Instituto Superior de Engenharia do Porto

Bridging the gap between closed and open data

System proposal for the Portuguese Legislation

Nuno Miguel Pereira Moniz

Dissertation to obtain Master degree in Computer Engineering, specialization in Networks, Architectures and Systems

Supervisor: Maria de Fátima Coutinho Rodrigues

Júri:

Presidente:

Doutor José António Reis Tavares, Instituto Superior de Engenharia do Porto

Vogais:

Doutor Luís Fernando Rainho Alves Torgo, Faculdade de Ciências da Universidade do Porto

Doutora Maria de Fátima Coutinho Rodrigues, Instituto Superior de Engenharia do Porto

Porto, October 2012

Resumo

Esta dissertação apresenta uma proposta de sistema capaz de preencher a lacuna entre documentos legislativos em formato PDF e documentos legislativos em formato aberto. O objetivo principal é mapear o conhecimento presente nesses documentos de maneira a representar essa coleção como informação interligada.

O sistema é composto por vários componentes responsáveis pela execução de três fases propostas: extração de dados, organização de conhecimento, acesso à informação.

A primeira fase propõe uma abordagem à extração de estrutura, texto e entidades de documentos PDF de maneira a obter a informação desejada, de acordo com a parametrização do utilizador. Esta abordagem usa dois métodos de extração diferentes, de acordo com as duas fases de processamento de documentos – análise de documento e compreensão de documento. O critério utilizado para agrupar objetos de texto é a fonte usada nos objetos de texto de acordo com a sua definição no código de fonte (Content Stream) do PDF. A abordagem está dividida em três partes: análise de documento, compreensão de documento e conjunção. A primeira parte da abordagem trata da extração de segmentos de texto, adotando uma abordagem geométrica. O resultado é uma lista de linhas do texto do documento; a segunda parte trata de agrupar os objetos de texto de acordo com o critério estipulado, produzindo um documento XML com o resultado dessa extração; a terceira e última fase junta os resultados das duas fases anteriores e aplica regras estruturais e lógicas no sentido de obter o documento XML final.

A segunda fase propõe uma ontologia no domínio legal capaz de organizar a informação extraída pelo processo de extração da primeira fase. Também é responsável pelo processo de indexação do texto dos documentos. A ontologia proposta apresenta três características: pequena, interoperável e partilhável. A primeira característica está relacionada com o facto da ontologia não estar focada na descrição pormenorizada dos conceitos presentes, propondo uma descrição mais abstrata das entidades presentes; a segunda característica é incorporada devido à necessidade de interoperabilidade com outras ontologias do domínio legal, mas também com as ontologias padrão que são utilizadas geralmente; a terceira característica é definida no sentido de permitir que o conhecimento traduzido, segundo a ontologia proposta, seja independente de vários fatores, tais como o país, a língua ou a jurisdição.

A terceira fase corresponde a uma resposta à questão do acesso e reutilização do conhecimento por utilizadores externos ao sistema através do desenvolvimento dum Web Service. Este componente permite o acesso à informação através da disponibilização de um grupo de recursos disponíveis a atores externos que desejem aceder à informação. O Web Service desenvolvido utiliza a arquitetura REST. Uma aplicação móvel Android também foi desenvolvida de maneira a providenciar visualizações dos pedidos de informação.

O resultado final é então o desenvolvimento de um sistema capaz de transformar coleções de documentos em formato PDF para coleções em formato aberto de maneira a permitir o

acesso e reutilização por outros utilizadores. Este sistema responde diretamente às questões da comunidade de dados abertos e de Governos, que possuem muitas coleções deste tipo, para as quais não existe a capacidade de raciocinar sobre a informação contida, e transformá-la em dados que os cidadãos e os profissionais possam visualizar e utilizar.

Palavras-chave: Extração de Texto, PDF, Recuperação de Informação, Ontologia, Domínio Legal, Dados Abertos

Abstract

This dissertation presents a system proposal capable of bridging the gap between legal documents in PDF format and open legislative documents. The objective is mainly to map the knowledge present in these documents in order to represent the collection as linked information.

The system contains various components responsible for the execution of three proposed phases of execution: data extraction, knowledge organization and information access.

The first phase proposes an approach to extract structure, text and entities from PDF documents in order to obtain the desired information in accordance with the user parameterization. The second phase proposes a legal domain ontology in order to organize the information extracted from the extraction process of the first phase and is also responsible for the indexing process of the legislative text of the documents. The third phase provides an answer to the access and reuse of the knowledge by third parties through the development of a Web Service. Additionally, an Android Mobile Application was developed to provide visualizations of the information requests.

The desired final outcome is thus the development of a system that transforms collections of PDF documents to open data format collections in a way that it should become accessible and reusable by third parties.

Keywords: Text Extraction, PDF, Information Retrieval, Ontology, Legal Domain, Open Data

Acknowledgments

Kind regards to Fátima Rodrigues who supervised this project in an excellent manner, willingly providing me the space to think, test and implement many ideas.

Regards to the faculty of the Computer Engineering Superior Institute (ISEP) for providing me the chance to learn and evolve both academically and personally. Also, a special mention and acknowledgment to João Coutinho.

A special thank you to João Carlos, João Mineiro, José Miranda, José Soeiro, Leonor Figueiredo, Miguel Heleno, Ricardo Lafuente, Ricardo Gomes and Ricardo Sá Ferreira. It wouldn't be possible without your friendship and support.

To Raquel.

To Bruno.

To Mário and Maria.

To Margarida, Manuel and Ana.

Index

Figures

Tables

Acronyms and Symbols

Acronyms List

CSV	Comma-Separated Values
DRE	Electronic Republic's Diary
EU	European Union
ICT	Information and Communications Technology
PDF	Portable Document Format
PSI	Public Sector Information
RDF	Resource Description Framework
SPARQL	SPARQL Protocol and RDF Query Language
UK	United Kingdom
URI	Universal Resource Identifiers
USA	United States of America
W3C	World Wide Web Consortium
XML	Extensible Markup Language

1 Introduction

1.1 Scope

The information published on the Internet has had an exponential increase for many years, with an increased acceleration in the last couple of years. Amongst the various types of information this dissertation refers to public and governmental information. The Internet has played a main role in the development of means for citizens' engagement and government transparency. Either the e-government or the open data/government initiatives have and are being developed in countries and by groups of people on an international level. But, while e-government mainly depends on the initiative of governments and its agencies in the development and maintenance of these types of initiatives, open data/government initiatives are mostly developed by citizen groups on a local, nation or international level, not disregarding the open data initiatives from governments such as United Kingdom, Brazil and United States of America.

The relationship between citizens and State institutions can be established in both directions. The measures and steps taken towards a more available and transparent public sector of information (PSI) also reflects this. Mainly, there are two types of initiatives from a State perspective: it may develop front-end systems that allows users to connect and interact with the institutions of the State (e.g. Law making processes, transparency watch, public planning and street planning, among others) or it may make available the data that the citizens can use to develop initiatives (e.g. crossing of information, better connection with a State structure, visualizations) for other citizens. In either case, the availability of the State in terms of sharing and publicly divulging that information is crucial regarding the development of open data initiatives.

Other than the question of openly divulging and sharing public information, there is still the question of how to post this information. The last few years have been rich in examples of governments sharing information in open formats, such as CSV[1], XML[2] and RDF[3], rather than the proprietary formats, such as PDF[4], Excel spread sheets and Word documents. This ensures that the access and reuse of the information will be much easier for the users, the citizens and the groups developing initiatives in this area, and moreover, it facilitates the legal issues of the use of this available information.

Therefore, there are two initial problems: publishing public information and the format in which it is published.

[1] Comma-Separated Values - http://tools.ietf.org/html/rfc4180
[2] Extensible Markup Language - http://www.w3.org/TR/2006/REC-xml11-20060816/
[3] Resource Description Framework - http://www.w3.org/RDF/
[4] Portable Document Format - http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=51502

This dissertation focuses on the legislative information present in the Portuguese Electronic Republic's Diary[5] (DRE). This site is managed by Imprensa Nacional – Casa da Moeda[6]. Its objective is to enable, through the Internet, the access to all Republic Diaries, as well as all the numbered diplomas published in the Republic Diary since the 5th of October of 1910. All the information is published in PDF format, a proprietary format.

One of the major challenges regarding Information and Communications Technology (ICT) and legislation is the question of how to engage the legislators and citizens in order to provide a more transparent process and dynamics between these two parts. Also, there is the ambition of international interconnection and in some way, interdependence. This part of the problem, addresses mainly the issue of interoperability and inter-linkage of various legislative actors in various countries in order to provide coordination of efforts.

According to the initial problem referred before, in fact the information required for this project is available, but not in an open format. Thus, the main objectives of this research are structured in the following manner: accessing the legislative information available in DRE; investigating how to translate that information from PDF format to an open format and, finally, how to make it accessible and reusable to others. But also, other objectives appear: linking of related legislative documents and entities present in the documents.

Furthermore, besides the question of accessing and understanding the documents, there is a question of organizing the information extracted from the previous described processes. This is a second area of research in this dissertation. The organization of knowledge that has a direct relation with conceptualizations more or less accepted along the years poses a question in the development of this project. By assuming as a challenge and objective to develop this project in the sphere of the Semantic Web [Berners-Lee, 2001], the use of databases based on ontologies is a starting point, as so the use of RDF language. In the following sub-section this subject will be better developed containing information regarding the different examples of similar projects.

The third area of this project is focused on the delivery of the organized information, both to access and reuse purposes. This poses a question in the area of delivery of information, of best practices and possibilities of integration and relation of information that is possible to obtain.

In accordance with this scope, the next section presents a background on the areas in which this dissertation is considered to be an integral part.

1.2 Background

During the past decade and up until now, there has been a significant increase in the importance of government transparency. Not only transparency related to public information but also related to the normal processes of a State. With the increased capabilities of the Internet and its use, and also

[5] http://www.dre.pt
[6] http://www.inmc.pt

with the increase importance that it has in the daily life of a significant part of sectors such as economy, finance, journalism and others, an increase of interest and also initiatives by groups and also by governments in these areas can be observed. To clarify the extent of this dissertation in terms of topics, it is necessary to describe their meaning in order to describe the boundaries of this document and the terms used.

The open data movement has had a considerable increase of popularity in recent years, mostly due to the birth of several projects such as data.gov[7] or data.gov.uk[8]. The open data movement, similar to other "open" movements such as open access, open source, and others, mainly advocate that data should be free to everyone to use and reuse without the risk of copyright infringements or other constraints in both referred actions. There is still no formalization of the concept "open" in open data, but the usual definition is pointed to the one described by Open Definition which states that "A piece of content or data is open if anyone is free to use, reuse, and redistribute it — subject only, at most, to the requirement to attribute and/or share-alike" [ODa, 2005].

Furthermore, the open data as a movement has its applications in various fields, such as science, education and government. The examples provided before to argument the "boom" of open data initiatives are an example of the latter that is the area in which this project is inserted: open government.

The Open Definition is also the main reference in the definition of Open Government, stating that its data and content should be "open" as defined before and should be "produced or commissioned by government or government controlled entities" [ODb, 2005]. The data produced by the Open Government initiatives, the Open Government Data, has many examples of applications, in various levels of the organization of a country or group of countries (local, national or others), in the subjects of transparency and democratic control, civil participation and governmental efficiency.

Regardless of the definitions, the use of Open Government and Open Data as classification for initiatives has fallen into some ambiguity in many cases. Yu and Robinson [Yu and Robinson, 2012] have addressed this ambiguity in "Open Government" by clarifying the roots of both Open Data and Open Government and their respective evolution as a concept. The authors also propose a stylized framework to consider both the "actual or anticipated benefits of the data disclosure" and "how is the disclosed data structured, organized and published". The former oscillates between service delivery and transparency and the latter between inert data and adaptable data. The framework is presented in Figure 1.

[7] http://data.gov
[8] http://data.gov.uk

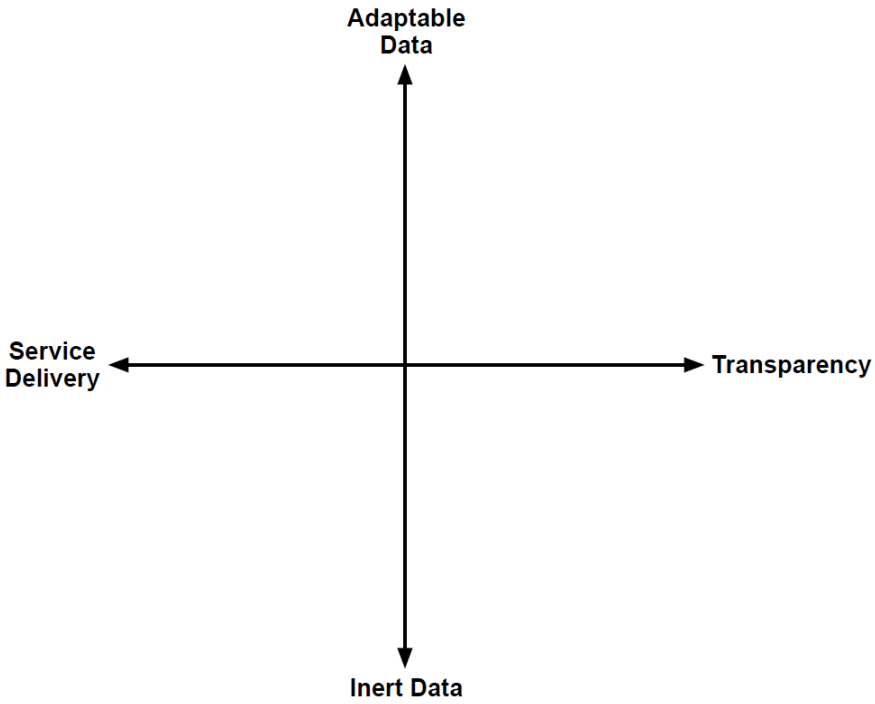


Figure 1 – Yu and Robinson stylized framework proposal

This framework is important in order to classify the current status of the available data and services provided in the legislative sector of the Portuguese State. Also, it is important to observe the difference between the existing classification and the intended new classification with the project described in this dissertation.

The Semantic Web [Berners-Lee et al., 2001] was introduced by Tim Berners-Lee and others stating that "the Semantic Web is not a separate Web but an extension of the current one, in which information is given well-defined meaning, better enabling computers and people to work in cooperation". It took the idea of interoperability to a new level, regarding the linkage of information in such a way that it would be all linked, thus providing basis to new areas of development and expertise. They forecasted that through the use of technologies such as XML, RDF and ontologies, providing the basis for inference of data "on the fly", and with everything being represent by an URI, the possibilities would be considerable in terms of evolution and technological advance, in many areas.

Five years after the public introduction of the concept Semantic Web, Shadbolt, Hall and Berners-Lee revisited [Shadbolt, N. et al, 2006] the idea and analyzed its evolution. Their conclusion was that, despite the *de facto* evolution in practical terms and in number of initiatives, the idea "remains largely unrealized". In contradiction with the original presentation which presented the realization of the Semantic Web as a somewhat straightforward path, the constraints and technological limitations played a significant role in the difficulties of the development of the Web of Data. The difficulty in the creation of standards and intelligent agents capable of translating heterogeneous data, the large-scale mediation and the closeness of data silos and the difficulties in translating and opening their data are some of the issues stated by the authors.

In the present, regarding government initiatives, the European Parliament and Council Directive on re-use of PSI [EU, 2003] had already been approved and published, and several countries developed various initiatives to implement the EU directive, such as the UK Office of Public Sector Information.

Related to one of the difficulties referred by Shadbolt et al. [Shadbolt et al., 2006], in the year 2009 the first government initiatives regarding the availability of various information in raw state were developed, with special attention to the United States of America (www.data.gov) and the United Kingdom (www.data.gov.uk) experience. In terms of national initiatives, it is estimated that by now there are twenty countries that made available a considerable amount of information in a format in compliancy with the definition of open, including Portugal[9]. Since 2009 the W3C[10] has developed guidelines on standards and best-practices on publishing Open Government Data [W3C, 2009]. It states three steps in order to accomplish the task:

1. The quickest and easiest way to make data available on the Internet is to publish the data in its raw form (e.g., an XML file of polling data from past elections). However, the data should be well-structured. Structure allows others to successfully make automated use of the data. Well-known formats or structures include XML, RDF and CSV. Formats that only allow the data to be seen, rather than extracted (for example, pictures of the data), are not useful and should be avoided.
2. Create an online catalogue of the raw data (complete with documentation) so people can discover what has been posted.
3. Make the data both human- and machine-readable

The document also states a few guidelines of no less importance regarding aspects of this process such as the identification of things, documentation, the linkage of data (Linked Data), preservation of data, exposure of interfaces, standard names and considerations on which data and format to publish and restrictions on its use.

Concerning the concept of Linked Data[11], this is defined by W3C as the interrelated datasets on the Web that make the Semantic Web a reality. It refers to the necessary explicit relationships between data, in addition to its access. In order to achieve this, the Semantic Web provides the basis and the technologies such as RDF enable these processes (on-the-fly access and conversion of existing data and databases). Related to the use of RDF, there are technologies that enable the setup of endpoints that provide access to the data, such as SPARQL [W3C, 2008]. One of the most common examples of Linked Data initiatives is DBPedia[12], which publishes Wikipedia[13] content in RDF.

Still related to the concept of Linked Data, Tim Berners-Lee [Berners-Lee, 2009] besides the considerations regarding the concept, sets out a star rating system for Linked Open Data, which is

[9] Portal de Dados Abertos de Portugal – www.dados.gov.pt
[10] World Wide Web Consortium – www.w3c.org
[11] World Wide Web Consortium (Linked Data) – http://www.w3.org/standards/semanticweb/data
[12] DBPedia - http://www.dbpedia.org/
[13] Wikipedia - http://www.wikipedia.org

the junction of Linked Data with the previously referred definition of open. The classification is as such:

★ Available on the web (whatever format) but with an open licence, to be Open Data
★★ Available as machine-readable structured data (e.g. excel instead of image scan of a table)
★★★ as two stars plus non-proprietary format (e.g. CSV instead of excel)
★★★★ All the above plus, Use open standards from W3C (RDF and SPARQL) to identify things, so that people can point at your stuff
★★★★★ All the above, plus: Link your data to other people data to provide context

The effort by governments to enter into the Internet era was mainly done by developing initiatives that provided State services or information online. These initiatives usually fall into the category of e-government initiatives. Electronic Government is the idea of providing online services and information that a Government and State provide physically.

The importance of mobile phone access has increased in direct relation with the increase in purchases of mobile phones with capacity to access the internet. In India it is foreseen that Mobile Internet will exceed PC access in that country by the end of this year [Russell, 2012]; the UK Office for National Statistics, in August of 2011 state [ONS, 2011] that there was an increase of "6 million people accessing the Internet over a mobile phone for the first time" and that "45 per cent of Internet users used a mobile phone to connect to the Internet" and in China it was estimated [CIW, 2011] in 2011 that "66 per cent of Internet users access the Internet through mobile phones".

With this increase in numbers and importance, the evolution of e-government initiatives went to the development of initiatives that take this increase in consideration, creating the concept of Mobile Government (m-government or mGovernment). According to the Mobile Government Consortium International[14] mGovernment is not just eGovernment using the new technologies and available channels. It is more an extension of eGovernment due to the growing numbers of mobile access to the Internet and the technological imposition of this fact. mGovernment mainly aims at providing services and applications that improve governments' fundamental functions [mGCI, 2011].

1.3 Approach

1.3.1 General Description

A significant part of the developments in the field of the legislative domain have been towards representing the process and relationship of legal documents with other legal documents. Our system is not designed for such a purpose. It is developed specifically for the process of extracting data in the legislative domain, from collections of PDF documents that do not enable the extraction of such knowledge. The purpose is therefore to bridge the gap between legal documents in PDF

[14] Mobile Government Consortium International - http://www.mgovernment.org

format and open legislation documents. By open legal documents we refer legislation text, its metadata, but also the entities present in that information. Our objective is, mainly, to map the knowledge present in these documents in order to represent the collection as linked information.

The desired final outcome is thus the development of a system that transforms collections of PDF documents to open data format collections in a way that it should become accessible and reusable by third parties. It addresses directly the problem for the open data community and for the Governments that mostly own these collections but don't provide the ability of reasoning the information contained in it, and transforming it into data that citizens and developers may access and reuse.

