A Semantic Web Platform for Legal Knowledge in Cloud

Claudia Gheorghiu, Andrei Panu and Lenuta Alboaei
Faculty of Computer Science
Alexandru Ioan Cuza University
Iasi, Romania
Email: {claudia.gheorghiu, andrei.panu, adria}@info.uaic.ro

Abstract—The paper presents roGSEM (Romanian Governmental Semantic Portal), a platform able to provide access to legal open governmental data: normative documents published by the Romanian state. The proposed solution is based on Semantic Web technologies and uses the information according to the Linked Data initiative. Also, roGSEM semantic data repository can be used by independent plug-ins, encouraging developers to create and deploy innovative applications and mashups that give an improved access to existing legal knowledge.

I. INTRODUCTION

Web 2.0 era promote the idea of open data, information, access or knowledge and sharing it among individuals in order to create active communities.

An important part of data nowadays is gathered and stored by governmental institutions worldwide. Some of it contains sensible information that concerns the state well-being and security, but the vast majority targets information of interest for the common citizen such as taxes, educational statistics or legal data. Therefore, this data can become open data, shared with all citizens, other institutions or commercial companies.

By adopting the open government initiative [1], countries around the world have thought of ways to publish and merge existing governmental data in order to provide data exchange capabilities among institutions but also for citizens. A special attention was given to legal information, and authorities published numerous datasets of digital legal content. This led to having the right of viewing legislation online being accepted in most countries these days. As the Web expands, so does the amount of available unstructured legal information and documents made available by governments, commercial organisations or legal portals. Information overload, together with demands for qualitative legal information knowledge search, discovery, and management have created the need to find and implement newer approaches for content management tools.

In this paper we present roGSEM, a Semantic Web platform that provides access to heterogeneous legal documents published by the Romanian state. The platform is able to support creating and deployment of mashups that are able to process and analyze semantic legal data.

According to the open government data initiative, roGSEM targets both end users and developers. For the end-users, roGSEM offers an interactive tool for browsing and filtering public governmental data, in order to visualize, analyze and enhance the results. For developers, the platform offers a simple integration mechanism of new mashups that use and process existing public sector information. The platform is addressed to any type of users, not necessarily ones that have a previous technical training in using Semantic Web technologies. Also, this tool is not restricted to a specific data source and is able to integrate different types of legal documents. Our platform allows registering of specific independent plugins, automatically incorporating their functionality into the application. It encourages developers to participate in building innovative applications, visualizations and mashups that offer citizens improved access to governmental data. The processed information is later presented in a coherent and user-friendly manner to the end user.

In previous work, we proposed SIGMA [2], a semantic platform that allowed users to browse through datasets published on the UK and US open governmental websites. The main focus was merging Linked Open Data from multiple sources, and not particularly to provide a Romanian equivalent for the US or UK portals. This approach is the main direction in our new project, roGSEM, which aims to extend and improve our previous platform.

The main purpose of roGSEM is to offer users a legal knowledge management tool and allow common users without any previous legal training to browse and find connections between Romanian normative documents.

The paper is structured as follows. In section 2, we outline the background of legal management knowledge and previous related work. Section 3 presents current solution architecture details. In section 3.1 we briefly describe technologies involved in the development of roGSEM while in 3.2 the emphasis is presenting how knowledge management is performed in our approach. Details about the user interface of our platform are found in section 3.3, followed by security and performance information. Section 4 illustrates a possible use case scenario and how a user could benefit from the advantages of such a platform. We conclude the results of our work and set future directions of development and further enhancements in the last part of this paper, section 5.

II. LEGAL KNOWLEDGE MANAGEMENT

In the context of international relations between countries and new forms of organization (such as the European Commission) more and more information is available and necessary if one intends to have a complete legal context about a specific
issue. Therefore it becomes increasingly difficult to reach the correct piece of information on time. Moreover, legal content management tools must take into consideration multilingual searches and retrieval of data, but also the difficult task of finding legal concepts equivalents among different countries.

To meet these problems, specific applications and frameworks are built [3]. Newer research in Legal Informatics [4] is represented by semantic analysis of content and reuse of existing knowledge by addressing the resource consumption bottleneck problem. Systems targeting automated summarization of legal texts, legal entity recognition tools or ontology-based law discovery are some of the most recent applications being developed.

