council of State	25	39840	1593
Legal Council	25	86305	3452
Total	75	178059	2374

Table 13. Evaluation results regarding structural elements identification.

	# of Basic Structural Elements	# of FC Structural Elements	# of PC Structural Elements	# of M Structural Elements
Supreme Civil and Criminal Court	148	139	8	1
Council of State	150	150	0	0
Legal Council	136	120	10	6
Total	434	409	18	7

In order to evaluate the quality of our legal citations' extraction and resolution method, we considered as correct only the detected legal references that were resolved correctly. Consequently, partially detected references or references with a wrong or partially correct value of the href attribute were classified as erroneous. We should note again that in this research effort we did not work with incomplete references and as a result such references were not taken into account in the evaluation process. The results of the assessment are shown in Table 14, while Table 15 contains the respective values of Precision, Recall, and F1 Score.

Table 14. Evaluation results regarding legal references identification and resolution.

	# of Simple Refs	# of Correct Simple Refs	# of Complex Refs	# of Correct Complex Refs	# of Total Refs	# of Correct Refs	# of False Positives
Supreme Civil and Criminal Court	138	121	339	252	477	373	2
Council of State	105	99	210	165	315	264	3
Legal Council	243	216	633	503	876	719	1
Total	486	436	1182	920	1668	1356	6

Table 15. Precision, recall, and F1 Score for legal references identification and resolution.

	Precision	Recall	F1 Score
Supreme Civil and Criminal Court	99.47%	78.20%	87.56%
Council of State	98.88%	83.81%	90.72%
Legal Council	99.86%	82.08%	90.10%
Total	99.56%	81.29%	89.50%

Several interesting remarks emerge from the inspection of these tables. First of all, the mean number of complete legal references is 22.24 references per document and considering that there are also many incomplete references, we confirm that court decisions and legal opinions are highly interconnected with each other and with other legal documents. It is obvious that in case of manual processing, a considerable workload is required for legal references' detection and markup. Legal opinions contain almost 2.5 times more references than court decisions, which is expected when considering the length of each document's type. Our grammar seems to perform very well regarding simple references, since almost 90% of them were successfully detected and resolved. The performance drops significantly for complex references to almost 78%, since there are much more citation styles for this category, while in total 81.29% (a percentage corresponding to the Recall metric) of all complete references were correctly extracted and resolved. Performance varies slightly between different legal bodies and documents from the Council of State show less variation in citing patterns than documents from the other two bodies. We should note that in contrast to legislative drafting, there are usually no official instructions for legal referencing when drafting court decisions and legal opinions, and as a result several different referencing styles can be found in these texts. Moreover, we noticed that the identification and resolution process was often failing due to typos or misspelling (e.g., mix of look-alike characters of the Latin and Greek alphabet or injection of unnecessary punctuation marks). Other reasons for failures in the evaluation process include the existence of extremely complex or ambiguous references (sometimes even difficult for a non-legal expert to resolve) and the use of rare citation patterns. The performance would be much higher if we had taken into account partially detected and resolved references. Considering that our base set that was used to identify the referencing patterns and create the grammar rules consisted of only 75 documents, we believe that the performance of the approach is satisfying and reduces significantly human effort to markup the rest of the legal references. A larger set of documents would have revealed more citation patterns, allowing for a more extended grammar and probably better evaluation results, since another frequent reason for failures was that some referencing patterns in the evaluation set were not present in the base set. At the same time, a larger base set would require higher workload and more time to inspect the documents, detect the patterns and create the set of grammar rules. Finally, we should highlight that the approach produced an extremely small number of false positive results (only six), which is the reason for the high value of the Precision metric.

9. Discussion

As discussed in the Introduction section, it is a common belief among researchers that releasing available government data as Open Data is a prerequisite for gaining the expected benefits. However, a long way is still lying ahead regarding data liberation, since in 2016 only 10% of data published at a global level were estimated to get released in an open format, a fact that significantly limits the potential for reuse and exploitation [47]. The situation seems even more disappointing when inspecting the application of the Open Data principles in the legal domain. According to the 4th edition of the Open Data Barometer [48], a global report of the World Wide Web Foundation on Open Government Data, while legislative datasets are considered important for government accountability, datasets related to legislation consist only 3% of open datasets published by all governments. Six years

ago, Marsden was stressing the need for Big Open Legal Data that would contribute to better access to legislation and better governance [49], however Open Data surveys like the Open Data Barometer show that we have not yet achieved this goal.