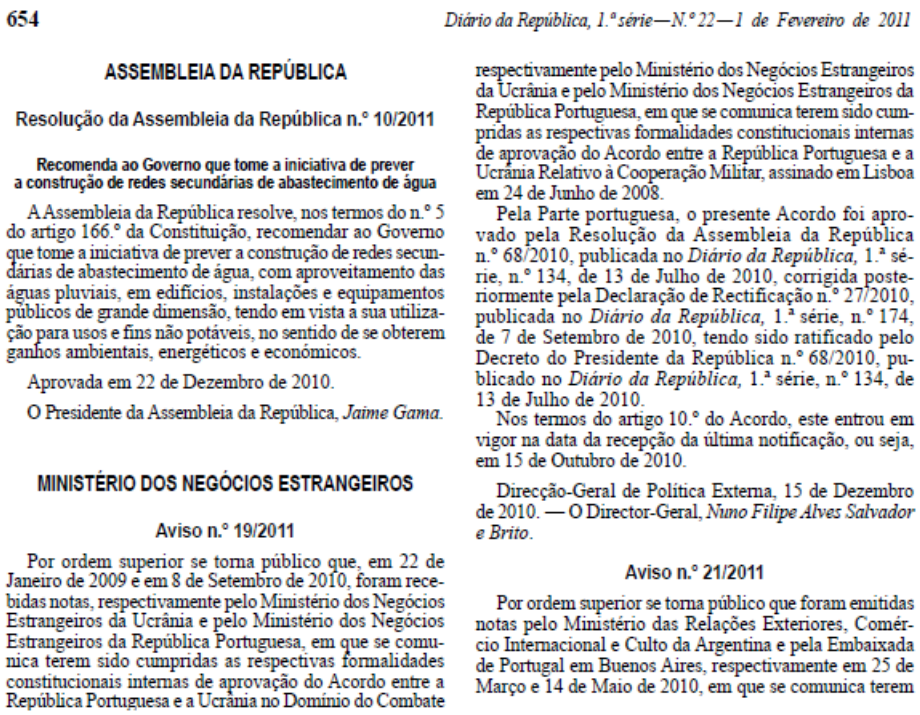
The development of this system poses three initial question related to the three phases in which this project is divided:

1. How to access and retrieve the data from the PDF documents collection?
2. How to organize the information extracted?
3. How to make it available for access and reuse?

1.3.2 Data Extraction – Phase 1

The first phase answers the question of how to access and retrieve the data from the PDF documents collection.

The objective of this first phase is to obtain the PDF document collection of Republic's Diaries in order to parse the various legislation documents present in each, and extract their metadata and named entities. To do so, first, it is necessary to ensure access to the Diaries; second, since the Republic's Diary may have more than one legislative text, it is necessary to divide the document into each of the units (legislative documents) that compose it; third, it is necessary to parse each of these units; fourth, after parsing each unit it is necessary to search and obtain the desired data. An example of a document is presented in Figure 2.



654 Diário da República, 1.ª série — N.º 22 — 1 de Fevereiro de 2011

ASSEMBLEIA DA REPÚBLICA

Resolução da Assembleia da República n.º 10/2011

Recomenda ao Governo que tome a iniciativa de prever a construção de redes secundárias de abastecimento de água

A Assembleia da República resolve, nos termos do n.º 5 do artigo 166.º da Constituição, recomendar ao Governo que tome a iniciativa de prever a construção de redes secundárias de abastecimento de água, com aproveitamento das águas pluviais, em edifícios, instalações e equipamentos públicos de grande dimensão, tendo em vista a sua utilização para usos e fins não potáveis, no sentido de se obterem ganhos ambientais, energéticos e económicos.

Aprovada em 22 de Dezembro de 2010.

O Presidente da Assembleia da República, *Jaime Gama*.

MINISTÉRIO DOS NEGÓCIOS ESTRANGEIROS

Aviso n.º 19/2011

Por ordem superior se torna público que, em 22 de Janeiro de 2009 e em 8 de Setembro de 2010, foram recebidas notas, respectivamente pelo Ministério dos Negócios Estrangeiros da Ucrânia e pelo Ministério dos Negócios Estrangeiros da República Portuguesa, em que se comunica terem sido cumpridas as respectivas formalidades constitucionais internas de aprovação do Acordo entre a República Portuguesa e a Ucrânia no Domínio do Combate respectivamente pelo Ministério dos Negócios Estrangeiros da Ucrânia e pelo Ministério dos Negócios Estrangeiros da República Portuguesa, em que se comunica terem sido cumpridas as respectivas formalidades constitucionais internas de aprovação do Acordo entre a República Portuguesa e a Ucrânia Relativo à Cooperação Militar, assinado em Lisboa em 24 de Junho de 2008.

Pela Parte portuguesa, o presente Acordo foi aprovado pela Resolução da Assembleia da República n.º 68/2010, publicada no *Diário da República*, 1.ª série, n.º 134, de 13 de Julho de 2010, corrigida posteriormente pela Declaração de Rectificação n.º 27/2010, publicada no *Diário da República*, 1.ª série, n.º 174, de 7 de Setembro de 2010, tendo sido ratificado pelo Decreto do Presidente da República n.º 68/2010, publicado no *Diário da República*, 1.ª série, n.º 134, de 13 de Julho de 2010.

Nos termos do artigo 10.º do Acordo, este entrou em vigor na data da recepção da última notificação, ou seja, em 15 de Outubro de 2010.

Direcção-Geral de Política Externa, 15 de Dezembro de 2010. — O Director-Geral, *Nuno Filipe Alves Salvador e Brito*.

Aviso n.º 21/2011

Por ordem superior se torna público que foram emitidas notas pelo Ministério das Relações Exteriores, Comércio Internacional e Culto da Argentina e pela Embaixada de Portugal em Buenos Aires, respectivamente em 25 de Março e 14 de Maio de 2010, em que se comunica terem

Figure 2 – Diary of the Republic document

After parsing the document and obtaining the information that is required, the second question is presented. The answer to the question of how should the information be organized in order to become useful is done by the second phase – Knowledge Organization.

1.3.3 Knowledge Organization – Phase 2

A global objective of this project is to develop a system that besides the description already provided, could be a project classified as an open government and open data initiative. Therefore, the concepts described in Section 1.2 are important for this explanation, especially the concept of Semantic Web and Linked Data.

In order to achieve this objective it is necessary that the organization of information is in accordance with the requirements of the Semantic Web. Furthermore, by handling data that is interlinked it is necessary that this can be translated in the organization of data. Therefore, in order to achieve this goal, the second phase aims at organizing the collected knowledge using an ontology and store that knowledge in a database assembled for this effect.

Still in the second phase, the development of the ontology is of the most importance and it will be detailed in Section 4.1. In general terms the development of the ontology has as a main goal to achieve interoperability with other already developed ontologies in the legal domain. Also, it is focused on representing the relation between documents and entities referred in the document text. An ontology was developed due to the nonexistence of a standard ontology for the legal domain. This ontology should incorporate three characteristics: small, interoperable and sharable.

Furthermore, this phase involved the investigation of possible database solutions to store the knowledge and its implementation. Also, to allow the connection between the system and the database, the connection between these two parts is developed and implemented.

1.3.4 Information Access – Phase 3

In the third and last phase, a response to the third question is developed, in order to transform the knowledge accessible and reusable in a manner that tallies with the definition of open data and open government concepts. Therefore, taking into account the description of the rising use of mobile technologies, this phase comprehends the development of a mobile solution that allows the use of this knowledge creating a different visualization and access to the information. In addition, a Web Service was developed in order for the knowledge to become accessible and reusable for other users.

1.4 Thesis

The previous and continuous use of proprietary formats in the process of divulging information poses a significant issue to the extraction of knowledge. The previously mentioned questions frame this issue in the terms set out by the previous sections:

1. How to access and retrieve the data from the PDF documents collection?
2. How to organize the information extracted?
3. How to make it available for visualization and reuse?

This dissertation is based on the idea that the use of proprietary format in the process of divulging information is not the best effort concerning posterior processing of that same information.

Furthermore, the developed system presents an answer to the previous questions and provides the output necessary to compare the possibilities of information and knowledge extraction from these silos of documents. The baseline of this thesis is the present conditions of access and reuse of the Electronic Republic's Diary and the comparison is made with the final results of the system.

It is expected that the translation of closed data silos into open data provides a better ability to access and reuse information. Additionally, it is expected that the translated open format documents provide better conditions for text processing and knowledge extraction from the study subject, the Portuguese Legislation.

1.5 Contributions

- Comprehensive investigation regarding information extraction from PDF documents
- Development of a methodology concerning extraction of text, structure, and entities in PDF documents

- Design and implementation of a system that enables the translation of PDF collections into non-proprietary formats (XML, RDF)
- Development of a Linked Data database of Portuguese Legislation that enables free access and information reutilization
- Comparative study of available solutions regarding ontologies in the legal domain
- Use of a legal domain ontology to store the information extracted from the Portuguese Republic's Diaries PDF documents
- Development of a Web Service capable of sharing the knowledge that the developed system extracted

1.6 Document Organization

This dissertation is organized as follows:

This first chapter outlines a scope of the problematic that is in the core of this project and background on the areas and topics in which this dissertation is circumscribed. Also, a general description of the approach used, and its different phases is presented. Lastly, the technologies and the intended contributions of this project are presented.

The second chapter presents and describes the developed system description. It provides a description of each of the components that are part of the system as well as the relation amongst them.

The third, fourth and fifth chapter describes thoroughly the three phases of the developed project: Data Extraction, Knowledge Organization and Information Access. In these chapters, the composition in terms of components referred in the last section is described for each of the phases. This is done so that the organization of the system and the connection and relation between them is made clear.

The sixth chapter shows and explains the constraints of the implementation of the developed system in an online server and the seventh chapter presents the results of the system. Chapter eight contains conclusions, further work and achievements.

2 System Description

As stated in the previous chapter, the main objectives of this research are: accessing the legislative information available in the DRE; investigating how to translate that information from PDF format to an open format and; finally, how to make it accessible and reusable to others. As also stated in the previous chapter, the project is divided in three phases: information extraction, knowledge organization, information access.

The project is to be constituted by various components, corresponding to the various phases and processes necessary to obtain the desired result. The desired result of this system is to extract, organize and enable the access and re-use of knowledge in the legal domain. This chapter will describe the necessary components and its functionalities.

2.1 Components

In accordance with the referred objectives the main processes that are required for the project to obtain its goals are presented:

- Accessing the Portuguese Republic's Diaries PDF documents
- Parsing of the PDF documents
- Extraction of text, structure and entities from the documents
- Organization and storing of the data extracted
- Indexing of the documents text
- Enable the access and reuse of the information
- Development of a mobile application to access the information

Through the analysis of the referred processes, it is possible to divide the implementation in three parts, regarding the components: the data extraction and knowledge organization system, the information provider component and the mobile application. The first part corresponds to the first two phases of the project: data extraction and knowledge organization; the second and third part corresponds to the information access phase.

Regarding the processes, the first phase, data extraction, corresponds to the first three processes pointed out in the previous list; the second phase, knowledge organization, corresponds to the processes four and five. The third phase, information access, corresponds to the two last processes.

The proposal of components that incorporate the described processes is the following: PDF Parser, Ontology, Index, Web Service and Mobile Application.

The PDF Parser is responsible for accessing the Portuguese Republic's Diaries PDF documents, parsing the PDF documents and the extraction of text, structure and entities from the documents;

the Ontology is responsible for the organization and storing of the data extracted; the Index is responsible for indexing the documents text; the Web Service is responsible for enabling access and reuse of the information to third parties; and the Mobile Application is the means of access to the information, through the use of the available resources in the Web Service.

In terms of general architecture, the components of the first part – PDF Parser, Ontology and Index – are connected in terms of input of output. The results of the three processes that compose the PDF Parser are inputs for the Ontology and Index. The Web Service uses the information from these two components in order to respond to the users' request of information. The general architecture is illustrated as follows.

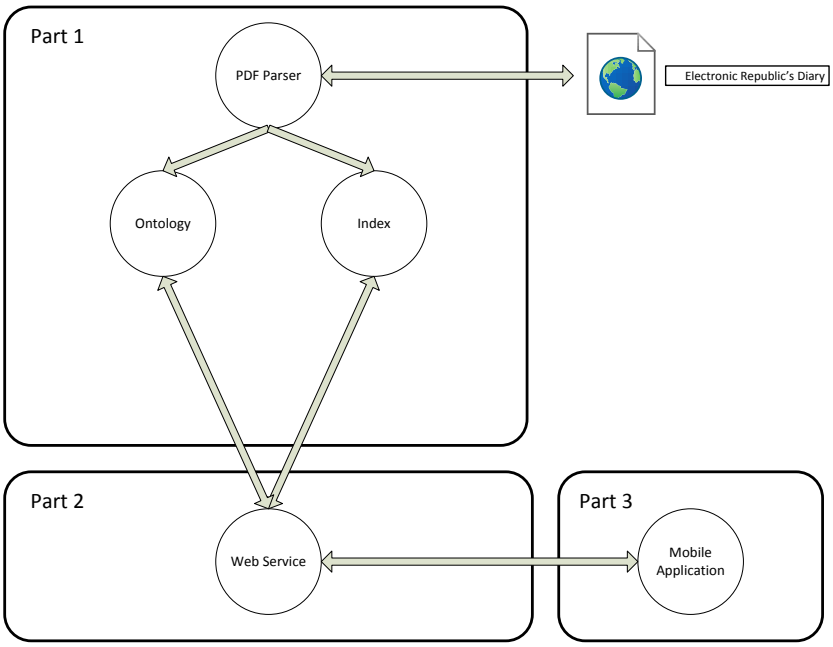


Figure 3 – General Architecture

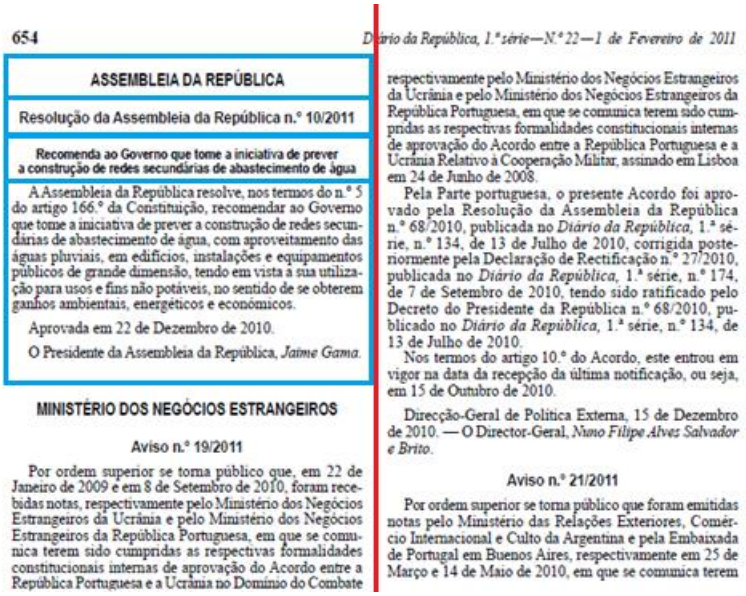
The previous illustration represents the coarse grain general architecture. It demonstrates the composition of the system and the connections between the various components. In the following sections, each of the components will be described in terms of functionalities, relation with other components and its necessary sub-components.

2.1.1 PDF Parser

The PDF Parser component integrates the first part of the implementation and represents the answer to the first phase, data extraction. This component should be executed given a specific time interval. Taking into account the observation of the legislative process, that time interval should be of one day. Besides the daily execution of the PDF Parser component, it is also necessary to extract information of previous Republic's Diaries; this component is also responsible for this action. The PDF Parser has as main goals the access to the Republic's Diaries documents, the parsing of the documents and the extraction of its structure, text and entities.

The obtained results represent the inputs of the components Ontology and Index. In order to do this sequentially, it is necessary for the interpretation of the PDF documents to be automatic.

Towards the implementation of that interpretation and data extraction, it is necessary to engage in an exhaustive study of the legislative documents in order to assert its composition in terms of information and structure. For example, the following illustration of Figure 2 contains a few indications concerning the document's structure.



654 Diário da República, 1.ª série — N.º 22 — 1 de Fevereiro de 2011

ASSEMBLEIA DA REPÚBLICA

Resolução da Assembleia da República n.º 10/2011

Recomenda ao Governo que tome a iniciativa de prever a construção de redes secundárias de abastecimento de água

A Assembleia da República resolve, nos termos do n.º 5 do artigo 166.º da Constituição, recomendar ao Governo que tome a iniciativa de prever a construção de redes secundárias de abastecimento de água, com aproveitamento das águas pluviais, em edifícios, instalações e equipamentos públicos de grande dimensão, tendo em vista a sua utilização para usos e fins não potáveis, no sentido de se obterem ganhos ambientais, energéticos e económicos.

Aprovada em 22 de Dezembro de 2010.

O Presidente da Assembleia da República, *Jaime Gama*.

MINISTÉRIO DOS NEGÓCIOS ESTRANGEIROS

Aviso n.º 19/2011

Por ordem superior se torna público que, em 22 de Janeiro de 2009 e em 8 de Setembro de 2010, foram recebidas notas, respectivamente pelo Ministério dos Negócios Estrangeiros da Ucrânia e pelo Ministério dos Negócios Estrangeiros da República Portuguesa, em que se comunica terem sido cumpridas as respectivas formalidades constitucionais internas de aprovação do Acordo entre a República Portuguesa e a Ucrânia no Domínio do Combate

respectivamente pelo Ministério dos Negócios Estrangeiros da Ucrânia e pelo Ministério dos Negócios Estrangeiros da República Portuguesa, em que se comunica terem sido cumpridas as respectivas formalidades constitucionais internas de aprovação do Acordo entre a República Portuguesa e a Ucrânia Relativo à Cooperação Militar, assinado em Lisboa em 24 de Junho de 2008.

Pela Parte portuguesa, o presente Acordo foi aprovado pela Resolução da Assembleia da República n.º 68/2010, publicada no *Diário da República*, 1.ª série, n.º 134, de 13 de Julho de 2010, corrigida posteriormente pela Declaração de Rectificação n.º 27/2010, publicada no *Diário da República*, 1.ª série, n.º 174, de 7 de Setembro de 2010, tendo sido ratificado pelo Decreto do Presidente da República n.º 68/2010, publicado no *Diário da República*, 1.ª série, n.º 134, de 13 de Julho de 2010.

Nos termos do artigo 10.º do Acordo, este entrou em vigor na data da recepção da última notificação, ou seja, em 15 de Outubro de 2010.

Direcção-Geral de Política Externa, 15 de Dezembro de 2010. — O Director-Geral, *Nuno Filipe Alves Salvador e Brito*.

Aviso n.º 21/2011

Por ordem superior se torna público que foram emitidas notas pelo Ministério das Relações Exteriores, Comércio Internacional e Culto da Argentina e pela Embaixada de Portugal em Buenos Aires, respectivamente em 25 de Março e 14 de Maio de 2010, em que se comunica terem

Figure 4 – Example of a Republic's Diary document structure

It is possible to observe that the documents content is organized in two columns. Furthermore, this document contains the indication of the entity responsible for the legislation. Responsible entity should be understood as the group of organisms responsible for the legislation. In the case above illustrated, the responsible entity is the Parliament (Assembleia da República); it contains the title of the legislation, the subtitle or description of the legislation and the text of the legislation.

After the investigation of the Republic's Diaries documents, using a group of 40 documents, it was concluded that some of these information's may or not be described in the legislation. This question will be addressed and further explained in Section 3.1.1. By understanding this structure, it is possible to implement the automatic extraction of data in order to proceed to its analysis, namely, to execute the recognition and extraction of entities.

Furthermore, in order to enable the access and reuse of legislative documents, it is necessary to create copies of the documents. Therefore, it is necessary that this component should translate the original PDF documents into an open format in order to comply with the open data definition referred before. Lastly, this information is used by the component Ontology and the component Index. The following illustration presents the organization of the PDF Parser component.