Regarding the incorporation of legal knowledge into the Web or into IT applications, or the more complex realization of the Legal Semantic Web, several directions have been taken, such as the development of XML standards for legal documentation and drafting (e.g. CEN MetaLex [5]), and the construction of legal ontologies. CEN MetaLex proposes a standard XML format for representing sources of law and references to sources of law from various levels of authority and different countries and languages. The service also allows users to query the data repository (almost all Dutch regulations in CEN MetaLex XML and as RDF Linked Data) by typing (part of) a title, and a date before which the document was published. By representing only sources and references, (without the actual content of legal documents) the system is restricted to being a legal reference tool, and not a legal management tool.

During the last decade, research on the use of legal ontologies as a technique to represent legal knowledge has increased. More recent contributions at that level include the Ontology of Fundamental Legal Concepts (one of the sources for the LKIF Core Ontology [6]). The Legal Knowledge Interchange Format (LKIF) [7] developed in the ESTRELLA project [8], is a Semantic Web based language for representing legal knowledge. The system is based on the LKIF Core Ontology, a standard vocabulary of basic legal terms and defines a set of approximately 200 concepts. The system is not focused on transforming existing unstructred data into the knowledge representation of the proposed ontological model, but to offer a groundwork for a legal knowledge based system. Therefore it doesn’t provide a semantic legal document repository or search features for public use.

Another proposal is the Norme in Rete (NIR) [9] portal, which enables access to legislation by providing a unique point of access to Italian and EU legal documents published on different web sites. The application downloads laws from institutional legal portals via dedicated spiders. The content is then parsed using the URN format of NIR (standard XML representation for Italian laws). The XML schemas used in the parsing process contained Italian legislative structural information as well as administrative and semantic metadata. The portal does not offer an API or the possibility for third party applications to reuse the stored legal content. It is intended for final users and not necessarily for developers.

III. USED TECHNOLOGIES

Nowadays, the World Wide Web is mainly composed of pages (digital documents containing markups) with information in the form of natural language text and multimedia created for human to read and understand. Computers are therefore generally used to render this information and not to reason about it. Information can no longer be intended for human readers only, but also to be processable by machines [10], [11]. This can be possible if using the advantages of the Semantic Web.

One of the central concepts of our project is Uniform Resource Identifiers (URIs). Each resource available on Web is uniquely addressed by using the Uniform Resource Identifiers. One of the most representative URI subset is the one regarding localization, Uniforms Resource Locator (URL) [12].

To express and process meta-data (data about data), the Resource Description Framework (RDF) model is used. RDF is a current standard of the World Wide Web Consortium [13] and an important brick of the Semantic Web, in order to represent information about Web resources. The RDF format is intended to capture and state the conceptual structure of existing information. The core concept is that of making statements about (Web) resources - the so-called RDF triples - of the form: “Subject has a property whose value is an object”. This way, RDF assertions - URI triples having the <subject, property, predicate> structure - can be interpreted as a semantic data representation model for describing semantic information that can be processed by computers.

roGSEM benefits from the advantages of RDF by using all of its processed information in RDF format. Furthermore, roGSEM components use SPARQL, an RDF query language [14]. This language allows accessing data stored as triples, and defining ways to expose information using web services - so called SPARQL endpoints.

The RDF files contain information structured according to concepts illustrated by several ontologies. The ontologies are created using Web Ontology Language (OWL) [15] that structures and characterizes resources and/or relations between them. By using OWL, the knowledge about the resources can be shared within a given community of practice.

The roGSEM platform is installed on the Microsoft cloud platform, Windows Azure, but it is not limited to using this cloud environment. It can be installed on any platform that provides .NET framework compatibility and Windows or Linux virtual machines.

The data stored in this environment are provided to the interrogation module through an SPARQL endpoint made available as a public web service. The interrogation module of roGSEM gathers its data by issuing HTTP requests to the Virtuoso [16] web service. The endpoint is able to provide data in different formats, such as XML, JSON, HTML, NTriples, RDF/XML, CSV.