On the other hand, simple opening of data is not enough for creating value from their exploitation. As Peristeras notes at the foreword of the book "Open Data Exposed" [50], there is additionally the need for policies that guarantee data quality, promote interoperability, and ensure compliance to established standards.

Working towards surpassing those problems that hinder the exploitation of available legal documents in digital format, we have so far focused our research on developing a methodology for automating the transformation of the existing corpus of legal documents into Open Data, taking advantage of the structural and semantic features of legal texts and the available natural language processing tools. Such an automation is necessary in order to reduce the need for human involvement in the process, saving valuable resources. As van Opijnen has noted in [51], before solving more challenging problems of the legal informatics domain, such as legal reasoning, we should fix the architectural flaws related to legal information publishing (e.g., formats, standardization of resources identification, interconnection, etc.).

Answering the dilemma between rule-based and machine learning approaches, we decided to proceed with a rule-based implementation. While machine learning is much more attractive to the academic community for information extraction related tasks, rule-based systems are easier to comprehend, maintain, incorporate domain knowledge and debug and are more popular between industrial vendors, even if recall and precision are often lower and rules' development is a tedious manual task [52].

In this context, we believe that our work makes several contributions:

Unlike other similar efforts found in the relevant literature, we wanted to incorporate our approach into well-established theoretical Open Data frameworks that are designed to promote data exploitation. The proposed methodology of our project is aligned with the ecosystem approach from the publisher perspective, an approach which, as already discussed, is considered appropriate to create value from Open Data, while at the same time contributes to higher Open Data maturity in practices of both data providers and consumers [23]. More specifically, in Section 2 we showed how the steps of our methodology are aligned with two ecosystem models: the extended Open Data lifecycle and the BOLD analytics lifecycle. Our approach is designed to cover the specific characteristics of documents produced within the Greek legal system; however, it can be easily adapted to cover the features and peculiarities of other legal systems. Even if one may argue that the Greek legal system is of little interest to the global legal research community due to the small size of the population that is able to read legal documents written in the Greek language, such projects are of high importance and provide useful insights, since according to Peristeras the application of the Open Data movement outside the Anglo-Saxon context "revealed specificities and special characteristics based on cultural, institutional and organisational factors that need to be carefully considered" [50].

Standardization and interoperability are important requirements for promoting innovation in the big data ecosystem that derives from public sector generated information [53]. Despite the existing efforts for legal documents standardization, the level of interoperability in this field is still low, while the poor level of semantic information attached to documents prevents conceptual interconnection and sharing of information [54]. Our work, contrary to a number of other research efforts that use custom solutions and schemas for legal data representation, builds on well-established international standards of the legal semantic web, such as the Akoma Ntoso document model, the Akoma Ntoso naming convention and ECLI. Akoma Ntoso provides flexibility and allows for local customization, while at the same time ensuring interoperability by defining common building blocks that can be applied to different judiciary systems.

Sannier et al. [12] have underlined the lack of research focusing on automatic structural markup of legal documents at large scale and stressed the need for more efforts in this field. Our work comes to fill

this gap, resulting at a large volume of court decisions and legal opinions being automatically marked up in the Akoma Ntoso format using legal language processing techniques. In total, more than 140,000 documents were automatically processed and marked up in XML in approximately 15 d (as already discussed, this performance could be much better if the Java instance of ANTLR was used) with a percentage of valid XML files around 90%, while manual markup would require several man-months or even man-years and would result at a much higher cost (we should note that the server used for the processing of the legal texts costs less than 1800 Euros). Moreover, we conducted an evaluation of the approach that treats legal language as a Domain Specific Language, since Koniaris et al. [20], who also employed the DSL approach in a similar research, do not provide evaluation results on the quality of the transformation. The quality evaluation of the markup that we performed confirms the conclusion of Sannier et al. [12], according to which automatic approaches for the markup are effective and the amount of required manual effort is relatively low. The evaluation results that we presented showed that around 95% of the effort needed to markup the structural parts of the legal documents can be fully undertaken by the NLP processing software, while only 1% of the effort is exclusively manual and around 4% is a combination of manual and machine work. Similarly, more than 80% of the effort needed to successfully markup complete legal references can be accomplished automatically, while a considerable portion of citations are detected but not resolved. As a result, the automated completion of the markup tasks takes on average around half a minute and only minimal manual intervention is needed to correct software failures, while our experience showed that marking up legal documents manually with an Akoma Ntoso editor requires up to 30 min depending on the length of the document. We should not forget that the designed pipeline includes also other time-consuming steps, such as preprocessing and validation, which when performed manually increase even more the required processing time. However, more research is needed to fully automate the process and a semi-automated approach with low manual intervention seems currently more realistic. For the time being, our work confirmed that automation cannot be fully accurate mostly due to the complexity and variation of legal language used and manual work remains inevitable, a conclusion also highlighted in [12].