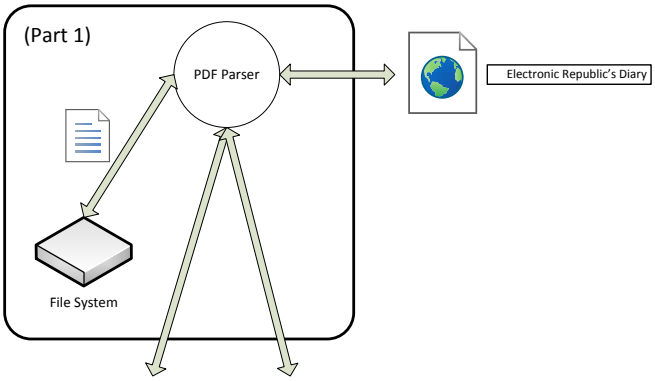


Figure 5 – PDF Parser component

Therefore, we can conclude that the processes made by this component should be as follows:

- Obtain the document
- Extract structure and text
- Recognize and extract entities
- Create an open format copy of the PDF document
- Insert the information into the component Ontology and component Index

This process is described in the following figure.

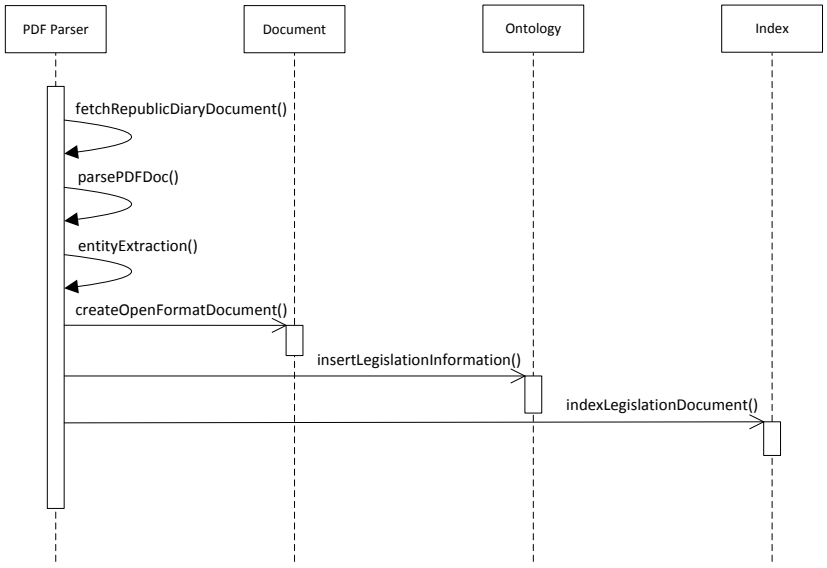


Figure 6 – Sequence Diagram of PDF Parser processes

2.1.2 Ontology

The Ontology component integrates the first part of the implementation and is integrated in the second phase, knowledge organization. This component should be permanently available either for the inputs from the component PDF Parser, or to provide response to the requests of the component Web Service. The Ontology component has as main goals the tasks of organizing and storing the legislative information. In order to pursue this objective it should have resources implemented and available that allows these operations, namely a database. In order to organize the knowledge in a manner that would be compliant with its use in a Semantic Web environment, the organization of knowledge is made through the use of ontologies.

The inputs that this component receives are the information feeds from the PDF Parser component. These feeds should be interpreted and stored in a database using a proper ontology given the domain that this project is inserted. The results represent inputs for the Web Service component according to the respective queries.

This component incorporates the results of the study of the Republic's Diaries documents. This is made through the development (or reuse) of an ontology capable of organizing and representing the knowledge extracted from the documents. In order to pursue this objective, it is necessary to study the legislation conceptually and investigate the available solutions regarding ontologies in the legal domain. The results of this investigation are presented in Section 4.1.

Regarding the conceptual study of the Portuguese Legislation through the observation of the Republic's Diary documents, it is asserted that from an abstract point of view, there are two main concepts represented in a legislation document that are of interest to our objectives: entities and documents.

In the context of this project the entity concept is defined as the reference to something that is formally understood as a person, a group of persons, an organization, locations, and others. The reference may or not include its name. For example, the reference to Prime-Minister is to be understood as a reference to the person and not the position. Regarding the document concept, it is defined as the reference to a work or a manifestation according to the IFLAs Functional Requirements for Bibliographic Records (FRBR) [Madison, 2000].

By integrating this conceptualization with the developed study of the Republic's Diary, it is concluded that both the entity and the document concept have a few different interpretations regarding the legal domain and its expression in the Portuguese legislation case. For example, in the case of the entity concept, it can represent people, groups of people or organizations; regarding the document concept, it can represent Republic's Diaries, legislation or legislation attachments.

The following figure illustrates the proposed organization of concepts according to the study of the Portuguese legislation and the objectives of the project.

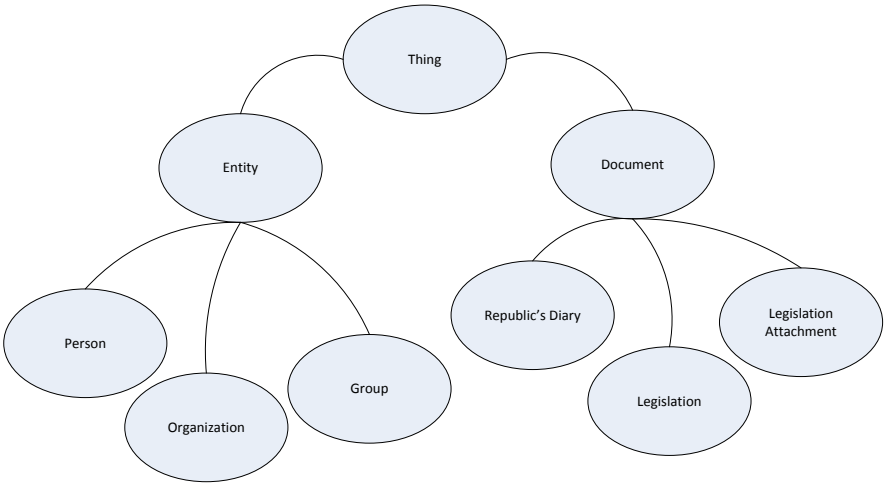


Figure 7 - Proposed organization of concepts

Therefore, the execution of this component is as such:

- Receive information feeds from the component PDF Parser
- Store the information in a database using an ontology

The following figure illustrates the organization of this component.

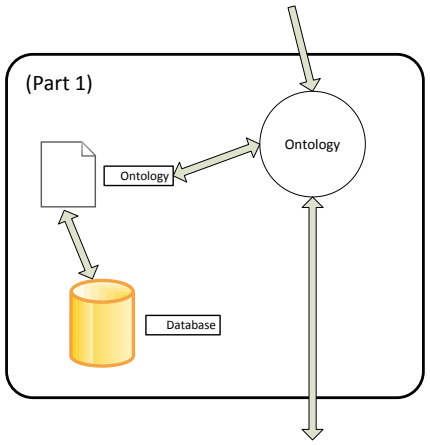


Figure 8 - Ontology component

2.1.3 Index

The Index component integrates the first part of the implementation and the second phase, knowledge organization. This component should be permanently available either for the inputs from the component PDF Parser, either to provide response to the requests of the component Web Service.

The main goal of the Index component is to index the legislative text of each of the legislation documents, which represents the input given by the PDF Parser component.

The results of the process of indexing are stored in a separate file. The results of this component represent inputs for the Web Service component, mainly related to requests of text-based search.

Therefore, the execution of this component is as such:

- Receive document text feeds from the component PDF Parser
- Index the document text
- Store the results in a separate file

The following figure illustrates the organization of this component.

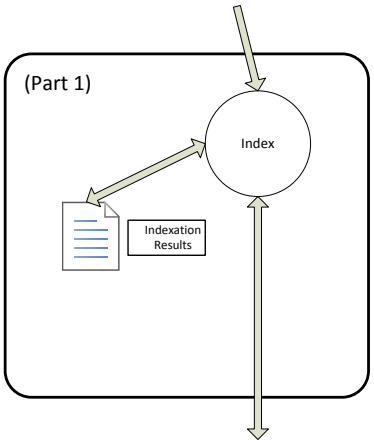


Figure 9 - Index component

2.1.4 Web Service

The Web Service component integrates the second part of the implementation and the third phase, information access. This component should be permanently available in order to respond to the requests addressed to it. Its main goal is to provide access to the knowledge regarding the Portuguese Legislation. It represents the point of access to the system in the user perspective. In order to pursue this objective, the Web Service publishes a list of procedures available, which the users can access in order to retrieve information.

This Web Service uses REST technology. The discussion concerning the options and the decision is presented in Section 5.1. This decision implies that the structure of the Web Service should be well formed, taking into consideration the desired functionalities of the component.

This component uses inputs from the components Ontology and Index. The component Index is only used in requests related to text-based searches. The results of this component represent the delivered responses to requesting users.

The main decision regarding the Web Service component is deciding the list of available procedures in order to access the information. Taking into consideration the previous remarks regarding the

composition of the Republic's Diary and the legislation documents and the organization of knowledge, a list of procedures is presented.

- Obtain the list of legislations present in a specific Republic's Diary
- Obtain the last ten published legislation
- Obtain information regarding a specific legislation
- Obtain information regarding a specific entity
- Obtain the list of entities referenced in the same documents as a specific entity
- Obtain all available versions of a specific legislation

After the presentation of the list above, it is possible to design the structure of the Web Service regarding the resources available for the requests. This is described in the following table.

Table 1 – List of Web Service resources

Resource	**Procedures**
http://.../republicdiary/[diary_id]	Obtain the list of legislations present in a specific Republic's Diary
http://.../legislation/last	Obtain the last ten published legislation
http://.../legislation/[legislation_id]	Obtain information regarding a specific legislation
http://.../entity/[entity_id]	Obtain information regarding a specific entity
http://.../entity/entities	Obtain the list of entities referenced in the same documents as a specific entity
http://.../legislation/[id]/version	Obtain all available versions of a specific legislation

Concerning the implementation of the Web Service component, it is expectable that the list should be different. In any case, this list represents the core procedures that this Web Service should make available. The following figure illustrates the organization of this component.

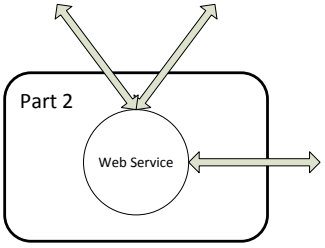


Figure 10 – Web Service component

2.1.5 Mobile Application

The Mobile Application component integrates the third part of the implementation and the third phase, information access. The main goal of this component is to serve as an interface to the Web Service component, and therefore to the knowledge stored in the system.

This component is responsible for the interaction between the user and the system. It should implement an interface capable of requesting and presenting the information available through the procedures available in the Web Service component. The component is developed for the Android platform[15].

The main goal of this component is implemented through the development of an interface capable of requesting information from the Web Service (see Procedures in Table 1) and the visualization of the results. The requests are made through the use of HTTP methods on the resources of the Web Service (see Resources in Table 1). The response is returned to the Mobile Application component in an open format document. The execution of this component is as such:

- Request information from the Web Service
- Interpret the responses and create visualizations for the information

The following figure illustrates the organization of this component.

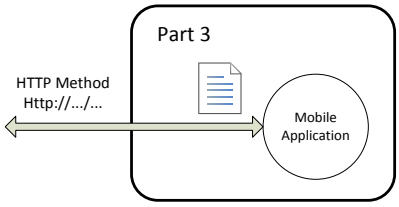


Figure 11 – Mobile Application component

[15] Android – http://www.android.com

This chapter presented the system description and the description of each of the components that compose it. The following figure illustrates the general architecture presented before, updated with the specifications of each component.

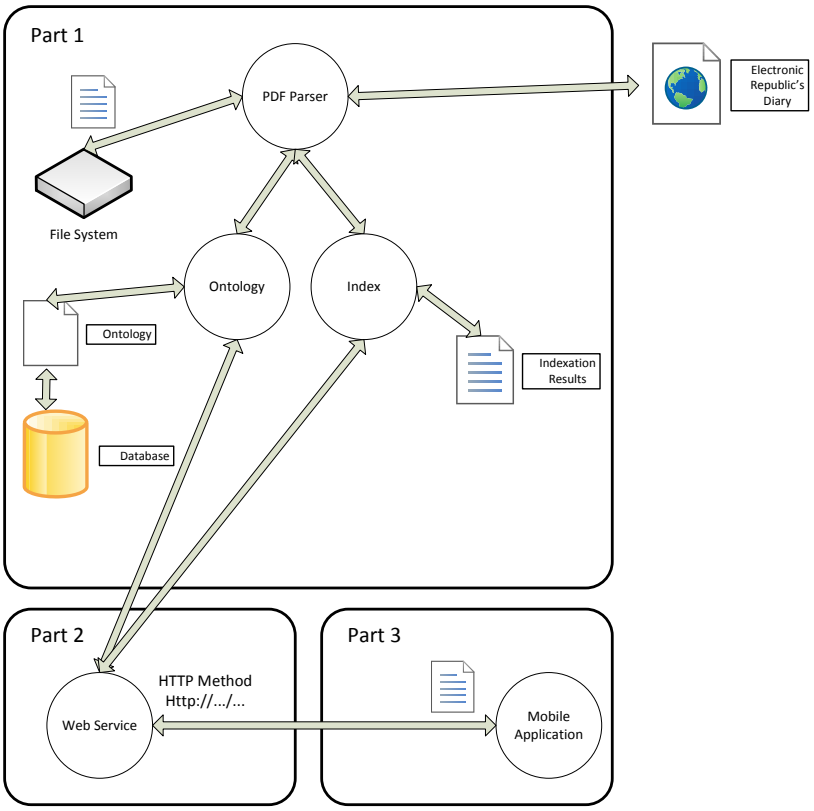


Figure 12 – Updated version of the General Architecture

3 Data Extraction – Phase 1

As the progressive surge of tools to markup XML in the legislative process continues, the problem of the previous published legislation still lingers. To this day, there are file systems of years containing legislation documents, namely in the Portuguese case. Therefore, in order to pursue the translation of proprietary formats into open access files, it is necessary to develop efforts in the translation of these databases. This is the main objective of this phase.

This phase comprehends the study of technologies, tools and strategies to extract information from PDF documents. It also comprehends the development of tools that using this study can analyse the Republic's Diary documents and extract data from them. In order to do so, it is necessary to study the composition of these documents. This study is presented in this chapter.

This phase is divided in three parts: investigation regarding information extraction from PDF documents, study of the Republic's Diary documents composition and development of tools in order to extract the required data.

The first part of this phase represents the development of an intensive study concerning the organization of information in the Republic's Diary documents with the objective of studying the information hierarchy [Taylor, 1999] of the documents. The second part comprehends the study of technologies, strategies and tools that have the ability to extract information from PDF documents. The third part of this phase incorporates the study made in the first and second part in order to develop the adequate tools required to extract information according to the objectives set out in terms of information. The type of information to extract is also defined in the study of the composition of Republic's Diary documents.

The final outcome of this phase is the division of the Republic's Diary documents into its various legislation documents, the extraction of data from each of those units and the creation of a copy of the document in an open format.

The collection of the Republic's Diary PDF documents is hosted online in the Electronic Republic's Diary (DRE) website[16]. A depiction of the website is presented in Figure 13. This website is hosted and managed by Instituto Nacional – Casa da Moeda[17]. The DRE website has been publishing the Republic's Diary documents in PDF format, and not a scanned version of the document, since the years 1996-1997. Nonetheless, due to technical issues we could only apply this project to the documents created since the year 2009. This technical issue will be developed in Section 3.2.

[16] Diário da República Electrónico – http://www.dre.pt
[17] Instituto Nacional – Casa da Moeda – http://www.incm.pt

Figure 13 – Depiction of the DRE website

The Republic's Diary is divided into three series, although the third has been discontinued since the 30th of June 2006. The series can be seen in the centre menu of Figure 13. The first series publishes the laws, decrees, resolutions, regiments, decisions and declarations from the Constitutional Court and others. The second series publishes the government members' dispatches, election results, budgets and all other documents that are legally obligated to be published besides all the previous ones. The focus of the development of this project will be only in the first series.

3.1 PDF Documents

This section presents the study of the Republic's Diary documents structure and the study of its information hierarchy. The first study comprehends the analysis and comprehension of the different parts that compose the Diary but also of the legislation itself. The final outcome is the composing parts of the documents. The second study concerns information hierarchy [Taylor, 1999], where the output of this study should be the conclusion on how the various parts of legislation are presented and its differences in terms of positioning and representation. Both of these studies are inputs for the development of tools that extract information from the PDF documents collection of the Republic's Diary.

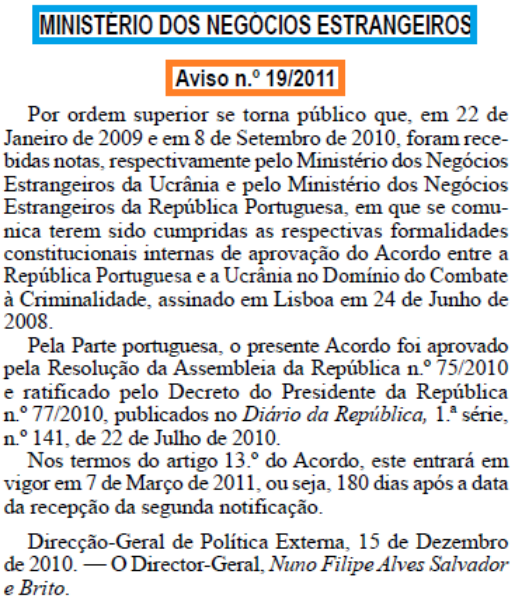
3.1.1 Document Structure

This section presents the analysis and comprehension of the different parts that compose the Republic's Diary documents and its composing legislation documents. This was done through the observation of a group of 40 Republic's Diary documents. The list of documents is presented in

Attachment 1. Although some of the elements of legislation documents are obligatory, others are optional. Nonetheless, some are used in most of the documents.

This study is of the most importance because, in order to transform the lack of formal indication of structure in PDF documents into knowledge it is necessary to understand it first. Only then it is possible to encode this knowledge and develop algorithms capable of bridging this gap between text documents and computer processable representations.

Amongst the group of documents that provide the basis of this study, three cases can be presented as paradigmatic which cover most of the results.



MINISTERIO DOS NEGÓCIOS ESTRANGEIROS

Aviso n.º 19/2011

Por ordem superior se torna público que, em 22 de Janeiro de 2009 e em 8 de Setembro de 2010, foram recebidas notas, respectivamente pelo Ministério dos Negócios Estrangeiros da Ucrânia e pelo Ministério dos Negócios Estrangeiros da República Portuguesa, em que se comunica terem sido cumpridas as respectivas formalidades constitucionais internas de aprovação do Acordo entre a República Portuguesa e a Ucrânia no Domínio do Combate à Criminalidade, assinado em Lisboa em 24 de Junho de 2008.

Pela Parte portuguesa, o presente Acordo foi aprovado pela Resolução da Assembleia da República n.º 75/2010 e ratificado pelo Decreto do Presidente da República n.º 77/2010, publicados no *Diário da República*, 1.ª série, n.º 141, de 22 de Julho de 2010.

Nos termos do artigo 13.º do Acordo, este entrará em vigor em 7 de Março de 2011, ou seja, 180 dias após a data da recepção da segunda notificação.

Direcção-Geral de Política Externa, 15 de Dezembro de 2010. — O Director-Geral, *Nuno Filipe Alves Salvador e Brito*.

Figure 14 – Example 1 of the study of document structure

The previous figure presents the most basic case found in the legislative documents. Regarding structure, it contains the issuing entity (in blue), the title of the legislation (in orange) and the legislative text. This case presents the units that are obligatory in a legislative document header.

REGIÃO AUTÓNOMA DOS AÇORES

Presidência do Governo

Decreto Regulamentar Regional n.º 4/2011/A

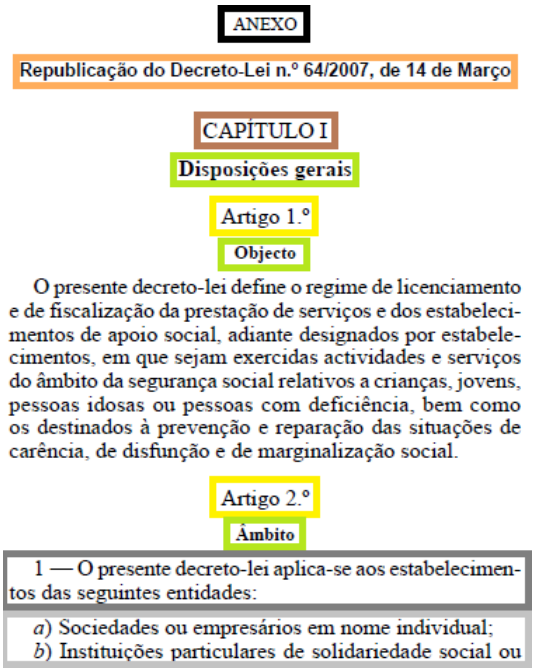
Aprova a orgânica da Secretaria Regional da Ciência, Tecnologia e Equipamentos

O Decreto Regulamentar Regional n.º 25/2008/A, de 31 de Dezembro, que aprovou a orgânica do X Governo Regional dos Açores, procedeu a vários ajustamentos na estrutura do Governo Regional numa perspectiva de adequação e eficiência dos seus órgãos e serviços em cada uma das áreas de intervenção governativa.