Due to a SOA like architecture, the web application can run in an cloud environment different from the one hosting the Virtuoso data server, components’ loose coupling being ensured also at the execution level.
IV. ARCHITECTURE

Our proposal, roGSEM, is a scalable platform, build using the Model-View-Controller architectural pattern. It enables integration of distributed or local functional components that process semantic data and provide the advantages of the Semantic Web. This section presents the main concepts and technologies that were used in the implementation of the roGSEM platform.

A. Knowledge Modeling in roGSEM

The goal of our platform is to provide access to Romanian governmental data, in particular to legislative documents. The current implementation of the platform has two main modules: the Semantic Annotator (called SA) and the Semantic Interrogation Tool (called SIT). The first module is responsible with creating the data repository needed by the platform and the later one with offering a user interface for browsing legal documents.

At the moment, the Romanian government does not publish data in semantic format, but does publish information in various formats, using institutions’ official websites. To gather and store available open governmental information we build the Semantic Annotator. One of the annotators’ components, the RoGov Crawler, parses the data from a raw format into a format suitable for being processed by the second component: the Semantic Generator.

The RoGov Crawler has a modularized architecture, scalable and flexible enough to be used on any type of website, having a template based approach. The crawler allows registering new specialized modules, to extract heterogeneous data from official public data sources.

The existing modules process information such as the Romanian Constitution Law and Romanian Official Gazette’s publications. The obtained information becomes the input for the Semantic Generator, a component specialized in creating RDF files by annotating the plain information with semantic metadata.

The first RoGov Crawler sub-module is designed to extract data from the Romanian Official Gazette found on the Romanian Chamber of Deputies’ official website. All information gathered is then translated into English in order to ensure a minimal internationalization and allow foreign citizens to analyse Romanian laws.

Because the vast majority of the articles published in the Official Gazette refer to the Romanian Constitution, we implemented a specific component to parse and transform the unstructured data into RDF files according to the Open Linked Principles [17]. This way, users could navigate through topics of interest in regulations published in the Official Gazette and also analyse the original text that the legislative item is referring to. Information found in the Romanian Constitution is available in three languages: Romanian, English and French.

For organizing the RDF files, several ontologies were also developed depending on the conceptual model needed. For the legal articles of interest, the ontology models concepts needed to illustrate a normative act. Normative documents are elaborated according to their hierarchy, their category and the competent public authority to adopt them. The main legislative techniques are found in Law no. 24/2000 updated on legislative drafting rules for drafting normative acts published in the Official Gazette no. 260 of 21.04.2010. Official Gazette is the official publication of the Romanian state, which contains specific regulations types. This is covered by the Constitution and refers to laws, ordinances, decrees and also other acts covered by the law of the organization and functioning of the Official Gazette.

In this approach, we model the most important documents published in the Official Gazette such as: Decisions, Ordinances, Emergency Ordinances, Presidential Decrees, Orders and Laws. Every legal act has a publisher, and depending on the publisher type (institution or state representative), the document is categorized according to the Romanian State legal system. The President issues decrees, the Romanian Parliament publishes laws and ordinances or emergency ordinances are only issued by the Government. Specific classes of Governmental agencies along with the Prime Minister can release decisions. Almost any official authority can issue orders but only orders with a general character are included in the official publication. Under the general provisions of Law no 24/2000, in the Official Gazette are not published classified laws and normative acts of individual scope issued by autonomous administrative authorities and central public administration bodies.

Another central topic of our normative act ontology is related to a document’s structure. The enacting is the actual content of the legal document, consisting of Titles, Sections, Chapters, Articles and Paragraphs. We organized legal document’s enacting structure by defining the contains object property with the corresponding restrictions, found in Law no. 24/2000. Each piece of legislation, document, article and paragraph has an identifier assigned with a Unique Reference Number, which makes it easy to link to the source at any level of granularity.

In order to fully represent a normative act depending on its publisher, a new ontology was created. This ontology conceptualizes basic administration information for the Romanian State, such as Organisation and Politician. The first class contains public institutions that can be further categorized according to their functional characteristics in executive or legislative organisations. Executive authorities include central agencies involved in the Romanian state administration. Politicians include entities like President, Prime Minister or Minister.