Our work confirmed most of the causes identified in [12] as barriers for the automation of the transformation process: heterogeneity of drafting styles, rare patterns for cross-references, incorrect use of characters that look identical, typographical errors, line breaks used for layout purposes etc. Since most of the above problems are not trivial to solve and require time investment in the pre-processing stage, confirming their impact to the quality of transformation is beneficial, as it highlights to candidate Open Data producers the need for setting up a strategy to tackle with them.

Transformation of data originally available in not machine-processable formats into open format and creation of descriptive metadata is a process that requires considerable costs that are part of the so-called "adaptation costs" [50]. According to Zuiderwijk et al. [33] the creation and maintenance of metadata is time-consuming and requires high investments and costs. Designing methodologies and tools for the automation of the process highly contributes to the removal of these Open Data barriers.

Court decisions are important base materials for legal professionals and other involved stakeholders in the legal domain, however their re-use as open data is still limited mainly due to technical barriers, such as documents' format [55]. The EU Council, in its Conclusions on the Online Publication of Court Decisions, urge for the publication of court decisions in machine readable format, along with their metadata, a practice that could facilitate data re-use [27]. Our work is one of the few in the research literature focusing on the publication of court decisions as Open Data. Additionally, as far as we know, it is the only research effort considering transformation of legal opinions (a legal document type with its own distinct characteristics), which are important documents for the application of the law, into an open format.

It is worth mentioning that one limitation of our work is that the evaluation results are not validated against other automated transformation approaches. Such a task would require the application of other

proposed parsing and processing methodologies on the same set of legal documents; unfortunately, this is not currently feasible due to the following reasons:

Approaches adapted to other legal systems and languages cannot be directly applied to our dataset, due to the peculiarities and the specific features of the Greek legal domain (including the unique characteristics of Greek court decisions and legal opinions) and the language used.

Most research works do not provide enough details to reproduce the adopted approaches. Especially in the case of rule-based systems (being the majority of automated markup efforts in the literature), the performance highly depends on the completeness of the employed set of rules.

In the Related Work section, we referenced two research projects oriented towards Greek legal documents that can be applied in texts written in the Greek language. In [20], neither the source code of the "Solon" platform nor the set of grammar rules used for legal text parsing are published. On the other hand, while the legal parser presented in [19] is available on Github, it is used to parse documents of the Government Gazette and detect their structure and cannot be used to extract structural elements of court decisions or legal opinions. Moreover, legal references in these documents follow a much stricter drafting style, due to rules set by a special committee, while judges are free to write legal citations according to their personal taste. As a result, a direct comparison of the performance of legal citations' extraction between the two approaches would not make sense.

We should stress at this point that it was not our primary goal to propose the optimal parsing approach for solving the automated transformation problem. Filling the gap that Sannier et al. underlined in [12], we wanted to validate the hypothesis that automated methodologies can be applied with satisfactory results in the large scale processing of legal documents, highly reducing the need for manual effort. As already discussed, the presented evaluation results successfully confirmed this hypothesis.

10. Conclusions

The legal domain is a domain where large volumes of legal documents are produced; however, usually the publishing bodies do not follow the Open Data paradigm, preventing data sharing and reuse and consequently limiting the potential benefits that could arise. We strongly believe that Legal Open Data publishing should be a responsibility and a priority of each government and we agree with the opinion of Charalabidis et al. who state that "rather than liberating data ex-post, the processes of data generation have to be open by design in order to minimize the cost of making them available to relevant stakeholders" [47]. Nevertheless, until the open-by-design model is adopted by the governments and the legal publishing bodies, we cannot but direct our research efforts on facilitating the ex-post transformation of raw legal documents into Open Data. Moving in this direction, we presented a methodology adapted to the ecosystem approach for the automatic transformation of court decisions and legal opinions into Open Data using legal language processing techniques.