A Secretaria Regional da Ciência, Tecnologia e Equipamentos emerge, assim, da referida reestruturação or-

Figure 15 – Example 2 of the study of document structure

The previous figure presents the most elaborate case found regarding the legislative documents header. Regarding structure, it contains the issuing entity (in blue), an issuing sub-entity (in dark blue), the title of the legislation (in orange), the description (in green) and the legislative text. This case presents the maximum units that are possible in a legislative document header.


ANEXO

Republicação do Decreto-Lei n.º 64/2007, de 14 de Março

CAPÍTULO I

Disposições gerais

Artigo 1.º

Objecto

O presente decreto-lei define o regime de licenciamento e de fiscalização da prestação de serviços e dos estabelecimentos de apoio social, adiante designados por estabelecimentos, em que sejam exercidas actividades e serviços do âmbito da segurança social relativos a crianças, jovens, pessoas idosas ou pessoas com deficiência, bem como os destinados à prevenção e reparação das situações de carência, de disfunção e de marginalização social.

Artigo 2.º

Âmbito

1 — O presente decreto-lei aplica-se aos estabelecimentos das seguintes entidades:

a) Sociedades ou empresários em nome individual;
b) Instituições particulares de solidariedade social ou

Figure 16 – Example 3 of the study of document structure

The previous figure presents an example of an attachment with most of the units that may be part of legislative documents. It contains indication of attachment (in black), subtitles (in light orange), chapter (in brown), articles (in yellow), paragraph (in dark grey), lines (in light grey) and legislative text. This case presents most of the examples of units that may be present in the body of legislative documents.

Additionally, there were encountered references to Section and Subsection units.

The following table presents the results of this study by listing the findings classifying them as being obligatory or optional.

Table 2 – Results of the study of document structure

Legislation Elements	**Obligatory**	**Optional**
Issuing Entity	■	
Issuing Sub-Entity		■
Title	■	
Description		■
Articles		■
Paragraphs		■
Lines		■
Chapter		■
Section		■
Sub-Section		■
Text	■	
Attachments		■
Attached Documents		■

Additionally, some final remarks concerning the document's structure:

- Republic's Diary documents contain a cover, index and one or more legislative documents
- The cover and index size is not one page in every Diary
- Documents may or not contain one or more attachments

3.1.2 Information Hierarchy

It is possible to state that a considerable number of public and private organizations that issue official documents regularly adopt well-defined layout structures. These standards include not only the geometric position of text but also its hierarchical structure - differenced fonts, styles and positioning. Using a combination of hereditary and acquired knowledge, we can understand the structure of complex documents without significant effort [Hassan, 2010].

This section presents the study of the hierarchical structure of the Republic's Diary documents, taking into consideration the findings in the study presented before concerning the document's structure. The objective of this study is to observe and describe the hierarchical structure of the documents taking into the account the geometric position, the font used in the text, as well as its size and style. This implies understanding the hierarchy of the document's structure.

In order to accomplish this, the information resulting from the previous study is used and the same group of 40 Republic's Diary documents (see Attachment 1) is analysed again, in order to observe and assess the hierarchical structure of the documents. The findings of this study are presented as follows.

The conclusions concerning hierarchy of the header of legislative documents in Republic's Diary documents are illustrated in the following figure.

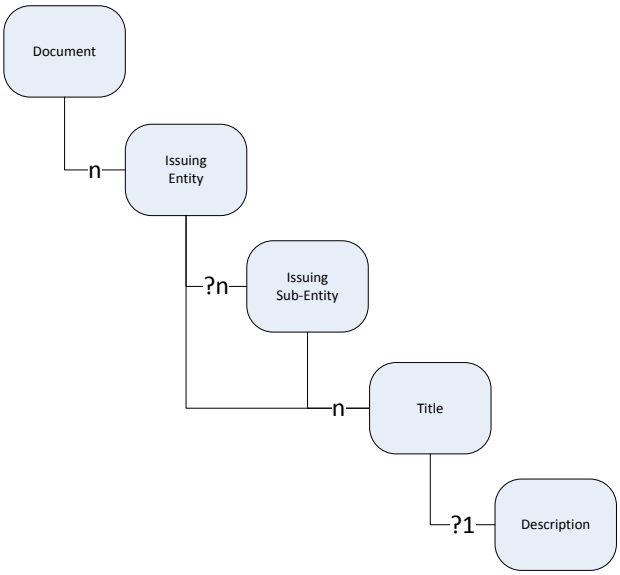


Figure 17 – Hierarchical Structure of a Republic's Diary document (header)

This illustration shows that a Republic's Diary document may contain one or more issuing entities; each of these issuing entities may or may not have one or more issuing sub-entities and in either case, each entity/sub-entity may have one or more titles; the titles represent legislative documents that may or may not have a description.

Concerning the body of the legislative documents in Republic's Diary documents, the following figure illustrates its hierarchical structure.

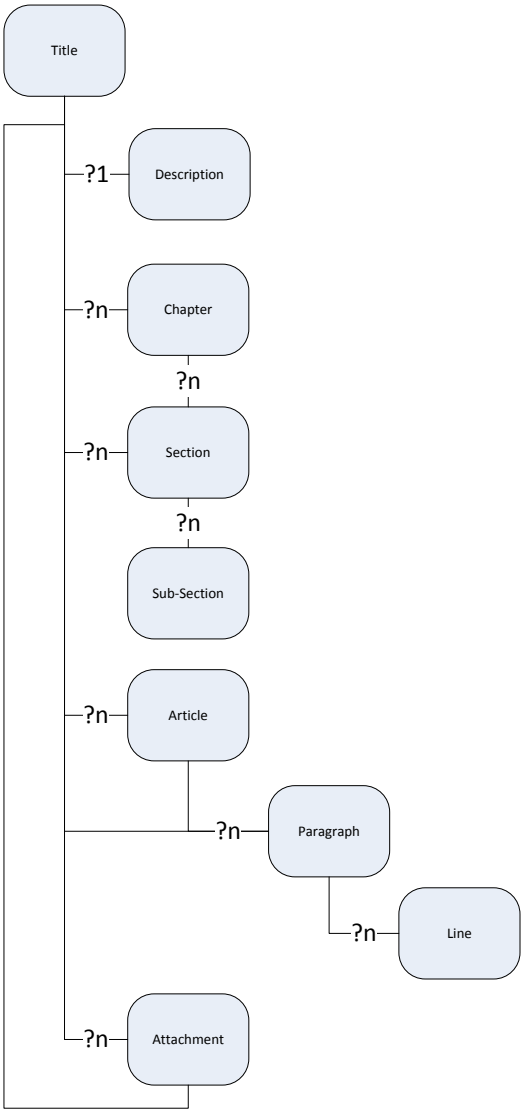


Figure 18 – Hierarchical Structure of a Republic's Diary document (body)

This illustration shows that the unit title which represents a legislative document may or may not have a description; may or may not have one or more chapters and/or sections; may or may not have one or more articles, paragraphs and attachments; each chapter may or may not contain one or more sections and each section may or may not contain one or more sub-sections; each article may or may not contain one or more paragraph and each paragraph may or may not contain one or more line; an attachment replicates the hierarchical structure described. Text units are always present in a legislation document.

3.2 Extraction Process – PDF Parser

Extracting text from a PDF document is not a direct and simple task. In our research we conclude that OCR [Mori et al., 1999] is the technology used in most cases ([Taylor et al., 1994], [Klink and Kieneger, 2001], [Todoran et al., 2001], [Hollingsworth et al., 2005]) due to the attempt to perform

text extraction on documents where there is no knowledge of its document's structure. However, in most of the cases mentioned it was concluded that OCR is time consuming and had issues in error recognition. The Republic's Diary document's structure was studied, and therefore it is not necessary the use of OCR technology. Today there is available technology to parse directly information from PDF documents. The discussion concerning technology used is presented later.

This approach and similar approaches of directly parsing information from PDF documents were used or described in the developed research ([Hassan and Baumgartner, 2005], [Antonacopoulos and Coenen, 1999], [Rosenfeld et al., 2002], [Siefkes, 2003]). For grouping text objects these approaches mainly use the font size as criterion for grouping text objects.

In this section an approach [Moniz and Rodrigues, 2012] for text processing of PDF documents with well-defined layout structures is presented. This approach uses two different extraction methods, according to the two stages of document processing - document analysis and document understanding [Hassan, 2010]. The criterion used for grouping the text objects is the PDF's content stream font definition used.

This description of the approach is organized as follows: the section PDF describes the relevant information concerning the PDF documents and its organization; the section Technologies describes the study regarding technologies available for extracting PDF documents content; the section General Description contains the overall description of the approach and the presentation and description of each of its three phases; section Implementation describes the phases and processes of the implementation of the approach; the last section, Background, presents the discussion concerning similar approaches studied.

3.2.1 PDF

PDF uses a structured binary file format described by a derivation of PostScript page description language. Objects are the basic data structure in a PDF file. For the purposes of this paper we elaborate some of the elements. The content stream is a stream object that contains the sequence of instructions that describe the graphical elements of the page. A dictionary object is an associative table containing key/value pairs of objects. A name object is an atomic symbol uniquely defined by a sequence of characters [Adobe, 2008]. In the structure of a PDF we are able to find the fonts dictionary in the resources dictionary. An example from Adobe Systems Incorporated [Adobe, 2008] is presented in the following figure.

```
3  0  obj
   <<  /Type  /Page
       /Parent  4 0 R
       /MediaBox  [0  0  612  792]
       /Resources  <<  /Font  <<  /F3  7 0 R
                                  /F5  9 0 R
                                  /F7  11 0 R
                               >>
                   /ProcSet  [/PDF]
               >>
       /Contents  12 0 R
       /Thumb  14 0 R
       /Annots  [  23 0 R
                   24 0 R
                ]
   >>
endobj
```

Figure 19 – Example of PDF Font Dictionary

In text objects from the content stream of a PDF file it is possible to find both objects: name and string. An example from Adobe Systems Incorporated [Adobe, 2008] is presented.

```
BT
   /F13  12  Tf
   288  720  Td
   (ABC)  Tj
ET
```

Figure 20 – Example of PDF Content Stream Text Object

3.2.2 Technologies

Towards the process of information extraction from PDF documents five alternatives were tested. These alternatives were programmed in languages Python[18] and Java[19]. The chosen alternative was programmed in Java. The following table lists the tested alternatives.

[18] Python Programming Language – http://www.python.org/
[19] Java – http://www.java.com

Table 3 – PDF parsing libraries

Solution	Programming Language	Link
PDFMiner	Python	http://www.unixuser.org/~euske/python/pdfminer/
Pdf-parser	Python	http://blog.rubypdf.com/2009/10/19/pdf-tools/
PDFTOHTML	Python	http://pdftohtml.sourceforge.net/
PDFBox	Java	http://pdfbox.apache.org
iText	Java	http://itextpdf.com

Tests were carried out using each of the studied libraries and the results were analysed. The results are presented as follows.

The alternatives PDFMiner and PDFTOHTML were discarded due to inability to support the information extraction from the Republic's Diary documents. The alternative Pdf-parser proved to be quite deprecated. It doesn't support some of the technical needs that will be necessary to extract information from the documents. The alternatives PDFBox and iText allow the extraction of information from the documents, and as such, more tests were carried out. The alternative PDFBox only allows the extraction of text although it requires the creation of models based on details such as font difference, character size and effects; the iText does not have the limitation that the previous alternative shows. Nonetheless, iText requires an import of other libraries that allows the type of access to the content stream of the documents that is necessary. Regarding the tests carried out and the former considerations the chosen alternative was the iText solution.

3.2.3 General Description

PDF document processing can be divided into two phases referring to the two structures in a document: document analysis in order to extract the layout structure and document understanding for mapping the layout structure into a logical structure [Klink et al., 2000]. Our approach is divided in three phases: the previous described phases and a third that combines the previous outputs.

3.2.3.1 Document Analysis

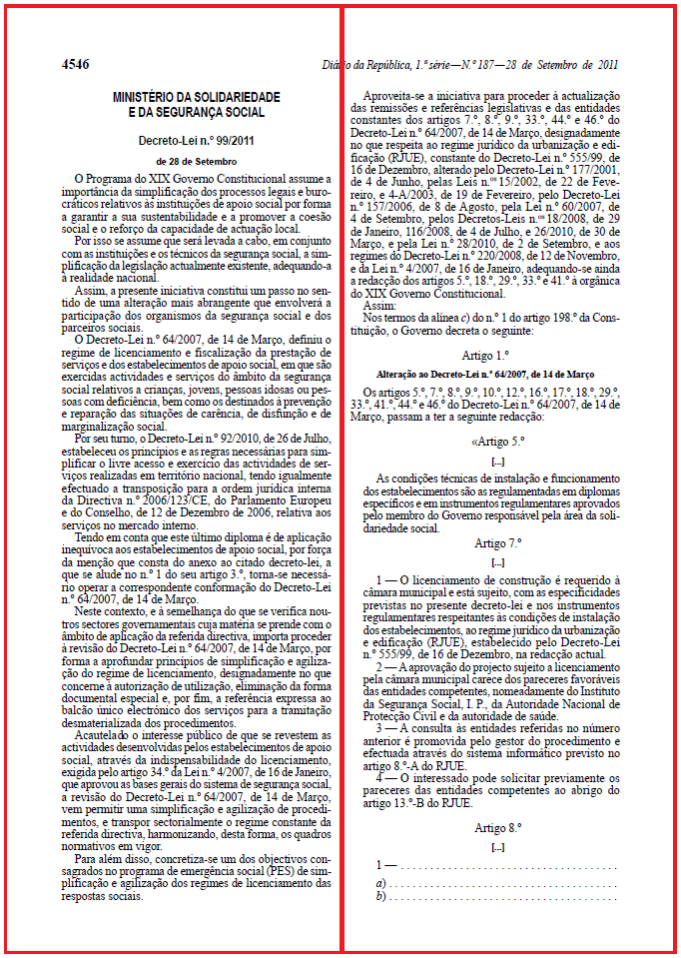
The first step in document analysis is layout analysis or segmentation. It consists on parsing a document into atomic blocks. According to our research there are two approaches for segmentation: top-down and bottom-up.

The top-down approach is an OCR simulation that usually makes use of whitespace density graphs or similar. This consists on parsing the documents along the x and y axis in order to find whitespace areas. Reports of block recognition problems in certain layouts were found [Hassan and Baumgartner, 2005].

The bottom-up approach can be described as a parsing and grouping process of the smallest segments that share a group of common characteristics such as font size [Hassan and Baumgartner, 2005].

In terms of region comparison, the discussion was based in research made by [Antonacopoulos and Coenen, 1999], where two categories of methods for region comparison are described: pixel-based and geometric. The geometric is described as the best approach, but the authors openly state their reservations of this approach due to the need of accurate descriptions of the regions.

Regarding segmentation our intended output is not a hierarchical structure but only the coarse-grained regions of each page of the documents, representing in our approach the two halves of the document page, as shown in Figure 21. This option will be elaborated further. In the given example, the graphic regions are defined by vertical ruling.



4546 *Diário da República, 1.ª série — N.º 187 — 28 de Setembro de 2011*

MINISTÉRIO DA SOLIDARIEDADE E DA SEGURANÇA SOCIAL

Decreto-Lei n.º 99/2011

de 28 de Setembro

O Programa do XIX Governo Constitucional assume a importância da simplificação dos processos legais e burocráticos relativos às instituições de apoio social por forma a garantir a sua sustentabilidade e a promover a coesão social e o reforço da capacidade de actuação local.

Por isso se assume que será levada a cabo, em conjunto com as instituições e os técnicos da segurança social, a simplificação da legislação actualmente existente, adequando-a à realidade nacional.

Assim, a presente iniciativa constitui um passo no sentido de uma alteração mais abrangente que envolverá a participação dos organismos da segurança social e dos parceiros sociais.

O Decreto-Lei n.º 64/2007, de 14 de Março, definiu o regime de licenciamento e fiscalização da prestação de serviços e dos estabelecimentos de apoio social, em que são exercidas actividades e serviços do âmbito da segurança social relativos a crianças, jovens, pessoas idosas ou pessoas com deficiência, bem como os destinados à prevenção e reparação das situações de carência, de disfunção e de marginalização social.

Por seu turno, o Decreto-Lei n.º 92/2010, de 26 de Julho, estabeleceu os princípios e as regras necessárias para simplificar o livre acesso e exercício das actividades de serviços realizadas em território nacional, tendo igualmente efectuado a transposição para a ordem jurídica interna da Directiva n.º 2006/123/CE, do Parlamento Europeu e do Conselho, de 12 de Dezembro de 2006, relativa aos serviços no mercado interno.

Tendo em conta que este último diploma é de aplicação inequívoca aos estabelecimentos de apoio social, por força da menção que consta do anexo ao citado decreto-lei, a que se alude no n.º 1 do seu artigo 3.º, torna-se necessário operar a correspondente conformação do Decreto-Lei n.º 64/2007, de 14 de Março.

Neste contexto, e à semelhança do que se verifica noutros sectores governamentais cuja matéria se prende com o âmbito de aplicação da referida directiva, importa proceder à revisão do Decreto-Lei n.º 64/2007, de 14 de Março, por forma a aprofundar princípios de simplificação e agilização do regime de licenciamento, designadamente no que concerne à autorização de utilização, eliminação da forma documental especial e, por fim, a referência expressa ao balcão único electrónico dos serviços para a tramitação desmaterializada dos procedimentos.

Acautelado o interesse público de que se revestem as actividades desenvolvidas pelos estabelecimentos de apoio social, através da indispensabilidade do licenciamento, exigida pelo artigo 34.º da Lei n.º 4/2007, de 16 de Janeiro, que aprovou as bases gerais do sistema de segurança social, a revisão do Decreto-Lei n.º 64/2007, de 14 de Março, vem permitir uma simplificação e agilização de procedimentos, e transpor sectorialmente o regime constante da referida directiva, harmonizando, desta forma, os quadros normativos em vigor.

Para além disso, concretiza-se um dos objectivos consagrados no programa de emergência social (PES) de simplificação e agilização dos regimes de licenciamento das respostas sociais.

Aproveita-se a iniciativa para proceder à actualização das remissões e referências legislativas e das entidades constantes dos artigos 7.º, 8.º, 9.º, 33.º, 44.º e 46.º do Decreto-Lei n.º 64/2007, de 14 de Março, designadamente no que respeita ao regime jurídico da urbanização e edificação (RJUE), constante do Decreto-Lei n.º 555/99, de 16 de Dezembro, alterado pelo Decreto-Lei n.º 177/2001, de 4 de Junho, pelas Leis n.os 15/2002, de 22 de Fevereiro, e 4-A/2003, de 19 de Fevereiro, pelo Decreto-Lei n.º 157/2006, de 8 de Agosto, pela Lei n.º 60/2007, de 4 de Setembro, pelos Decretos-Leis n.os 18/2008, de 29 de Janeiro, 116/2008, de 4 de Julho, e 26/2010, de 30 de Março, e pela Lei n.º 28/2010, de 2 de Setembro, e aos regimes do Decreto-Lei n.º 220/2008, de 12 de Novembro, e da Lei n.º 4/2007, de 16 de Janeiro, adequando-se ainda a redacção dos artigos 5.º, 18.º, 29.º, 33.º e 41.º à orgânica do XIX Governo Constitucional.

Assim:

Nos termos da alínea *c*) do n.º 1 do artigo 198.º da Constituição, o Governo decreta o seguinte:

Artigo 1.º

Alteração ao Decreto-Lei n.º 64/2007, de 14 de Março

Os artigos 5.º, 7.º, 8.º, 9.º, 10.º, 12.º, 16.º, 17.º, 18.º, 29.º, 33.º, 41.º, 44.º e 46.º do Decreto-Lei n.º 64/2007, de 14 de Março, passam a ter a seguinte redacção:

«Artigo 5.º

[...]