The relation between a document and its issuer is represented using the hasPublisher object property. In order to map every document class and its publisher, ontology’s classes define restrictions over the hasPublisher property using OWL axioms. Each restriction defines the specificity of the property domain through owl:allValuesFrom, owl:someValuesFrom or owl:oneOf axioms according to the drafting specifications mentioned above.

Both ontologies were linked to DBpedia classes using the equivalentClass OWL axiom in order to ensure that the represented concepts are included in the Linked Data Cloud.

The Semantic Generator uses the previously described ontologies to annotate the files extracted by the crawler with
the corresponding semantic concepts. The resulted files are further analyzed using OpenCalais [18] service to automatically generate semantic meta-data. The OpenCalais service uses natural language processing algorithms, machine learning and other text parsing techniques, analyzing documents and identifying embedded entities: persons, companies, organizations, geographical data, books, albums, authors, etc. It can also identify facts (statuses, alliances, educational studies, political aspirations, etc) or events (sport events, managerial changes, etc). This way the RDF data is enhanced with the result of legal entity recognition techniques provided by OpenCalais. For example, if processing the following excerpt from Government Emergency Ordinance no. 75/2005: “The policies promoted by the Ministry of Education and their implementing specific strategies contribute to the assessment, assurance, control and continuous improvement of quality of education. Public finance of education is, first of all, depending on quality. [...] (4) Policies to ensure the quality of education in Romania are permanently linked to the actions promoted at European and world level.”, we identify as organization the Ministry of Education, public finance as an industry term, the Romanian country, but also a generic relation between the Ministry of Education and the policies promoted by this authority.

To allow a better filtering and an overview of document’s scope, we annotated every legal article with topics and tags provided by the natural language processor. SocialTags is a feature of OpenCalais that attempts to emulate how a person would tag a specific piece of content, feature that can be used for organizing and navigating. Another available functionality is document categorization of the submitted content input - the topic or topics that the content discusses according to Calais Categorization taxonomy. By using these features the platform adds more metadata and provides the possibility to perform advanced inferences. This way, the previous example is further enhanced with education, accreditation and quality assurance tags while the main topic of the fragment is identified as education.

B. roGSEM Modules

roGSEM user interface functionality is based on independent modules and plugins. The plugins offer users functional tools to manipulate semantic data in a user-friendly manner.

By having a SOA architecture the platform can be easily extended. At any time, a new component can be added to the list of existing ones, already registered in the application. The process of adding a new component does not imply previous configuration or redeployment of the platform. Any developer can implement new functionalities, register them in the platform, and they will be automatically loaded into roGSEM.

Fig. 1 illustrates the general architecture of the roGSEM platform, with all its components: the Semantic Annotator, the data repository module and the user interface tool.

The Semantic Interrogation Tool (called SIT) is responsible for communication with the end-user. It allows users to dynamically query legal governmental data stored in the repository. All user input is translated into SPARQL interrogations over the RDF files, de facto format for semantic web application data sources. The SIT module then renders the results in a coherent manner, and, depending on the nature of data, further performs specific analysis. Data can be merged with other data and become a new data source for other functional components.

The process of registering a new plugin requires the developer to specify the component’s URL (hosting address), a name and information necessary to associate the mashup with existing entities in the endpoint, which represent the processed information.
Registered plugins communicate with the platform and also with other active plugins using the roGSEM API. Using elements URIs, components can further load and perform complex inferences in order to display better and complete results.

At every moment, a user can browse through all governmental records found in the platform’s endpoint. The user can select several subjects and specify the property that he intends to have analyzed. The corresponding plugin list will be instantiated and every plugin will be displayed allowing users to interact with the selected information.

Starting from the platform’s data sources, we have implemented a functional component and included it into roGSEM. The plugin analyses connections between documents based on the normative act ontology. It uses item’s structural hierarchy described by the ontological model and allows users to browse through Constitution Law contents. Semantic entities are highlighted by the view and details are available by clicking on every entity’s link. When users pass their mouse over terms held in the ontology, popup previews appear with ontological denitions and hyperlinks to further information. Chapters or articles are connected based on elements URI’s created by the Semantic Generator. The plugin generates the semantic network represented by the RDF files and offers users a tool to easily navigate through complex legal items, by pointing to other possibly connected documents.