In the future, we plan to extend our grammar, adding more legal texts in the base set used to extract the DSL rules and study how the markup quality is influenced. One should expect the performance to increase, since more language patterns are identified, until reaching a saturation level where adding more texts only slightly improves the results, as Sannier et al. note in [42]. Moreover, in order to attach more accurate semantic information to the legal text, we intend to investigate possible improvements of the applied NER approach or the application of other approaches, such as the one implemented by Angelidis et al. [56] that seems to provide very good results for Greek legal texts. Towards this direction, we are currently working on distinguishing the different roles involved in the judiciary process (e.g., judge, lawyer, defendant, appellant), usually present at the header of the legal documents.

Another direction for further research is the application of the ecosystem approach from the consumer side. As already discussed, data opening does not guarantee data exploitation. To achieve this, we plan to collaborate with the legal community (e.g., lawyers and juries) in order to understand the needs of the involved stakeholders and propose a strategy for the exploitation of available Legal Open Data. Generating value from Open Data is a complex process and requires the collaboration

of all involved actors. In [57], we have already identified some usage scenarios that take advantage of legal open datasets and we intend to build on top of this work. Working in this direction, we also want to implement a web platform that would provide enhanced access (employing visualization techniques and providing a REST API) to the generated open dataset.

Author Contributions: Conceptualization, J.G., K.P. and A.P.; Data curation, K.P. and P.S.; Formal analysis, A.P. and P.S.; Funding acquisition, J.G., K.P., A.P. and P.S.; Investigation, P.S.; Methodology, J.G. and A.P.; Project administration, A.P.; Resources, K.P. and P.S.; Software, K.P.; Supervision, J.G.; Validation, K.P., A.P. and P.S.; Writing—original draft, A.P. All authors have read and agreed to the published version of the manuscript.

Funding: The project "Automated Analysis and Processing of Legal Texts for their Transformation into Legal Open Data" is implemented through the Operational Program "Human Resources Development, Education and Lifelong Learning" and is co-financed by the European Union (European Social Fund) and Greek national funds.

Acknowledgments: We would like to thank Haris Papageorgiou and his research team at the Institute for Language & Speech Processing of the "ATHENA" Research Center for applying their Named Entity Recognizer on the legal corpus and extracting the desired entities.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

1. Bing, J. Celebrating Gnaeus Flavius and Open Access to Law. *J. Open Access Law* **2013**, *1*, 1.
2. Peruginelli, G. Law Belongs to the People: Access to Law and Justice. *Leg. Inf. Manag.* **2016**, *16*, 107–110. [CrossRef]
3. Greenleaf, G.; Mowbray, A.; Chung, P. The Meaning of "Free Access to Legal Information": A Twenty Year Evolution. *J. Open Access Law* **2013**, *1*. [CrossRef]
4. Agnoloni, T.; Sagri, M.T.; Tiscornia, D. Opening Public Data: A Path towards Innovative Legal Services. 2011. Available online: www.hklii.hk/conference/paper/2D2.pdf (accessed on 21 December 2019).
5. Wass, C. openlaws.eu–Building Your Personal Legal Network. *J. Open Access Law* **2017**, *5*, 1.
6. Custers, B. Methods of data research for law. In *Research Handbook in Data Science and Law*; Edward Elgar Publishing: Cheltenham, UK, 2018.
7. Casanovas, P.; Palmirani, M.; Peroni, S.; van Engers, T.; Vitali, F. Semantic Web for the Legal Domain: The next step. *Semantic Web* **2016**, *7*, 213–227. [CrossRef]
8. Janssen, M.; Charalabidis, Y.; Zuiderwijk, A. Benefits, Adoption Barriers and Myths of Open Data and Open Government. *Inf. Syst. Manag.* **2012**, *29*, 258–268. [CrossRef]
9. Janssen, M.; Matheus, R.; Longo, J.; Weerakkody, V. Transparency-by-design as a foundation for open government. *Transform. Gov. People Process Policy* **2017**, *11*, 2–8. [CrossRef]
10. Zuiderwijk, A.; Janssen, M.; Choenni, S.; Meijer, R.; Alibaks, R.S. Socio-technical Impediments of Open Data. *Electron. J. E-Gov.* **2012**, *10*, 156–172.
11. Palmirani, M.; Vitali, F. Legislative drafting systems. In *Usability in Government Systems*; Elsevier: Amsterdam, The Netherlands, 2012; pp. 133–151.
12. Sannier, N.; Adedjouma, M.; Sabetzadeh, M.; Briand, L.; Dann, J.; Hisette, M.; Thill, P. Legal Markup Generation in the Large: An Experience Report. In Proceedings of the 2017 IEEE 25th International Requirements Engineering Conference (RE), Lisbon, Portugal, 4–8 September 2017; pp. 302–311.
13. Dragoni, M.; Villata, S.; Rizzi, W.; Governatori, G. Combining NLP Approaches for Rule Extraction from Legal Documents. In Proceedings of the 1st Workshop on MIning and REasoning with Legal texts (MIREL 2016), Sophia Antipolis, France, 16 December 2016.
14. Gibbons, J.P. *Language and the Law*; Routledge: Abingdon, UK, 2014; ISBN 978-1-315-84432-9.
15. Nazarenko, A.; Wyner, A. Legal NLP Introduction. *TAL* **2017**, *58*, 7–19.
16. Boella, G.; Di Caro, L.; Graziadei, M.; Cupi, L.; Salaroglio, C.E.; Humphreys, L.; Konstantinov, H.; Marko, K.; Robaldo, L.; Ruffini, C.; et al. Linking Legal Open Data: Breaking the Accessibility and Language Barrier in European Legislation and Case Law. In Proceedings of the 15th International Conference on Artificial Intelligence and Law, San Diego, CA, USA, 8–12 June 2015; ACM: New York, NY, USA, 2015; pp. 171–175.