As condições técnicas de instalação e funcionamento dos estabelecimentos são as regulamentadas em diplomas específicos e em instrumentos regulamentares aprovados pelo membro do Governo responsável pela área da solidariedade social.

Artigo 7.º

[...]

1 — O licenciamento de construção é requerido à câmara municipal e está sujeito, com as especificidades previstas no presente decreto-lei e nos instrumentos regulamentares respeitantes às condições de instalação dos estabelecimentos, ao regime jurídico da urbanização e edificação (RJUE), estabelecido pelo Decreto-Lei n.º 555/99, de 16 de Dezembro, na redacção actual.

2 — A aprovação do projecto sujeito a licenciamento pela câmara municipal carece dos pareceres favoráveis das entidades competentes, nomeadamente do Instituto da Segurança Social, I. P., da Autoridade Nacional de Protecção Civil e da autoridade de saúde.

3 — A consulta às entidades referidas no número anterior é promovida pelo gestor do procedimento e efectuada através do sistema informático previsto no artigo 8.º-A do RJUE.

4 — O interessado pode solicitar previamente os pareceres das entidades competentes ao abrigo do artigo 13.º-B do RJUE.

Artigo 8.º

[...]

1 — . . . . . .

a) . . . . . .

b) . . . . . .

Figure 21 – Resulting regions of segmentation process

As stated, our approach is destined for known and fixed-structured documents. Therefore, the top-down approach and the geometric region comparison method are considered the most appropriate for this step.

The second step is to extract text from the regions resulting from segmentation. Using the iText library mentioned before, the determination of areas to extract text within is possible.

Note that the layout objects extracted are solely for the purpose of extracting text from the PDF file. The output of this phase is an array of text segments, according to the reading order but without any explicit logical structure.

The following figure illustrates the process of the Document Analysis phase.

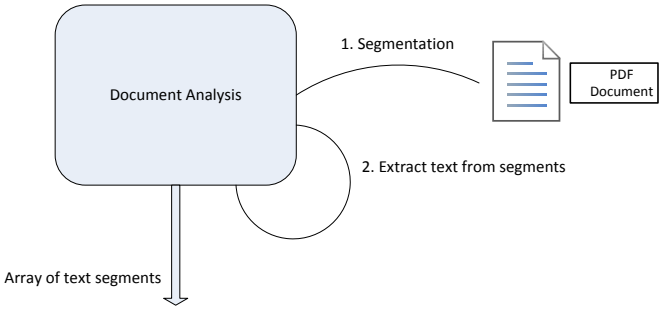


Figure 22 – Document Analysis phase

3.2.3.2 Document Understanding

According to [Todoran et al., 2001], document understanding can be divided into two other phases: the process of grouping the layout document objects in order to classify the logical objects; and the process of determining their logical relations.

In order to complete the first phase the best criteria for grouping the layout objects is similar to the perceptual grouping referred by [Rosenfeld et al., 2002]. Rosenfeld used spatial knowledge to aggregate primitive text objects and create groups of text (line, paragraphs and columns). In our approach the font used in the text segments was used as criteria.

Like the work of [Giuffrida et al., 2000], [Hu et al., 2005] and [Hassan and Baumgartner, 2005] the use of fonts is present in our approach, although the criterion for grouping objects is different. We implemented a similar approach, but defined the criteria as the font itself, as defined by the content stream of a document.

Therefore, we have the objects that are required for grouping according to our criteria. Using the previously presented Figure 20 as example, the operators *BT* and *ET* represent the beginning and the end of the text object. The second line sets the font and the fourth line prints the string.

Based on these findings it is possible to state that by extracting text objects from a PDF file we are able to group strings by font used. The result is then translated into a XML. Note that this result has no guarantee of being in the correct reading order.

The second phase of document understanding is integrated in the third and final phase of document processing, described in the next section. The following figure illustrates the described process of the Document Understanding phase.

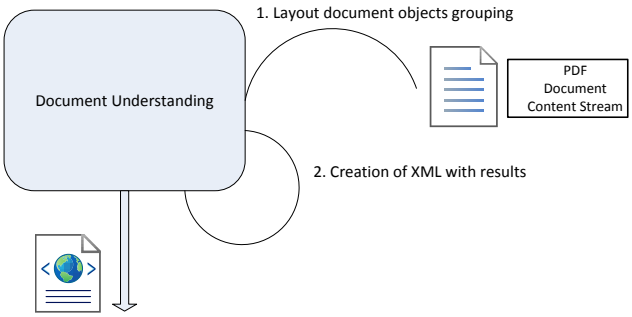


Figure 23 – Document Understanding phase

3.2.3.3 Merging Phase

At this point we have two outputs from previous phases: a complete text description in correct reading order and a XML file with strings tagged and grouped by font used.

Therefore, two processes are missing: joining the two outputs in order to have a XML file that contains the tagged string groups in the correct reading order and, it is necessary to apply the second phase in document understanding, described as the process of determining the logical relations between the groups of objects.

In this approach one logical relation that is dealt from the beginning, as stated above, is the reading order. Other logical relations have to be inputted by the user of the system, such as the structural relationships between segments (e.g., a paragraph contains lines). Our approach is based on two sets of rules: structural and logical rules. Structural rules are mainly applied in order to classify and create new groups of XML tags or to re-label existing ones; syntactical rules are used. Logical rules are applied in order to establish logical relations between groups. Both structural rules and logical rules have their own specific syntax. In the following section this will be explained in further detail.

The expected output of our approach is a XML file containing the text description of the PDF file, in correct reading order, tagged accordingly and containing logical relations set out by the user.

The following figure illustrates the procedure of the Merging phase.

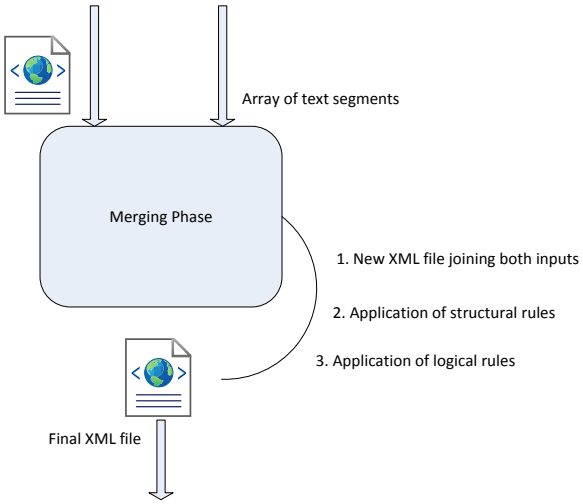


Figure 24 – Merging phase

3.2.4 Implementation

The implementation is divided in two phases: the extraction phase and the analysis phase. Despite the former presentation of phases, the implementation of the approach does not follow the presented order.

The extraction phase contains three processes: extraction of information from the PDF's content stream, extraction of text using geometric positioning and merging the output of the two previous processes into a XML file. The analysis phase contains two processes: application of structural rules and application of logical rules. The system output is a XML file that contains the mapping of the layout structure to a logical structure of the PDF document.

The extraction of text within tables and the extraction of images were not implemented. Nonetheless, the text within tables is analysed and entity recognition is executed.

The coarse-grained description of the implementation is illustrated in the following image.

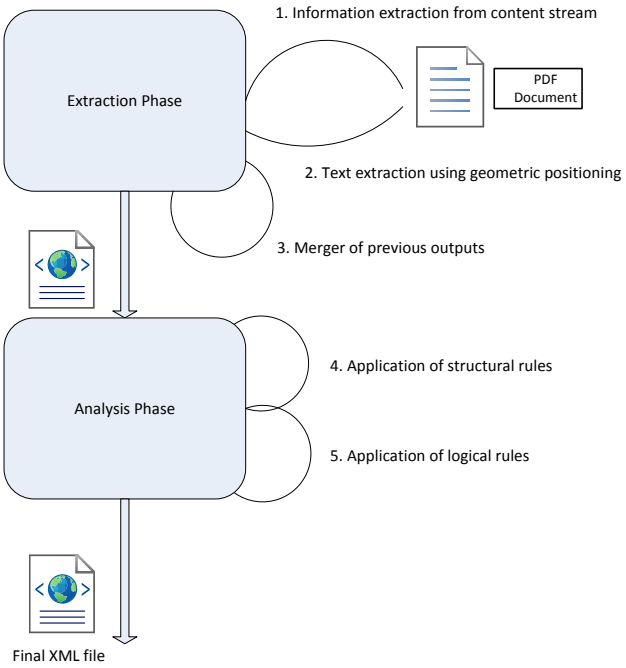


Figure 25 – Coarse-grain description of implementation

3.2.4.1 Extraction from content stream

As explained before, a PDF document is composed by objects. Regarding this section the reference is specifically to text object.

The objective of this process is to extract strings labelled with the font resource declared for its use. This is done by parsing sequentially the content stream extracting each text object and parsing its font and string. Sequential strings that have the same font are grouped. As the results are obtained, they are appended in a XML structure. After this, two procedures are called: one to extract explicit entities and another to clean the XML.

In the first procedure, as explained before, we use a single criterion of font used. By analysing the fonts used in the Portuguese Republic's Diary, we found that the italic style is most often used to refer to an entity. Therefore, this process consists on the extraction of these explicit entities and its relabeling.

In the second procedure cleaning operations are made e.g. cleaning empty tags, joining two consecutive objects with the same tag. Also, in this procedure tables are removed. However, before this operation, a regular expression for entity recognition is applied in the text within, in order to extract the entities present in the documents tables.

The following figure illustrates this process.

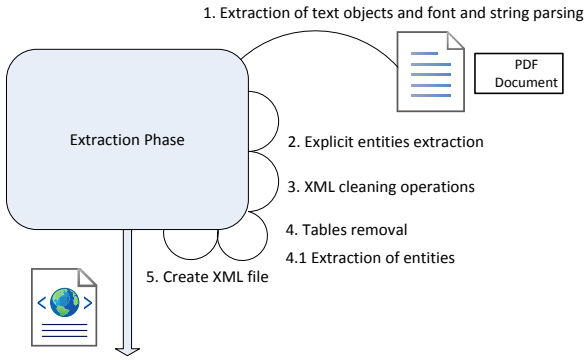
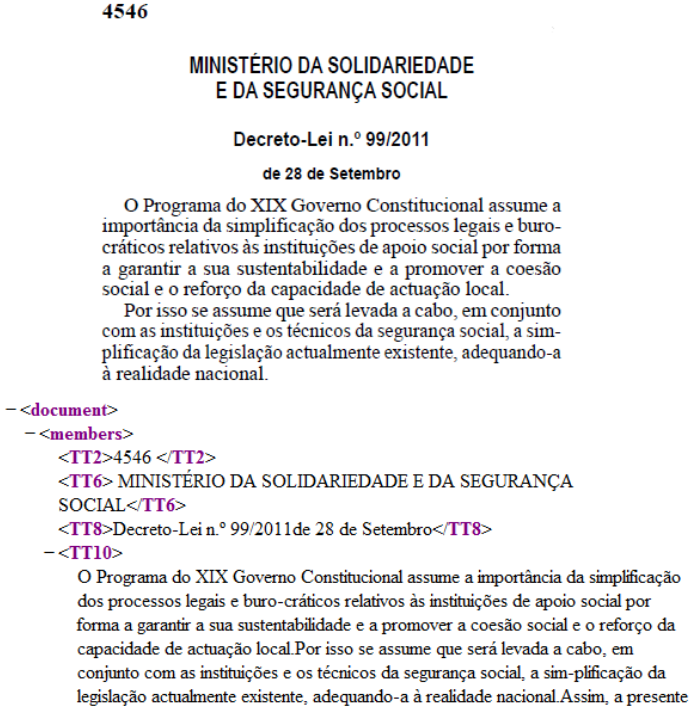


Figure 26 – Extraction from content stream process

In the following figure an excerpt of the auxiliary XML file created to store this information and the respective PDF document is presented.



4546

MINISTÉRIO DA SOLIDARIEDADE
E DA SEGURANÇA SOCIAL

Decreto-Lei n.º 99/2011

de 28 de Setembro

O Programa do XIX Governo Constitucional assume a importância da simplificação dos processos legais e burocráticos relativos às instituições de apoio social por forma a garantir a sua sustentabilidade e a promover a coesão social e o reforço da capacidade de actuação local.

Por isso se assume que será levada a cabo, em conjunto com as instituições e os técnicos da segurança social, a simplificação da legislação actualmente existente, adequando-a à realidade nacional.

```
−<document>
  −<members>
     <TT2>4546 </TT2>
     <TT6> MINISTÉRIO DA SOLIDARIEDADE E DA SEGURANÇA
     SOCIAL</TT6>
     <TT8>Decreto-Lei n.º 99/2011de 28 de Setembro</TT8>
    −<TT10>
       O Programa do XIX Governo Constitucional assume a importância da simplificação
       dos processos legais e buro-cráticos relativos às instituições de apoio social por
       forma a garantir a sua sustentabilidade e a promover a coesão social e o reforço da
       capacidade de actuação local.Por isso se assume que será levada a cabo, em
       conjunto com as instituições e os técnicos da segurança social, a sim-plificação da
       legislação actualmente existente, adequando-a à realidade nacional.Assim, a presente
```

Figure 27 – Excerpt of the auxiliary XML

3.2.4.2 Extraction from Layout

In this process the text from the PDF document is extracted using region filters present in the iText library. We therefore extract text from a known location. The documents we refer are organized in double columns; therefore, we extract text by setting a vertical ruling in the middle of the page as shown in Figure 21.

The output of this operation consists on arrays of strings which are joined in order to produce a unique array. This array represents a sequential list, in accordance with the correct reading order of the text. Each string of the array contains a line of a column.

The sole purpose of this process is to extract text in the correct reading order.

3.2.4.3 XML Output

This is the final process after both extraction processes. It consists of sequential comparisons between the previous extractions.

For each line obtained from the Extraction from Layout, a lookup in the auxiliary XML of the Extraction from content stream process is made. The resulting matches are appended into the final XML. When resulting matches are inconsistent (more than one result) the algorithm provides various options in order to obtain a consistent result. If the lookup does not provide any match, the algorithm operates text splits in the respective line in a final attempt to provide a consistent response. The algorithm responsible for this procedure (see Attachment 2) is the result of accumulated experiences and knowledge regarding the organization of text in these documents and of an intensive period of tests.

In the following figure a bit from the auxiliary XML (not in correct reading order) is presented, where as in Figure 29 a bit from the XML output is presented. It is possible to denote that the numbers in the TT8 tag are not sequential. We do not detain the necessary information to specify why the iText library is unable to parse the text objects from the content stream in correct reading-order. However, we assume this could be either due to the content stream not having all of its text objects in a sequential manner or due to the use of misleading character recognition due to the use of vectors in that process.

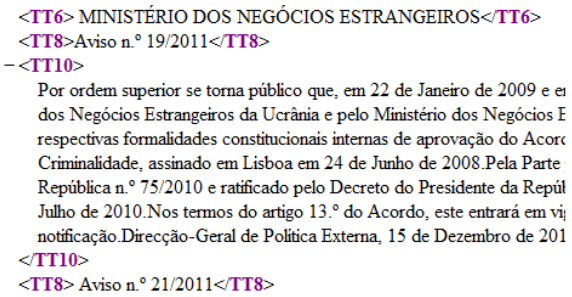


Figure 28 – Bit of auxiliary XML

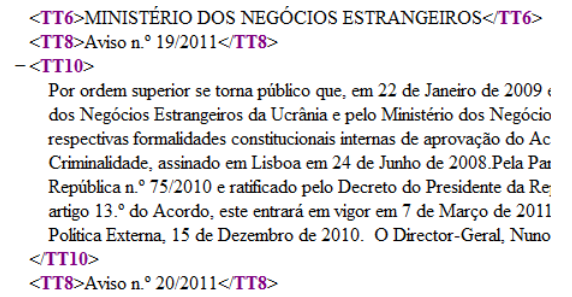


Figure 29 – Bit of XML output

3.2.4.4 Application of Structural and Logical Rules

These two processes (application of structural rules and application of logical rules) compose the analysis phase.

Although these processes are separated, they are implemented using the same paradigm. It consists on a rule based system that is applied according to the syntax defined for each set of rules (structural and logical).

In order to apply these rules it is necessary a user input. This input is done by the declaration of rules in four text files containing the respective structural and logical rules. The system embeds operations that enable the application of these rules. The operations are the result of the knowledge acquired from the analysis of the XML output file – the output of the previous phase.

Regarding the application of structural rules, these obey pre-defined types of operations. The structural rules are defined in two separate files. The first file contains rules relating to operations that include recognition of structure entities (articles, lines, chapters, sections and others), deletion of structure entities and recognition of entities (see Table 4 for shortlist). The second file contains rules to alter original XML tag names to tag names with meaning (see Table 5 for shortlist).

The syntax for the specification of rules in the first file is as follows: '`RegExp:::operation`'. The operation bit represents an internal process encoded in our system. As mentioned, the domain knowledge is based uniquely in the Portuguese Republic's Diary documents. Some examples of these internal processes are insertion after or before the present tag and recognition of structural elements within text objects.

In the following figure an XML output is presented without the application of any rule.

<TT2>Regulamentação específica</TT2>
−<TT10>
As condições técnicas de instalação e funcionamento dos estabelecimentos são as regulamentadas em diplomas específicos e em instrumentos regulamentares aprovados pelo membro do Governo responsável pela área da soli-dariedade social.CAPÍTULO II

Figure 30 – XML output without structural rules

In the following figure the same case is presented with the application of an example rule from the first structural rules file, related to the recognition of chapter elements: '`CAPÍTULO\s[IVXLCM]+):::chapter`'.

<TT2>Regulamentação específica</TT2>
−<TT10>
As condições técnicas de instalação e funcionamento dos estabelecimentos são as regulamentadas em diplomas específicos e em instrumentos regulamentares aprovados pelo membro do Governo responsável pela área da soli-dariedade social.
</TT10>
<chapter>CAPÍTULO II</chapter>

Figure 31 – XML output with structural rules (First file)

As an important remark, the extraction of entities is processed at this point with the application of a structural rule. The successful results obtained have no implication in the structure of the text or document; they are stored separately.

Table 4 – Structural rules (First file)

Rule	**Description**
[\.\s]{4,}+ ::: (Actual corpo);	Series of four or more ". " strings are transformed into string "(Actual corpo);"
((?:\W\s)?(Artigo \d+\W{2})(\W[A-Z])$) :::article	Calls the internal process "article" to handle references to articles
((?:\W\s)?(Artigo \d+\W{2}))(.+$) :::articleBetweenText	Calls the internal process "articleBetweenText" to handle references to articles between two different text objects
(\d+\.+)*\d\s{1}[\W&&[^%\+]]\s{1} :::paragraph	Calls the internal process "paragraph" to handle references of paragraphs
(?<=(\;|\:))([a-z]\)\s).+?(?:\;|\.$\|[a-z\)]$|$) :::line	Calls the internal process "line" to handle references of line
^Correio electrónico: dre@incm.pt.+$ :::delete	Calls the internal process "delete" to erase all references from the XML output

The second file of structural rules consists on a list of rules that deal solely with altering the initial XML tag names into the user specified desired tag names.

The syntax for the specification of rules in the second file is as follows:
'`RegExp:::previoustag:::newtag`'. The rule may or may not contain regular expressions.

Table 5 – Structural rules (Second file)

Rule	**Previous XML tag**	**New XML tag**
Republica.+$:::TT8:::LexDocumentRepublish	TT8	LexDocumentRepublish
^\s?[A-ZÁ-Ü]{2}.+$:::TT6:::LexEntity	TT6	LexEntity
.+:::TT20:::LexSubEntity	TT20	LexSubEntity
.+:::TT12:::LexSubEntity	TT12	LexSubEntity
.+\d/.*:::TT8:::LexDocument	TT8	LexDocument
.+:::TT8:::Title	TT8	Title

In Figure 32 an example rule from the second file is applied to the bit previously presented in the following figure where tag *<TT6>* is replaced by tag *<govEntity>*: `'^\s?[A-ZÁ-Ü]{2}.+$:::TT6:::govEntity'`.