C. Performance and Security Considerations

roGSEM is built as a platform. Each time the information needs to be displayed, it searches for available plug-ins that are designed to load the specified type of data. If one (or more) plug-ins are available, it will be loaded and the needed data will be passed to the plug-in (the URI of the predicate and the URI of the resource). Once the plug-in receives this information, it can start loading the records and display them in a custom manner. Beside the RDF data, a plug-in can also use other external data-sources or APIs, independent from the main application. Using this approach, the platform can be extended by adding new plug-ins created by third party developers.

Because roGSEM relies completely on a remote endpoint for retrieving data, the performance of each operation depends on the quality of the services offered by the data provider.

From the security point of view, the implementation uses the primary security model of .NET platform. Having the platform in a cloud environment, the security is enhanced by the cloud services provider.

V. USE CASE

We consider the situation when a user wants to know which are the steps needed to be followed when setting up a kindergarten in Romania. Multiple normative acts published by Romanian authorities describe different details regarding pre-school education institutions. Therefore the user needs to have access to various legal documents in order to have a complete understanding of the processes involved in creating and managing such an organization.

He would like to see which is the legal framework of private education in Romania. For this, he would also want to analyse the fundamental law of the state, the Constitution, but also any regulations on this topic.

In order to see available legal items, the user must select the normative act entity class displayed by roGSEM. The corresponding plugin will be instantiated, displaying an input search for browsing regulations. When searching for “education”, the user finds as a result the National Educational Law. The application displays the entire content of the legislative document, but also the possibility for the user to further search for elements of interest within the selected document. Relevant articles of the law are displayed when searching for “private education”. The user can analyse details about preschoolers’ rights, how salaries are determined or mentions about the management structure of the institution. Internal references to articles or paragraphs are automatically determined by the application and allow users to easily preview the content of the reference, without actually leaving the current paragraph. Some results refer to the Constitution, a concept highlighted by the plugin as an external reference. By clicking on the hyperlink, the user navigates to the document in question and here he can further refine his search scope. The user can always return to any step of his current session and continue or modify search criteria as needed.

Information regarding accreditation of a preschool institution is available in several documents having legal dispositions on accreditation, namely the Government Emergency Ordinance no. 75/2005 on quality assurance in education, approved with amendments by Law no. 87/2006. Again, the platform automatically identifies these external references when the user browses through the Education Law and allows direct navigation.

When opening Law no. 87/2006 the user learns more details about the institutions responsible for ensuring quality in education, such as their internal departments, their responsibilities and also the steps needed to obtain the accreditation. The law explains that the first step must be obtaining an authorization for functioning. The second step is the accreditation to release official education documents. The procedure for every stage is described in an exact and brief manner.

VI. CONCLUSIONS AND FUTURE WORK

We present in this paper a platform, roGSEM, able to create mash-ups providing access to open legislative data. Our approach is based on the existing semantic Web technologies and uses public distributed portals conforming to the linked data initiative [17].

roGSEM performs queries by using a SPARQL endpoint, retrieving the available data and displaying it in a human readable and user friendly way. Functional plug-ins can perform complex data analysis and correlations. roGSEM platform is addressed to any kind of users that want to use the governmental information published on administration sites [19]. The platform does not require previous knowledge about creating SPARQL queries nor about the ontologies used to model and organize these kinds of data. All the transformation and filtering is automatically performed by the platform, so that users can visualize the output of their request in a comprehensive way. Also, the input data-sources can be easily modified, roGSEM not being constrained to a specific data type. From
the architecture point of view, the application is extensible via independent plug-ins.

We shall consider several future developments - e.g., performing inferences and aggregations in order to obtain a relevant result regarding specific topics of interest and integrating new functional plug-ins.

ACKNOWLEDGMENT

The authors would like to thank Mr. Radu Nicolescu for his valuable contribution to this article. This work was supported by a grant of the Romanian National Authority for Scientific Research, CNDDI - UEFISCDI, project number PN-II-IN-CI-2012-1-0200.

REFERENCES