17. Virkar, S.; Udokwu, C.; Novak, A.-S.; Tsekeridou, S. Facilitating Public Access to Legal Information. In Proceedings of the 2nd International Data Science Conference, iDSC2019, Puch/Salzburg, Austria, 22–24 May 2019; pp. 77–82.
18. Cifuentes-Silva, F.; Labra Gayo, J.E. Legislative Document Content Extraction Based on Semantic Web Technologies. In Proceedings of the Semantic Web, ESWC 2019, Portorož, Slovenia, 2–6 June 2019; pp. 558–573.
19. Chalkidis, I.; Nikolaou, C.; Soursos, P.; Koubarakis, M. Modeling and Querying Greek Legislation Using Semantic Web Technologies. In Proceedings of the Semantic Web, ESWC 2017, Portorož, Slovenia, 28 May–1 June 2017; pp. 591–606.
20. Koniaris, M.; Papastefanatos, G.; Anagnostopoulos, I. Solon: A Holistic Approach for Modelling, Managing and Mining Legal Sources. *Algorithms* **2018**, *11*, 196. [CrossRef]
21. Charalabidis, Y.; Zuiderwijk, A.; Alexopoulos, C.; Janssen, M.; Lampoltshammer, T.; Ferro, E. The Multiple Life Cycles of Open Data Creation and Use. In *The World of Open Data: Concepts, Methods, Tools and Experiences*; Charalabidis, Y., Zuiderwijk, A., Alexopoulos, C., Janssen, M., Lampoltshammer, T., Ferro, E., Eds.; Springer: Cham, Switzerland, 2018; pp. 11–31.
22. Zuiderwijk, A.; Janssen, M.; Davis, C. Innovation with open data: Essential elements of open data ecosystems. *Inf. Polity* **2014**, *19*, 17–33. [CrossRef]
23. Charalabidis, Y.; Alexopoulos, C.; Loukis, E. A taxonomy of open government data research areas and topics. *J. Organ. Comput. Electron. Commer.* **2016**, *26*, 41–63. [CrossRef]
24. Lnenicka, M.; Komarkova, J. Big and open linked data analytics ecosystem: Theoretical background and essential elements. *Gov. Inf. Q.* **2019**, *36*, 129–144. [CrossRef]
25. Garofalakis, J.; Plessas, K.; Plessas, A.; Spiliopoulou, P. A Project for the Transformation of Greek Legal Documents into Legal Open Data. In Proceedings of the 22nd Pan-Hellenic Conference on Informatics, Athens, Greece, 29 November–1 December 2018; pp. 144–149.
26. Francesconi, E. A Review of Systems and Projects: Management of Legislative Resources. In *Legislative XML for the Semantic Web: Principles, Models, Standards for Document Management*; Sartor, G., Palmirani, M., Francesconi, E., Biasiotti, M.A., Eds.; Springer: Dordrecht, The Netherlands, 2011; pp. 173–188.
27. Van Opijnen, M. The EU Council Conclusions on the Online Publication of Court Decisions. In *Knowledge of the Law in the Big Data Age*; Frontiers in Artificial Intelligence and Applications; IOS Press: Amsterdam, The Netherlands, 2019; pp. 81–90.
28. Pełech-Pilichowski, T.; Cyrul, W.; Potiopa, P. On Problems of Automatic Legal Texts Processing and Information Acquiring from Normative Acts. In *Advances in Business ICT*; Mach-Król, M., Pełech-Pilichowski, T., Eds.; Springer: Cham, Switzerland, 2014; pp. 53–67.
29. Palmirani, M.; Vitali, F. Akoma-Ntoso for Legal Documents. In *Legislative XML for the Semantic Web: Principles, Models, Standards for Document Management*; Sartor, G., Palmirani, M., Francesconi, E., Biasiotti, M.A., Eds.; Springer: Dordrecht, The Netherlands, 2011; pp. 75–100.
30. Boer, A.; van Engers, T. A MetaLex and Metadata Primer: Concepts, Use, and Implementation. In *Legislative XML for the Semantic Web: Principles, Models, Standards for Document Management*; Sartor, G., Palmirani, M., Francesconi, E., Biasiotti, M.A., Eds.; Springer: Dordrecht, the Netherlands, 2011; pp. 131–149.
31. Biasiotti, M.; Francesconi, E.; Palmirani, M.; Sartor, G.; Vitali, F. *Legal Informatics and Management of Legislative Documents*; Global Centre for ICT in Parliament Working Paper No. 2; IPU: Geneva, Switzerland, 2008.
32. Tillett, B. What is FRBR? A conceptual model for the bibliographic universe. *Aust. Libr. J.* **2005**, *54*, 24–30. [CrossRef]
33. Zuiderwijk, A.; Jeffery, K.; Janssen, M. The Potential of Metadata for Linked Open Data and its Value for Users and Publishers. *JeDEM* **2012**, *4*, 222–244. [CrossRef]
34. Barabucci, G.; Cervone, L.; Palmirani, M.; Peroni, S.; Vitali, F. Multi-layer Markup and Ontological Structures in Akoma Ntoso. In Proceedings of the International Workshop on AI Approaches to the Complexity of Legal Systems. Complex Systems, the Semantic Web, Ontologies, Argumentation, and Dialogue, Beijing, China, 19 September 2009; pp. 133–149.
35. Van Opijnen, M. European Case Law Identifier: Indispensable Asset for Legal Information Retrieval. In Proceedings of the Workshop: From Information to Knowledge—Online Access to Legal Information, Florence, Italy, 6 May 2011; pp. 91–103.