<govEntity>MINISTÉRIO DOS NEGÓCIOS ESTRANGEIROS</govEntity>
<TT8>Aviso n.º 19/2011</TT8>
−<TT10>
Por ordem superior se torna público que, em 22 de Janeiro de 2009 e em 8 de Setembro de 2010, foram rece-bidas notas, respectivamente pelo Ministério dos Negócios Estrangeiros da Ucrânia e pelo Ministério dos Negócios Estrangeiros da República Portuguesa, em que se comu-nica terem sido cumpridas as respectivas formalidades constitucionais internas de aprovação do Acordo entre a República Portuguesa e a Ucrânia no Domínio do Combate à Criminalidade, assinado em Lisboa em 24 de Junho de 2008.Pela Parte portuguesa, o presente Acordo foi aprovado pela Resolução da Assembleia da República n.º 75/2010 e ratificado pelo Decreto do Presidente da República n.º 77/2010, publicados no n.º 141, de 22 de

Figure 32 – XML output with tag structure rules (Second file)

Regarding the application of logical rules, the rules are defined in two separate files as well.

The logical rules intend to structure the final XML file in order to replicate the information hierarchy present in the original PDF document. This replication takes into consideration the user decisions concerning the desired granularity. This process requires a previous user analysis in order to specify the correct options. For our example, in terms of information hierarchy we find that the Legislation Entity is the most important element in the Portuguese Republic's Diary; each Legislation Entity may or may not have a Sub-Entity; these Entities issue Legislation Documents; a Legislation Document may or may not have a Description; a Legislation Document may or may not be organized by Articles, etc. This information is detailed in Section 3.1.

In order to reproduce that hierarchy two types of processing are required: a first process where a specific tag appends all the following objects until a similar tag is found; a second process that appends the objects of a specific tag onto another preceding it.

The first logical rules file represents the rules applied for the first process (see Table 6 for list); the second file contains the rules that are applied in order to perform the second process (see Table 7 for list).

The first logical rules file represents a top-down approach of aggregation. It appends every tag onto a specific user defined tag. The syntax for these rules is as follows: '`firstTag:::aggregationTag`'. The *firstTag* field represents the parent tag, and the *aggregationTag* represents the tag to which the following will be appended. This process is used primarily with the objects that have higher importance in the structure or contains most of the text (for example Legislation Entities and Legislative Documents).

Table 6 – Logical rules (first file)

Rule	Description
members:::LexEntity	Every object under an object with “LexEntity” tag is appended onto it. The objects with “LexEntity” tag are appended onto the object with tag “members”
LexEntity:::LexDocument	Every object under an object with “LexDocument” tag is appended onto it. The objects with “LexDocument” tag are appended onto the precedent object with tag “LexEntity”
LexEntity:::LexSubEntity	Every object under each object with “LexSubEntity” tag is appended onto it. The objects with “LexSubEntity” tag are appended onto the precedent object with tag “LexEntity”

In the following figure a bit of a XML output is presented with the application of an example rule ‘`LexEntity:::LexDocument`’, from the first logical rules file. In the figure it is possible to observe the application of a rule that follows what was stated before concerning information hierarchy.

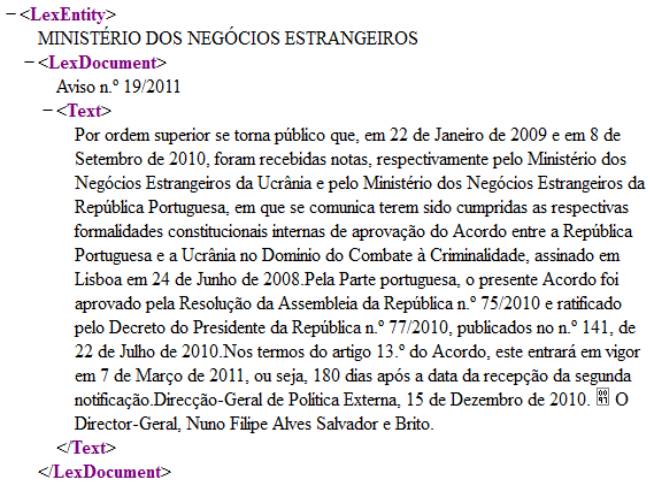


Figure 33 – XML output with logical rules (Example 1)

The second logical rules file contains rules that have the objective of appending objects with a specific tag onto another user defined tag. The syntax for these rules is as follows: ‘`parentTag:::tagToAppend`’. The *parentTag* field represents the tag onto which the objects will be appended; the second field represents the tag to be appended.

Table 7 – Logical rules (second file)

Rule	Description
line:::paragraph	Every object with a tag “line” is appended onto the preceding object with tag “paragraph”

In the following figure a bit of a XML output is presented with the application of the rule ‘`line:::paragraph`’, from the second logical rules file.

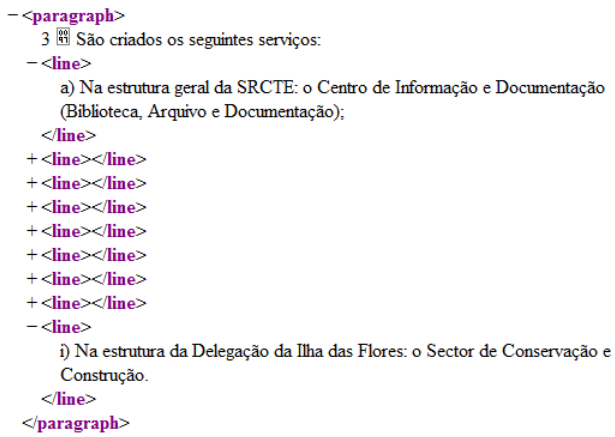


Figure 34 – XML output with logical rules (Example 2)

After the application of both structural and logical rules a procedure is called to examine the resulting XML output and to operate a cleaning operation (empty tags).

This is the final process of the analysis phase. The output of this phase is the final XML file that contains the mapping of the layout structure to logical structure of a PDF document.

3.2.5 Background

In order to be clear about the influence of the studied approaches, Table 8 represents the mapping of our processes with what we consider to be correspondent to both following descriptions.

Niyogi [Niyogi, 1994] presents a description of a computational model for extracting the logical structure of a document, described as follows:

1. a procedure for classifying all the distinct blocks in an image;
2. a procedure for grouping these blocks into logical units;
3. a procedure for determining the read-order of the text blocks within each logical unit;
4. a control mechanism that monitors the above processes and creates the logical representation of the document;
5. a knowledge base containing knowledge about document layout and structure and;
6. a global data structure that maintains the domain and controls data.

Taylor et al. [Taylor et al., 1994] presents four phases in his implementation:

1. Physical Analysis
2. Logical Analysis
3. Functional Analysis
4. Topical Analysis

Table 8 – Mapping of implemented processes with previous research

Processes	**[Niyogi, 1994]**	**[Taylor et al., 1994]**
Extraction from Content Stream	1) and 2)	1)
Extraction from Layout	3)	1)
XML Output	3)	2)
Application of Structural Rules	4)	3)
Application of Logical Rules	4)	4)

The former table presents a general idea of the correspondence of processes present in our approach and previous research.

Giuffrida et al. [Giuffrida et al., 2000] used spatial knowledge of a given domain knowledge to encode a rule-based system for automatically extracting metadata from research papers; they used spatial knowledge to create a rule; the metadata was extracted from PostScript files and formatting information was used.

Hu et al. [Hu et al., 2005] proposed a machine learning approach to title extraction from general documents; tests were made with Word and PowerPoint documents. This method mainly utilizes formatting information such as font size in the models.

Both approaches use formatting information, such as the font used. We use the font as declared in the content stream of PDF documents as criteria for perceptual grouping.

We assumed this option due to the often presence of different styles within text segments of the same font size. Usually this represents an entity; therefore, using the content stream font description as criteria instead of the font size, enables a better information extraction process.

4 Knowledge Organization – Phase 2

This phase comprehends the development of artefacts that enable the storage of information extracted in the previous phase. Namely, this phase comprehends the study and development of an adequate ontology concerning the legal domain and the objectives of the implemented system, a database capable of storing such knowledge and the required connection between the system and the database to enable the storage and querying of the data. This phase also comprehends the indexing process of the legislation text present in the documents.

Therefore, this phase is divided into three parts. The first part contains the study of previously developed ontologies in the legal domain, the desired characteristics of an ontology definition and the presentation of the developed ontology. The second part describes the construction and implementation of the database that enables the storage of information according to the developed ontology and the description of process of communication between the system and the database. The third and final part presents the development of the indexing process.

The first part presents the results of the research concerning ontologies in the legal domain. During this research, it was found that a standard for this domain is not yet developed, and therefore it is necessary to study the already existing ontologies. Results are presented and discussed.

The decision of developing a new ontology is taken. The main idea behind this new ontology is allowing an easier use of the ontology for other users, ensuring the interoperability with the existing ontologies and guaranteeing its share ability. Furthermore it addresses the description of the relation between legislation documents and entities referenced in their text.

The second part presents the details concerning the database which will store information according to the ontology. The third part presents the details concerning the indexing process.

4.1 Ontology

The legal domain can contain a various number of types of documents. Nonetheless, this section will focus on legislation.

Many ontologies have already been developed in order to tackle various issues, such as knowledge organization in the legal domain, standardization of markup languages of legislative texts or interoperability with other ontologies. In the course of this project, the discussion will be focused on legislation, and the issue of how to organize knowledge in the legal domain.

Some ontologies in this domain, through its dissemination, became somewhat a reference in terms of share ability and standardization. Nonetheless, there is still no standard concerning ontologies in the legal domain. However, some ontologies present the partial standardizations, such as MetaLex [Boer et al., 2007]. Others are used internationally, such as Akoma Ntoso [Vitali and Zeni, 2007] and

therefore need to be studied and evaluated. This will be done in Section 4.1.1. However, before this discussion is presented, it is necessary to define what are the established requirements of an ontology in the scope of this project.

The information available from the extraction phase presented earlier does not permit the same degree of results that experiences and cases of countries that by their own governments' initiative developed e-legislation or open legislation projects in order to provide this information directly from their source.

The extraction phase concentrates specifically in the extraction of documents, information regarding these documents that enable its characterization and, finally, the recognition and extraction of entities. Therefore, in abstract terms, the extraction phase focuses on two concepts: Documents and Entities. These concepts will be described in detail further.

However, despite the narrowing of the conceptual scope, a central focus point of the ontology required for this project is interoperability. Therefore, notwithstanding the small size required for the ontology, the interoperable capability of the used ontology is non-dispensable. This is a requirement that opens the definition of a third characteristic required: share ability. Related to the previous considerations, the ontology is required to be country, language and jurisdiction-independent, and thus able to be shared with less restrictions.

Thus, the intended ontology is defined as having at least three characteristics: small, interoperable and shareable.

4.1.1 Background

This section provides the background concerning ontologies and its classification, but also the study of different ontologies in the legal domain and its fitting regarding the intended ontology set forth in the former section.

In 1995, Guarino and Giaretta [Guarino and Giaretta, 1995] in a discussion concerning the definition of ontology and Thomas R. Gruber [Gruber, 1993] [Gruber, 1995] proposal, defined ontology as "a logical theory which gives an explicit, partial account of a *conceptualization*", and conceptualization as being "an intensional semantic structure which encodes the implicit rules constraining the structure of a piece of reality".

Later, in 1998, Guarino [Guarino, 1998] defined ontology, refining Gruber's definition by making clear the difference between ontology and conceptualization. This is the definition used in this dissertation. The definition is as such:

"An ontology is a logical theory accounting for the *intended meaning* of a formal vocabulary, i.e. its *ontological commitment* to a particular *conceptualization* of the world. The intended models of a logical language using such a vocabulary are constrained by its ontological commitment. An ontology indirectly reflects this commitment (and the underlying conceptualization) by approximating these intended models".

In the same reference, Guarino presents two types of ontologies "according to their *accuracy* in characterizing the conceptualization they commit to": fine and coarse-grained ontologies. Fine-grained ontologies are more specific regarding the intended meaning of a vocabulary but due to number of axioms and the expressiveness of the language adopted "may be hard to develop and to reason on"; a coarse-grained ontology "may consist of a minimal set of axioms written in a language of minimal expressivity, to support only a limited set of specific services, intended to be shared among users which *already agree* on the underlying conceptualization". Guarino distinguishes both types by referring to them as detailed *reference ontologies* (fine-grained) and coarse *shareable ontologies* (coarse-grained). Regarding these definitions, the ontology required for the developed system complies with the definition of coarse *shareable ontologies* (coarse-grained ontology).

Still concerning the types of ontologies, Guarino develops the categorization of ontologies according to their level of dependence on a particular task or point of view [Guarino, 1997] and specifies four kinds of ontologies: top-level ontologies, domain and task ontologies and application ontologies.

Top-level ontologies should "describe very general concepts", "which are independent of a particular problem or domain"; a domain or task ontology describes, respectively, "the vocabulary related to a generic domain or a generic task or activity by specializing the terms introduced in the top-level ontology"; and an application ontology describes "concepts depending both on a particular domain and task, which are often specializations of *both* the related ontologies". Regarding the former definition, Guarino adds: "These concepts often correspond to *roles* played by domain entities while performing a certain activity, like *replaceable unit* or *spare component*". Regarding these definitions, the ontology suggested for the developed system complies with definition of domain or task ontology. The designation used for the purposes of this dissertation is domain ontology.

Giovanni Sartor [Sartor, 2007] defines legal semantic web as "legal (legislative) information partly machine understandable, automatically processable according to its legal meaning". He lists a few opportunities for legislation regarding legal semantic web:

- Maintenance of legal sources
- Improvement of legal drafting
- Legislation based upon knowledge and dialogue
- Publicity of procedures and information
- Dialogue between sub-national, national, and international institutions

Sartor also states that compliance with shared of standards is a "precondition for this opportunity to be realised". The existence of standards is of the most importance regarding the Semantic Web. This is stressed by Biasiotti [Biasiotti et al., 2008] when enumerating trends in the provision of legal information states that "the available information has to be coded and decoded according to shared machine-readable standards or protocols". Biasiotti continues by stating that "the determination of machine-processable standards is a crucial issue for public policy in the information society" and that "the shared adoption of appropriate open standard greatly facilitates technological progress, cooperation and competition in the framework of the knowledge society".

Concerning the available options regarding ontologies in the legal domain, this research is based on the study of Barabucci [Barabucci et al., 2012]. According to the authors, legal ontologies are usually divided in two categories: core and domain-specific. Core ontologies describe the “common conceptual denominator of the field” [Despress and Szulman, 2007] and “provide definitions of general legal concepts”, while domain-specific ontologies “cover more context-oriented concepts and similar abstract ideas peculiar to certain field”. Several ontologies are presented and classified according to the orthogonal classification of legal ontologies proposed by the authors. According to the authors this orthogonal classification is “centered around their relation to legal sources and external entities”. Therefore, besides the classification of ontologies according to the previously stated categories, core and domain-specific but also top/upper ontologies, the authors divide them in three groups: document-centric, content-centric and integration-centric.

According to Barabucci et al., document-centric ontologies are defined as those “whose main goal is to describe the documental part of the legal documents” and are “used to model the evolution of legal sources”; content-centric ontologies “primarily define legal concepts carried by legal sources”; and integration-centric ontologies are those that “give much relevance to the integration of legal sources and concepts to external entities, that exist outside the legal domain and are independent of it”. The results of the study carried out by Barabucci et al. are presented in the following table.

Table 9 – Classification of existing ontologies [Barabucci et al., 2012]

Ontology	**Top**	**Core**	**Domain**	**Document**	**Content**	**Integration**
Ontology of Greek Public Administration [Savvas and Bassiliades, 2009]			X	X		X
Legal Case Ontology [Shen et al., 2008]			X	X	X	X
CLIME [Winkels et al., 2002]			X		X	
OntoPrivacy [Cappelli et al., 2007]			X		X	
ALIS IP [Laukyte et al., 2008]			X		X	
OCL.NL [Breuker et al., 2002]			X		X	
LTS [Ajani et al., 2007]			X			X
PARL [PARL, 2012]			X			X
CGOV [CGOV, 2012]			X			X
OCD [OCD, 2011]			X			X
Legislation.gov.uk Ontology [UK Legislation, 2012]		X		X		
MetaLex [Boer et al., 2010]		X		X		
Core Legal Ontology [Gangemi et al., 2003]		X			X	
FOLaw [Breuker and Hoekstra, 2004]		X			X	
LKIF-Core [Hoekstra, 2009]		X			X	
LRI-Core [Breuker and Hoekstra, 2004]		X			X	
Ontology of Fundamental Legal Concepts [Rubino et al., 2006]		X			X	
Jur-IWN [Sagri and Tiscornia, 2003]		X				X
Micro-ontologies [Despress and Szulman, 2007]		X				X
FRBR [Madison, 2000]	X		X	X		
DOLCE [Masolo et al., 2004]	X				X	
SUMO [Sowa, 2000]	X				X	
SKOS [Bechhofer and Miles, 2009]	X					X
FOAF [Brickley and Miller, 2010]	X					X

As stated before, the objective of the organization of knowledge extracted from the Republic's Diary documents is not to portrait the implication of legal concepts in legal texts, but to describe the legal documents bibliographically (organization of the documents) including the several versions of a given entry, and to detect external entities (entity extraction) in order to create the ability to relate the legal documents with the external entities referred in its text. Therefore, the ontology required for the necessary task in this system should be a domain and document and integration-centric ontology.

Analysing Table 9 we conclude that the only ontology with these characteristics (document and integration-centric) is the Greek Public Administration Ontology [Savvas and Bassiliades, 2009]. This ontology is defined as an ontology to represent the public administration and the document flow

amongst them. The authors state that the differentiation concerning previous proposals is that it refers specifically to “administrative entities, procedures and documents rather than legal hierarchies”. Therefore, the fact that classifies the Greek Public Administration Ontology as being integration-centric is its ability to model the Greek Public Administration and the people involved in the various processes.

Considering the previous requirements of the ontology solution and its purposes, the Greek Public Administration Ontology appears to be inadequate. According to the orthogonal classification of Barabucci et al., this solution would comply with the objectives of this project, but the focus of this solution does not capture the conceptualization required, namely, the concept of Entity. In this solution the concept of Entity is focused on the organizations and specific persons with high responsibility in the Public Administration. The desired ontology must have the capability of organizing knowledge referring to heterogeneous external entities.

Although none of the options in Barabucci evaluation of a group besides the Greek Public Administration Ontology presents the double classification expected, the study of some of the ontologies in the presented list appear to have strong contributions for the desired ontology of the developed system, namely, the Legislation.gov.uk ontology which uses MetaLex.

The Legislation.gov.uk developed a small ontology that uses vocabularies from two other ontologies: MetaLex [Boer et al., 2010] and FRBR [Madison, 2000].

MetaLex is a generic and extensible framework for the XML encoding of the structure of, and meta-data about, legal sources. The RDF ontology of MetaLex classifies bibliographic entities (work, expression, manifestation and item level, and content models), types of reference between bibliographic entities, events, and others. This classification is made through the use of the FRBR ontology. Since 2002, the first version [Boer et al., 2002] of MetaLex was redesigned and took into consideration the experiences of Norme in Rete [Biagioli et al., 2003] and Akoma Ntoso [Vitali and Zeni, 2007] as referred by Boer et al. [Boer et al., 2007]. MetaLex was submitted as a standard proposal and a partial CEN Workshop Agreement now exists.

The IFLA’s Functional Requirements for Bibliographic Records (FRBR) is a general model for describing the evolution of any document. It works for both physical and digital resources, and it is not tied up with a particular metadata schema. This ontology distinguishes the concepts described in MetaLex. The concepts are work, a distinguishable intellectual or artistic creation; expression, the intellectual or artistic form that a work takes each time it is realized; manifestation, a physical embodiment of an expression of a work; and item, a single exemplar of a manifestation of an expression.

This presents a powerful combination that enables an ontology to classify documents in bibliographic terms in a very high level of assertion. The Legislation.gov.uk project ontology is developed on this distinction basis: legislation is defined as work; different versions of the legislation are expressions; different publishing formats for that expression are manifestations; while specific copies of those files are items.

Furthermore, the Legislation.gov.uk ontology uses Dublin Core Metadata Initiative Metadata Terms [DCMI, 2012] to create pointers between versions (Expression in FRBR) of a given legislation (Work in FRBR). This is the approach used to enable versioning of legislations: dated versions may have a `dct:replaces` pointer to the previous version of the item and a `dct:isReplacedBy` pointer to the next version of the item.