36. Sandoval, A.M. Text Analytics: the convergence of Big Data and Artificial Intelligence. *Int. J. Interact. Multimed. Artif. Intell.* **2016**, *3*, 57–64.
37. Venturi, G. Legal Language and Legal Knowledge Management Applications. In *Semantic Processing of Legal Texts: Where the Language of Law Meets the Law of Language*; Francesconi, E., Montemagni, S., Peters, W., Tiscornia, D., Eds.; Springer: Berlin/Heidelberg, Germany, 2010; pp. 3–26.
38. Strembeck, M.; Zdun, U. An approach for the systematic development of domain-specific languages. *Softw. Pract. Exp.* **2009**, *39*, 1253–1292. [CrossRef]
39. Parr, T.; Harwell, S.; Fisher, K. Adaptive LL(*) Parsing: The Power of Dynamic Analysis. In Proceedings of the 2014 ACM International Conference on Object Oriented Programming Systems Languages & Applications, Portland, OR, USA, 20–24 October 2014; pp. 579–598.
40. Bacci, L.; Agnoloni, T.; Marchetti, C.; Battistoni, R. Improving Public Access to Legislation through Legal Citations Detection: The Linkoln Project at the Italian Senate. In Proceedings of the Law via the Internet 2018, Florence, Italy, 11–12 October 2018; pp. 149–158.
41. Agnoloni, T.; Venturi, G. Semantic Processing of Legal Texts. In *Handbook of Communication in the Legal Sphere*; De Gruyter Mouton: Berlin, Germany/Boston, MA, USA, 2018.
42. Sannier, N.; Adedjouma, M.; Sabetzadeh, M.; Briand, L. An automated framework for detection and resolution of cross references in legal texts. *Requir. Eng.* **2017**, *22*, 215–237. [CrossRef]
43. Stenetorp, P.; Pyysalo, S.; Topić, G.; Ohta, T.; Ananiadou, S.; Tsujii, J. BRAT: A Web-based Tool for NLP-assisted Text Annotation. In Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics, Avignon, France, 23–27 April 2012; pp. 102–107.
44. De Maat, E.; Winkels, R.; van Engers, T. Automated Detection of Reference Structures in Law. In Proceedings of the 19th Annual Conference on Legal Knowledge and Information Systems: JURIX 2006, Paris, France, 7–9 December 2006; pp. 41–50.
45. Giouli, V.; Konstandinidis, A.; Desypri, E.; Papageorgiou, H. Multi-domain Multi-lingual Named Entity Recognition: Revisiting & Grounding the resources issue. In Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06), Genoa, Italy, 24–26 May 2006.
46. Tange, O. GNU Parallel 2018. Available online: https://zenodo.org/record/1146014#.Xf2abI8RVPY (accessed on 21 December 2019).
47. Charalabidis, Y.; Zuiderwijk, A.; Alexopoulos, C.; Janssen, M.; Lampoltshammer, T.; Ferro, E. Open Data Value and Business Models. In *The World of Open Data: Concepts, Methods, Tools and Experiences*; Charalabidis, Y., Zuiderwijk, A., Alexopoulos, C., Janssen, M., Lampoltshammer, T., Ferro, E., Eds.; Springer: Cham, Switzerland, 2018; pp. 115–136.
48. *The Open Data Barometer*, 4th ed.; World Wide Web Foundation: Washington, DC, USA, 2017.
49. Marsden, C. Twenty Years of the Public Internet: Assessing Online Developments in Good Law and Better Regulation. In Proceedings of the Law via the Internet 2013, Jersey (Channel Islands), UK, 26–27 September 2013.
50. Van Loenen, B.; Vancauwenberghe, G.; Crompvoets, J.; Dalla Corte, L. Open Data Exposed. In *Open Data Exposed*; van Loenen, B., Vancauwenberghe, G., Crompvoets, J., Eds.; T.M.C. Asser Press: The Hague, The Netherlands, 2018; pp. 1–10.
51. Opijnen, M. The European Legal Semantic Web: Completed Building Blocks and Future Work. In Proceedings of the European Legal Access Conference, Paris, France, 21–23 November 2012.
52. Chiticariu, L.; Li, Y.; Reiss, F.R. Rule-Based Information Extraction is Dead! Long Live Rule-Based Information Extraction Systems! In Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, Seattle, WA, USA, 18–21 October 2013; pp. 827–832.
53. Munné, R. Big Data in the Public Sector. In *New Horizons for a Data-Driven Economy: A Roadmap for Usage and Exploitation of Big Data in Europe*; Cavanillas, J.M., Curry, E., Wahlster, W., Eds.; Springer: Cham, Switzerland, 2016; pp. 195–208.
54. Tiscornia, D.; Fernández-Barrera, M. Knowing the Law as a Prerequisite to Participative eGovernment: The Role of Semantic Technologies. In *Empowering Open and Collaborative Governance: Technologies and Methods for Online Citizen Engagement in Public Policy Making*; Charalabidis, Y., Koussouris, S., Eds.; Springer: Berlin/Heidelberg, Germany, 2012; pp. 119–138.
55. Van Opijnen, M.; Peruginelli, G.; Kefali, E.; Palmirani, M. Online Publication of Court Decisions in Europe. *Leg. Inf. Manag.* **2017**, *17*, 136–145. [CrossRef]

56. Angelidis, I.; Chalkidis, I.; Koubarakis, M. Named Entity Recognition, Linking and Generation for Greek Legislation. In Proceedings of the Thirty-first Annual Conference on Legal Knowledge and Information Systems, JURIX 2018, Groningen, The Netherlands, 12–14 December 2018; pp. 1–10.
57. Garofalakis, J.; Plessas, K.; Plessas, A.; Spiliopoulou, P. Modelling Legal Documents for Their Exploitation as Open Data. In Proceedings of the 22nd International Conference on Business Information Systems, Seville, Spain, 26–28 June 2019; pp. 30–44.

© 2019 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).