The Dublin Core Metadata Initiative (DCMI) [Powell et al., 2005] is a standardized[20] provider of core metadata vocabularies in support of interoperable solutions for discovering and managing resources.

Also, through the use of FOAF [Brickley and Miller, 2010] a link between the legislation item (Work in FRBR) and the documents that are particular versions of legislation can be established through the use of properties `foaf:isPrimaryTopic`, and `foaf:primaryTopic` in the opposite situation.

FOAF (Friend of a Friend) specifies a language capable of modelling and linking people and information using the Web. It integrates three kinds of network: social networks, representational networks and information networks.

Through the use of vastly shared and used ontologies such as FOAF and Dublin Core, but also through the refinement of FRBR levels in this sort of documents using the FRBR ontology, the Legislation.gov.uk ontology proves to be simple but much effective in the task of creating a bibliographic record of legislation and maintaining the coherence of these records concerning its evolution (e.g. drafting new versions of a given legislation). The use of MetaLex ontology, provides a basis for interoperability with other knowledge bases in the legal domain. To the extent of this research, MetaLex is currently used in the UK Legislation and in the Dutch Legislation.

The Legislation.gov.uk Ontology can be described as being small and interoperable ontology, which makes an excellent use of well known, shared and used ontologies such as FRBR, FOAF and Dublin Core. Through its use of FRBR levels it enables the possibility of versioning of legislation. The use of MetaLex, a partial standard at this point, provides a basis for easy interoperability with other legislation silos.

However, this ontology doesn't provide an answer to one of the desired ontology requirements: the ability to relate the legal documents with the external entities referred in its text. Actually, this could've been achieved through the use of the available vocabulary in the FOAF ontology. It provides concepts such as Person, Group and Organization, that could guarantee a mapping of references between the documents and external entities.

Due to the non-existence, to the extent of this research, of an ontology capable of complying entirely with the requirements stated for the desired ontology, a new ontology is proposed.

[20] The Dublin Core Metadata Element Set has been published as IETF RFC 5013, ANSI/NISO Standard Z39.85-2007 and ISO Standard 15836:2009

4.1.2 SL Ontology – A Simple Ontology for Legislation

In this section, the SL Ontology (A Simple Ontology for Legislation) is presented. The previous section showed that the studied ontologies in the legal domain lack, in a general sense, the focus on modeling the relationship between documents and external entities that are mentioned in the legislation text. However, both of the presented ontologies in the previous section give an excellent idea of the basis necessary for a new ontology that should provide the vocabulary needed to ensure this idea.

As presented before, the main characteristics defined for the ontology necessary for the organization of knowledge acquired by the system are: being small, interoperable and shareable.

Through the previous study it is possible to assess that the ontologies presented show a shared underlying conceptualization, at least from an abstract point of view. According to this, the small characteristic referred points to the necessity of a coarse-grained ontology. The SL Ontology defines a minimal set of core services and axioms required to organize ad describe the knowledge extracted from the Republic's Diary documents.

The Legislation.gov.uk shows an excellent example of the modelling of legislation and the versioning of these documents, as stated before. However, in terms of interoperability, it has made the choice of using the MetaLex vocabulary. The required interoperable characteristic is not entirely fulfilled with the option. The SL Ontology does not discard the interoperability with the knowledge bases that are organized according to MetaLex, but acknowledges the existence of other important vocabularies shared in the legal domain, namely, Pan-African Akoma Ntoso, that should be covered.

Akoma Ntoso (Architecture for Knowledge-Oriented Management of African Normative Texts using Open Standards and Ontologies) [Vitali and Zeni, 2007] is an operating framework and set of guidelines for driving e-Parliament services in a Pan-African context by formalizing and harmonizing the storage, publication and exchange of Parliamentary documents using a precise, common and easy to understand data format based on XML.

The share ability of the SL Ontology is centred in the independence that the vocabulary and the organization of knowledge has from a group of factors. The SL Ontology is country, language and jurisdiction-independent. This is achieved through a coarse-grained approach in the development of this ontology. Through the condensation of the requirements in terms of size of vocabulary and interoperability, it is possible to achieve a higher independence regarding the factors mentioned before.

The SL Ontology is organized in layers: the core and implementation layers. The following figure illustrates the organization of the ontology.

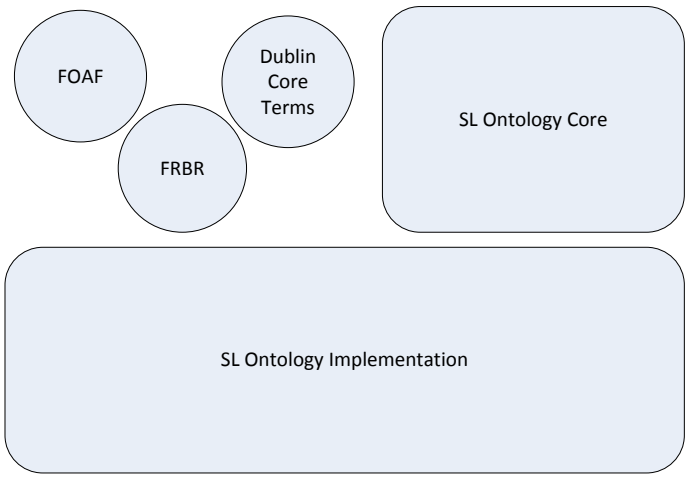


Figure 35 – SL Ontology layer organization

The core layer is composed by OWL classes that represent Top Level Classes (TLC) of the ontology. The following table presents the list of TLC.

Table 10 – SL Ontology TLC

Top Level Class	**Description**
Document	Represents any document and is linked to the FRBR group-1 abstraction levels (Work)
Entity	A human being or a formal or informal group of human beings
File	Represents a version of a given document and is linked to the FRBR group-1 abstraction levels (Expression)

Regarding the SL Ontology development, these three OWL classes represent the minimal requirements for the knowledge organization of a bibliographic record of legislation documents and the description of the connection between the documents and external entities.

The implementation layer represents the main element of the SL Ontology. The Implementation layer contains the specification of the Ontology as it will be used by the users.

This layer has two tasks: 1) enable the knowledge organized according to the SL Ontology to be interoperable with the Semantic Web and other datasets of Linked Data, and 2) to give an accurate definition to the TLC present in the core layer.

In general, this layer has the objective of mapping the TLC with the classes from other ontologies that will be used and have been discussed in the previous section. For example, the TLC Entity will be linked to the FOAF Agent class and the Document and File TLC will be linked to the corresponding

FRBR classes, Work and Expression, respectively. In addition, two other OWL classes are implemented: Legislation and Container. These classes are a specification of the TLC Document. They represent two types of documents that are alone or together in a legislative system. In the Portuguese case, the Republic's Diary is the container of published legislation.

The illustration of the SL Ontology Implementation layer is depicted in the following figure.

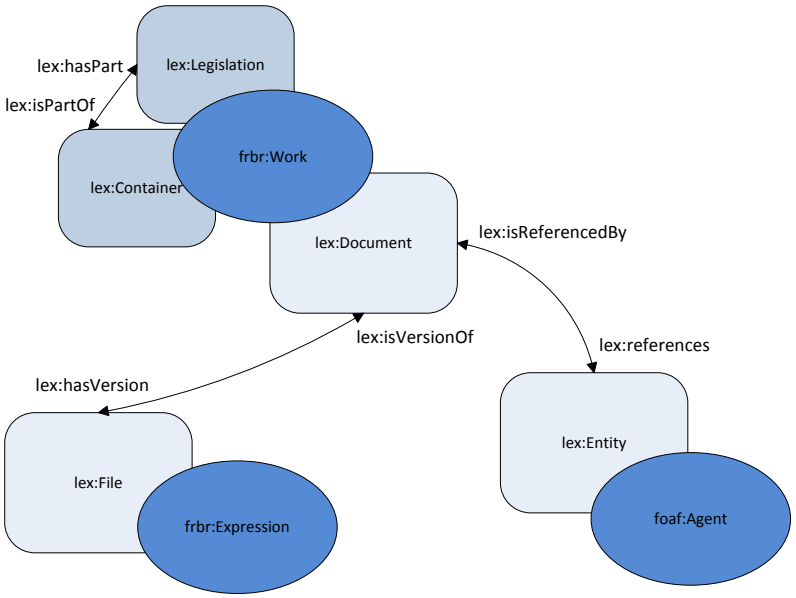


Figure 36 – SL Ontology Implementation layer

From the previous picture it is possible to observe that, besides the organization of the OWL classes in the SL Ontology, the connection between these classes is made through the use of object properties. Each of the object properties has an equivalent with an equal vocabulary as the Dublin Core Metadata Initiative Metadata Terms.

The description of the classes in the SL Ontology, including object and data properties are presented in Attachment 3.

4.1.3 Interoperability

The study carried out regarding ontologies in the legal domain and the study of particular ontologies focused on legislation presents relevant information concerning the interoperability required for this ontology.

The interoperability characteristic of SL Ontology is mainly carried out through ontology alignment [Noy and Musen, 1999]. Noy and Musen define ontology alignment as follows: ontology alignment consists in establishing links between ontologies and allowing the aligned ontologies to reuse information one from another.

From the study of the Greek Public Administration Ontology and the Legislation.gov.uk project ontology it is possible to assess some facts: 1) Dublin Core is the main standard concerning metadata; 2) FRBR is an important reference regarding the organization of bibliographic records and is included in both described frameworks (MetaLex and Akoma Ntoso); 3) FOAF is the main reference concerning the modeling of entities, at least regarding people, organizations and groups.

Concerning interoperability the second fact is of the most importance. In order to develop a strong interoperability capacity for the SL Ontology it necessary to enable the exchange of information between the ontologies referenced.

Therefore, the linkage of the knowledge stored by this system is accessible through the consumption of information according to well-known and used ontologies. The following figure illustrates this characteristic.

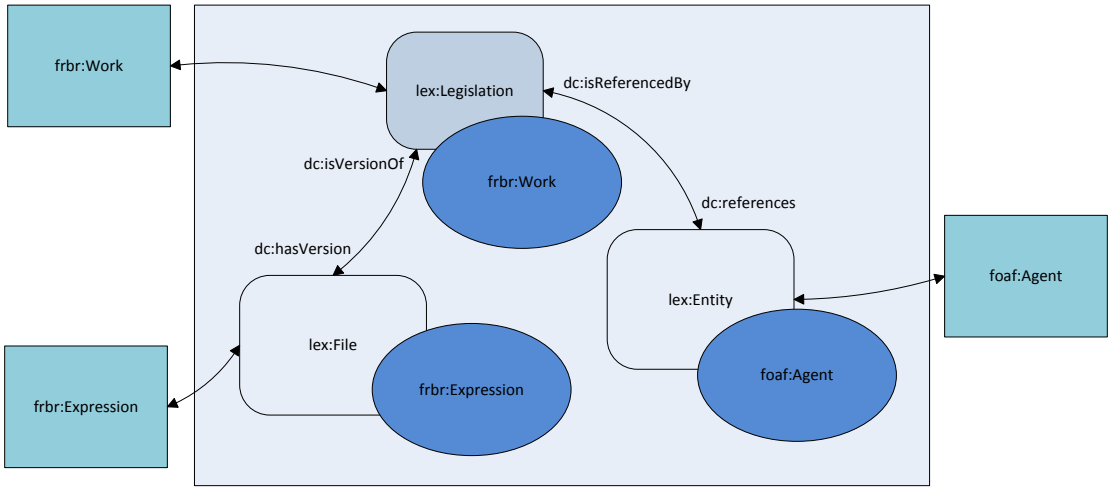


Figure 37 – Interoperability characteristic

Primarily, the SL Ontology classes do not instantiate individuals according to every legal domain standard. It is considered that this would imply a continuous effort to update the ontology. By developing the SL Ontology as a coarse-grained ontology it is considered that the level of abstraction required does not imply this effort.

4.2 Information Storage

This section describes the necessary components to ensure the storage of information extracted from the legislation documents. This is composed by two components: the database and the connection between the system and the database. Both are described as follows.

4.2.1 Database

In order to store the information extracted from the extraction phase, using the SL Ontology, it is necessary to use a compatible database. For this effect a triple store was implemented. A triple

store is a framework used for storing and querying RDF data. It provides a mechanism for persistent storage and access to RDF graphs. The triple store implemented in this system is Jena SDB is a non-native non memory triple store. This type of triple store is set up to run on third party databases. In this system, the database used is MySQL[21].

Jena SDB is a Java framework for building Semantic Web applications. Jena[22] provides a collection of tools and Java libraries that support the development of semantic web and linked data apps, tools and servers. SDB[23] is a component of Jena for RDF storage and query specifically to support SPARQL The storage is provided by an SQL database and many databases are supported, both Open Source and proprietary. An SDB store can be accessed and managed with the provided command line scripts and via the Jena API.

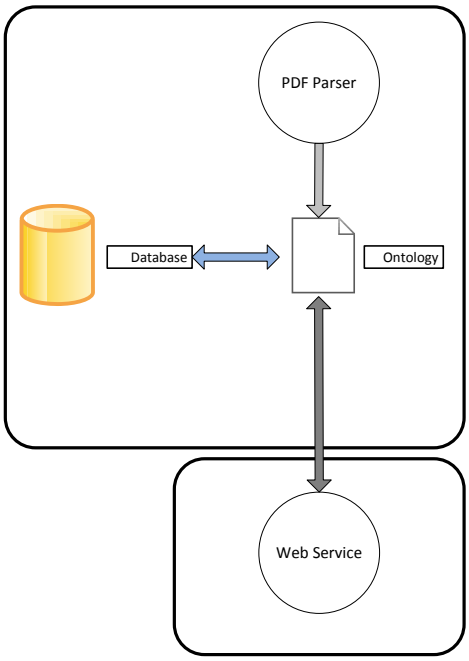


Figure 38 – Database flow of information

The database is accessed in two processes: 1) the storage of information extracted from the PDF Parser component and 2) the intake of information from the Web Service in order to provide responses to the user requests.

In the first process, the PDF Parser feeds information that is translated according to the SL Ontology and then stored in the database (triple store). In the second process, the Web Service component consumes knowledge from the database according the SL Ontology. The response from the second process is in the form of N-triples [W3Ca, 2004]. N-Triples is a line-based, plain text format for

[21] MySQL 5.5.8 - http://www.mysql.com/ [last access: Sep 2012]
[22] Apache Jena - http://jena.apache.org/ [last access: Sep 2012]
[23] Apache Jena SDB - http://jena.apache.org/documentation/sdb/index.html [last access: 2012]

encoding an RDF graph. Each line represents one statement of information and is contains three parts: subject, predicate and object.

An example [W3Ca, 2004] of a N-Triples file is presented as follows.

```
<http://www.w3.org/2001/08/rdf-test/> <http://purl.org/dc/elements/1.1/creator> "Dave Beckett" .
<http://www.w3.org/2001/08/rdf-test/> http://purl.org/dc/elements/1.1/creator "Jan Grant" .
http://www.w3.org/2001/08/rdf-test/ http://purl.org/dc/elements/1.1/publisher _:a .
_:a <http://purl.org/dc/elements/1.1/title> "World Wide Web Consortium" .
_:a <http://purl.org/dc/elements/1.1/source>     <http://www.w3.org/> .
```

4.2.2 Connection between System and Database

The feeding of information from the PDF Parser component to the database is done through the insertion of information using the SL Ontology for translation.

After the extraction and storage of information the PDF Parser component uses a connector to ensure the connection between the component and the database. The connector used is MySQL Connector/J[24] which provides connectivity for client applications developed in the Java programming language through a JDBC driver. MySQL Connector/J is a JDBC Type 4 driver, which means that it is a pure-Java implementation of the MySQL protocol and does not rely on the MySQL client libraries.

After ensuring a connection with the database, the feeding of information requires two artefacts: an ontology and a reasoned. The ontology used, the SL Ontology, has been described. The reasoner used is Pellet[25]. A reasoner is an artefact used to represent and reason about information. Pellet is described as a complete OWL-DL reasoner with very good performance, extensive middleware, with extensive support for reasoning with individuals and user-defined data types [Sirin et al., 2011].

Concerning the queries of information made by the Web Service, the process is equivalent. The database queries response is presented in the form of N-triples, according to the SL Ontology, using Pellet to reason about the information and presenting it.

4.3 Indexing Documents – Index

In order to provide a search engine capable of enabling the finding of information in the legislative text, it is necessary to create an index of that information. This is done by the component Index. This process requires the input of the legislative text by the PDF Parser component. Then, the Index component parses the text in order to create an index of that information to enable queries.

[24] MySQL Connector/J 5.1.16 - http://dev.mysql.com/doc/refman/5.1/en/connector-j-overview.html [last access: Sep 2012]
[25] Pellet: OWL 2 Reasoner for Java - http://clarkparsia.com/pellet [last access: Sep 2012]

In order to accomplish this process, the framework Apache Lucene[26] is used. Lucene is a high-performance, full-featured, open source text search engine library written entirely in Java. It is suitable for nearly any application that requires full-text search, especially cross-platform.

The PDF Parser component feeds information concerning entities and legislation. Concerning entities, the name of the entity and the correspondent resource URL; concerning legislation, the name of the issuing entity, the name of the issuing sub-entity if provided, the title of the document, the description if provided, the URL of the resource and the document's full text.

Lucene provides specifications to comply with the user's necessities. Namely, it provides the user with the option of storage and indexing. In the following table these options are presented.

Table 11 – Lucene specifications

Type	Field	Store	Index
Legislation	Issuing Entity	Yes	Yes
Legislation	Issuing Sub-Entity	Yes	Yes
Legislation	Title	Yes	Yes
Legislation	Description	Yes	Yes
Legislation	Text	No	Yes
Legislation	URL	Yes	No
Entity	Entity	Yes	Yes
Entity	URL	Yes	No

Lucene provides the choice of storage medium for the resulting index. It may be stored in a database or in a given file. For the purposes of this system, the index is stored in a file in the system. This choice was decided due to the facilitation of copies of the index (for backup purposes for example) and due to avoidance of oversize or overloading issues of the database.

[26] Lucene 3.6.1 - http://lucene.apache.org/core/ [last access: Sep 2012]

5 Information Access – Phase 3

The Information Access phase is the last phase of this project. It is composed by the development of two components: the Web Service and the Mobile Application. This phase represents the part of the system that external actors (e.g. users, systems) may use in order to access and reuse the knowledge stored in the previous phase.

The Web Service is the point-of-access to the system and represents the component that allows the reuse of the knowledge stored. The Mobile Application is developed in order to enable users to access to the knowledge contained in the system. The following section describes the development and implementation of the referred components.

5.1 Web Service

According to the definition of W3C, a Web Service is a software system designed to support interoperable machine-to-machine interaction over a network. It has an interface described in a machine-processable format (specifically WSDL [W3C, 2001]). Other systems interact with the Web service in a manner prescribed by its description using SOAP-messages, typically conveyed using HTTP with an XML serialization in conjunction with other Web-related standards [W3Cb, 2004].

Also, W3C identifies two major classes of Web Services: *REST-compliant Web Services*, in which the primary purpose of the service is to manipulate XML representations of Web resources using a uniform set of "stateless" operations; and *arbitrary Web services,* in which the service may expose an arbitrary set of operations [W3Cc, 2004].

Muehlen et al. [Muehlen et al., 2005] presents a discussion between the options of a REST [Fielding, 2000] or SOAP-compliant Web Service in which this system decision is based. The following table presents the advantages and disadvantages presented by the authors.

Table 12 – Characteristics of REST and SOAP [Muehlen et al., 2005]

	REST	SOAP
Characteristics	Operations are defined in the messages Unique address for every process instance Each object supports the defined (standard) operations Loose coupling of components	Operations are defined as WSDL ports Unique address for every operation Multiple process instances share the same operation Tight coupling of components
Self-declared advantages	Late binding is possible Process instances are created explicitly Client needs no routing information beyond the initial process factory URI Client can have one generic listener interface for notifications	Debugging is possible Complex operations can be hidden behind façade Wrapping existing APIs is straightforward Increased privacy
Possible disadvantages	Large number of objects Managing the URI namespace can become cumbersome	Client needs to know operations and their semantics beforehand Client needs dedicated ports for different types of notification Process instances are created implicitly

Considering the general description of the system and the phases and processes required, as well as the usage and reusability characteristics required, a REST-compliant Web Service was developed.

It is considered that according to the definition of each framework and the discussion provided by Muehlen et al., that the REST framework is resource-oriented and SOAP framework is process/operation-oriented. Given that one of the objectives of the system is to provide a reusable-compliant access, it is considered that a REST Web Service solution is the more appropriate. This decision values the specific REST characteristics of being *stateless* and having a unique address for every process instance.

A REST Web Service is implemented using the API framework for Java, Restlet[27]. This API provides the ability to design the REST Web Service according to the desired structure organization of the requests, which will be presented further.

5.1.1 Structure

The structure of the Web Service must comply with the organization of the knowledge stored in the previous phase. Therefore it must provide access to the legislative documents and entities, as well as lists of both. Additionally, it enables full-text search of the legislative documents and therefore it implements this process as well.

[27] Restlet 2.0 - http://www.restlet.org/

The following table presents the organization of the Web Service regarding the structure of the resources available.

Table 13 – Web Service Structure

Resource	**HTTP Method**	**Description**
.../legislation	GET	List of all legislation documents
.../legislation/latest	GET	List of the ten most recent legislation documents
.../legislation/[lex_id]	GET	Information concerning a given legislation
.../legislation/[lex_id]/version	GET	Versions of the text of a given legislation
.../legislation/[lex_id]/references	GET	List of references of a given legislation
…/legislation/[lex_id]/referrals	GET	List of referrals to a given legislation
.../entity	GET	List of all entities
.../entity/[ent_id]	GET	Information concerning a given entity
.../entity/[ent_id]/entities	GET	List of entities mentioned together with a given entity in the legislation documents
.../search	POST	Text search of the legislation documents text

Since there is no objective of allowing the users to modify the knowledge stored and the resources available enable its access, this Web Service only uses two HTTP methods: GET and POST.

The HTTP requests to these resources require queries to the Database and the responses are encoded in XML format according to the pre-defined data structure. These subjects are presented as follows.

5.1.2 Processes

The Web Service component processes involve three other components: User/Mobile Application, Database (Ontology system component) and Index (Ontology system component).

Regarding the access to information, either lists of legislation or entities and information regarding a given legislation or entity, the processes involve the User/Mobile Application and the Database. Regarding the search procedure, it involves the User/Mobile Application and the Index. The following Sequence Diagrams illustrate an example of both these processes.

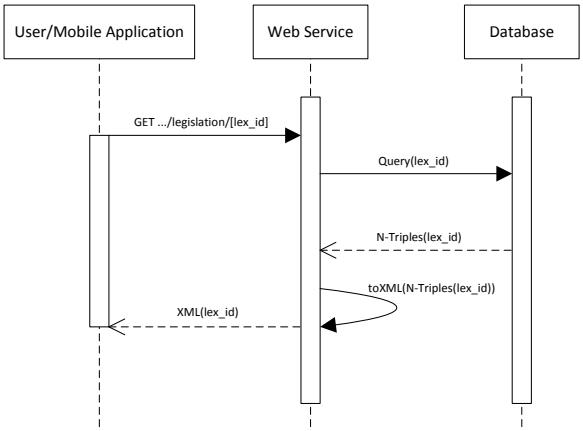


Figure 39 – Sequence Diagram of knowledge concerning a given legislation request example

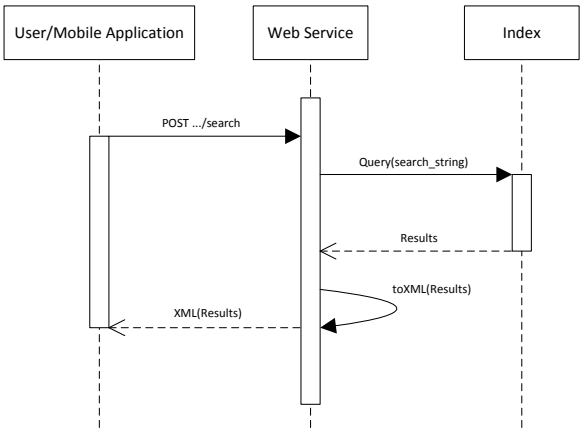


Figure 40 – Sequence Diagram of text search request example

In the first example a request is done using the GET HTTP Method to a resource in the Web Service; the Web Service then queries the Database concerning information regarding a given legislation (using its legislation identifier); the Database responds to the Web Service in the form of N-Triples, containing the referred information; the Web Service encodes that information into XML according to a developed data structure (presented in the following section); and returns the final results to the User/Mobile Application.

In the second example a request is done using the POST HTTP Method to a resource in the Web Service; the Web Service queries the Index concerning the presence of a given string (contained in the POST request) in the legislation documents text; the results are interpreted by the Web Service and encoded into XML according to a developed data structure; and the Web Service returns the search request results to the User/Mobile Application.

The following section presents the data structure developed and used in the responses of the Web Service.

5.1.3 Data Structure

The Web Service component is responsible for encoding the responses from the Database and the Index in order to produce a machine-readable response for the Mobile Application or a given user that accesses the component directly.

Concerning the responses from the Database and the Index, the former is accessed through a query containing the data correspondent to the request; the response from the Database is encoded in triples (subject, predicate and object). The latter is accessed directly from the Web Service and thus the processing of the request is made in the scope of the component; the results are extracted from a data structure.

The response of the Web Service is encoded in XML. Other formats could be used, such as JSON[28]. Regarding JSON, although it is a widely used format in terms of communication between components, it is considered that its potential is more focused on carrying messages that contain data structure objects. In comparison with XML, and considering that the communications in this system include the representation of documents, the option to use XML messages was decided.

Regarding the Database, the following figure illustrates an example response given a request from the Web Service in order to obtain the latest legislation documents that the Database contains.

[28] JavaScript Object Notation (JSON) – http://www.json.org/ [last access: Sep 2012]

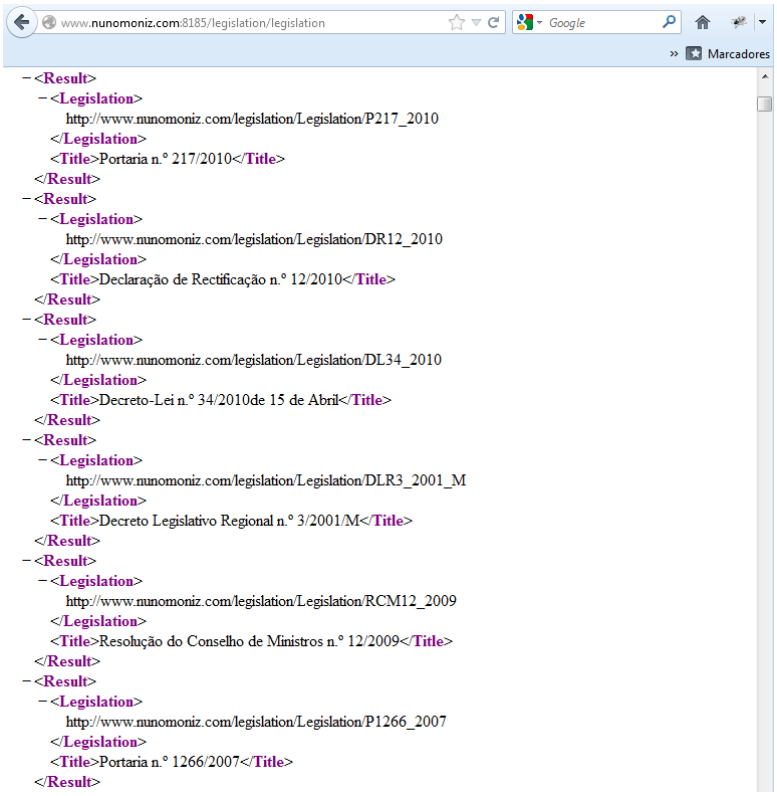


Figure 41 – Example of XML response

The basic structure of the XML responses corresponds to an encapsulation of “Result” objects in the XML parent node, “Results”.

The “Result” objects contain objects with the description of each result obtained from the query. In the former case, concerning legislation documents, it is composed by two other objects: “Legislation” object, containing the URL of the resource; and a “Title” object, containing the title of the legislation document.

Concerning the implementation of this system, the following table depicts the composition of each “Result” object correspondent to each of the processes available in the Web Service.

Table 14 – Composition of “Result” objects in Web Service responses

Resource	Legislation	Entity	Title	Name	Relation	Object	Link	hasVersionIn	RelatedEntity
.../legislation	x		x						
.../legislation/latest	x		x						
.../legislation/[lex_id]	x				x	x			
.../legislation/[lex_id]/version	x							x	
.../legislation/[lex_id]/references		x		x					
.../legislation/[lex_id]/referrals	x		x						
.../entity		x		x					
.../entity/[ent_id]		x			x	x			
.../entity/[ent_id]/entities									x
.../search			x				x		

The request regarding versions of a legislation include, under each “hasVersionIn” object, the correspondent XML representation of the version of a given legislation document.

5.2 Mobile Application

This section describes the Android Mobile Application developed to provide a means for visualization of the knowledge and to test its presentation.

The access to the information is carried out through the available resources in the Web Service component described previously. The Mobile Application requests the information from the Web Service pending the available actions in the application; formats the responses and produces visualization. The following sections describe the development of the application and its designed layouts.

5.2.1 Description

The developed Mobile Application produced visualization to a given set of available information requests by the Web Service. The Mobile Application enables the visualization of:

- Latest legislation documents
- Complete list of legislation documents
- Complete list of entities
- Legislation documents details
- Entities details
- Entities referenced by a given legislation

- Referrals to a given legislation
- Related entities of a given entity
- Legislation documents text

The complete project tree is presented in Attachment 5.

5.2.2 Interface

For each of visualizations described in the previous section, a layout was developed.

Some of the visualizations required represent lists either of legislation documents or entities. For each of these situations a respective layout was developed. For legislation documents, concerning the visualization of the latest legislation documents, complete list of legislation documents and referrals to a given legislation, a layout was developed that simply contains a `ListView`. The following figure presents a depiction of this layout. The developed layout for lists of entities is identical.

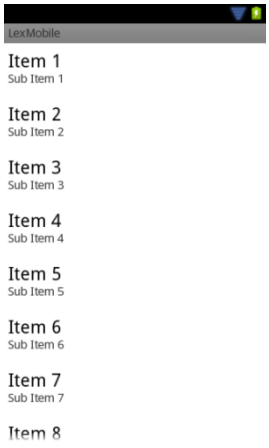


Figure 42 – Layout for visualization of legislation documents lists

The layout developed to provide visualization for the details of legislation is presented in the following figure. The visualization contains the title, the date the legislation was made available in the system, the date it was issued, the link of the resource and the link to the original PDF document. The option to visualize the references of the legislation document as well as the referrals and the most recent version of the legislation document text is also provided.

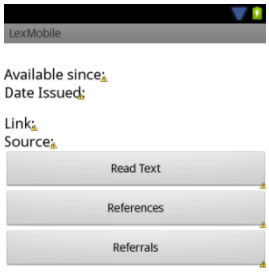


Figure 43 – Layout for visualization of legislation documents details

The following figure presents the layout for visualization of entities. The presented information concerning an entity includes its name and the link of the resource. It also provides the option to visualize related entities of a given entity.

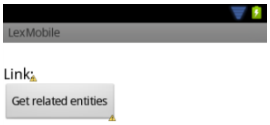


Figure 44 – Layout for visualization of entities details

Furthermore, the Mobile Application provides a visualization for the text of a given legislation document. It is composed by a single `TextView`. It also provides a visualization of the latest legislation documents present in the system, present in the main layout. This layout is identical to the legislation and entity lists with the addition of a tab bar.

6 Implementation

This chapter presents the details of implementation of the described system. The system was deployed in an online environment and tested using both the Mobile Application and a browser for the web environment tests.

This chapter describes the organization of the server concerning the various components of the system. The implemented system in the server contains the components PDF Parser, Ontology (SL Ontology and Database), Index and Web Service. This chapter also contains the description of the execution of the several components and the flow of information between them.

6.1 Server Design

This system was implemented in a specific domain (www.nunomoniz.com). Regarding the system design, it contains all the components described previously except for the Mobile Application. The following figure depicts the design of the system and its channels of communication.

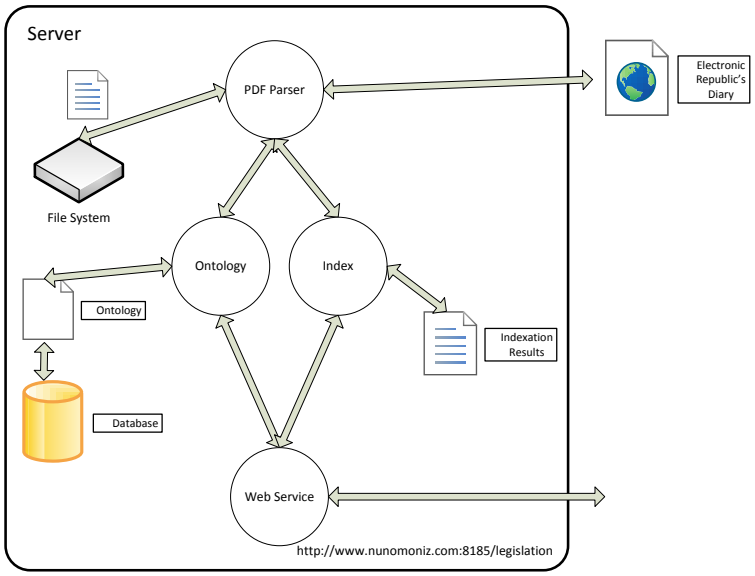


Figure 45 – Implemented Server Design

The previous figure shows the organization of the system and provides information concerning the Web Service. The Web Service is implemented in the server and serves the requests that are addressed to the port 8185 in the location `http://www.nunomoniz.com:8185/legislation`.

According to this implementation description, the following table provides the location of the resources of the Web Service correspondent to the available processes.

Table 15 – Web Service Structure implemented

Resource	HTTP Method
http://www.nunomoniz.com:8185/legislation/legislation	GET
http://www.nunomoniz.com:8185/legislation /legislation/latest	GET
http://www.nunomoniz.com:8185/legislation /legislation/[lex_id]	GET
http://www.nunomoniz.com:8185/legislation /legislation/[lex_id]/version	GET
http://www.nunomoniz.com:8185/legislation /legislation/[lex_id]/references	GET
http://www.nunomoniz.com:8185/legislation /legislation/[lex_id]/referrals	GET
http://www.nunomoniz.com:8185/legislation /entity	GET
http://www.nunomoniz.com:8185/legislation /entity/[ent_id]	GET
http://www.nunomoniz.com:8185/legislation /entity/[ent_id]/entities	GET
http://www.nunomoniz.com:8185/legislation /search	POST

The namespace of the ontology was defined in accordance with this implementation also.

Additionally, a SPARQL server was implemented in order to provide the ability to query the triple store directly. Joseki[29] is a SPARQL Server for Jena. It contains a HTTP engine that supports the SPARQL Protocol and the SPARQL RDF Query language. It is accessible in the URL http://www.nunomoniz.com:2020. It provides two options concerning the execution of queries: through an interface or through a direct request using an URL.

The following figures depict the accessible interface and the response.

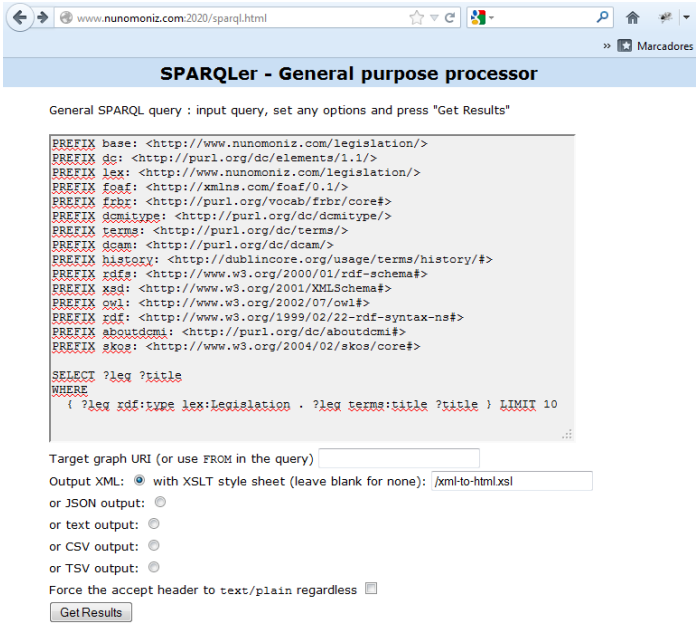


Figure 46 – Joseki interface

[29] Joseki – A SPARQL Server for Jena http://www.joseki.org/ [last access: Out 2012]

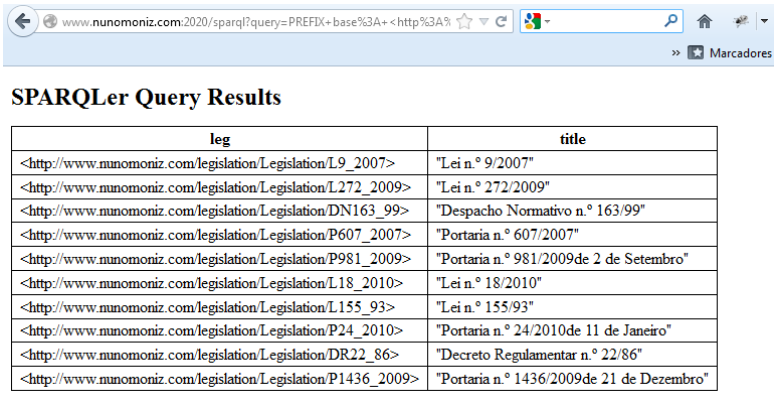


Figure 47 – Joseki query results[30]

Also, an auxiliary component was developed in order to integrate information from DBPedia. DBPedia is a project aiming to extract the structured content of Wikipedia. In order to integrate the information present in our database, this component queries the DBPedia SPARQL endpoint in an attempt to link the information in both silos. This component is executed after the extraction and storage of information from the Republic's Diary PDF documents.

The primary objective of this auxiliary component is to discover the link of the DBPedia resource that refers to a given entity in our triple store. A query example could be `SELECT ?subject WHERE { ?subject rdfs:label 'Jaime Gama'@pt }` using the available Virtuoso SPARQL Query Editor[31]. The following figure illustrates the response of this example query.

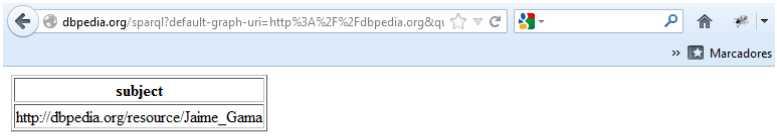


Figure 48 – DBPedia SPARQL endpoint response example

6.2 Information Flow

This section describes the flow of information between the components and the external parties such as the Electronic Republic's Diary site.

[30] The simplified direct link for the query is http://www.nunomoniz.com:2020/sparql?query=PREFIX+lex%3A+%3Chttp%3A%2F%2Fwww.nunomoniz.com%2Flegislation%2F%3E%0D%0APREFIX+terms%3A+%3Chttp%3A%2F%2Fpurl.org%2Fdc%2Fterms%2F%3E%0D%0APREFIX+rdf%3A+%3Chttp%3A%2F%2Fwww.w3.org%2F1999%2F02%2F22-rdf-syntax-ns%23%3E%0D%0ASELECT+%3Fleg+%3Ftitle%0D%0AWHERE+%0D%0A++{+%3Fleg+rdf%3Atype+lex%3ALegislation+.+%3Fleg+terms%3Atitle+%3Ftitle+}+LIMIT+10&default-graph-uri=&output=xml&stylesheet=%2Fxml-to-html.xsl

[31] Virtuoso SPARQL Query Editor – http://dbpedia.org/sparql

Regarding the flow of information between the components, there is no significant modification in comparison with the formerly described. The PDF Parser feeds the information produced by itself to the Database (using the developed SL Ontology) and to the Index. To the Database, the PDF Parser feeds the metadata concerning a legislation document; to the Index, it feeds the text of the respective document. The Database is queried by the Web Service according to the request made; the Web Service also queries the Index results in order to obtain information regard text search requests.

Due to the fact that the PDF Parser component handles text and XML documents, some issues occurred due to default character encoding (UTF-8[32]). Tests were carried out in order to decide on the appropriate character encoding to be used in the implemented server. Tests show that the character encoding ISO-8859-1[33] is the best choice. Therefore the command used[34] to execute the PDF Parser component is adjusted in order to comply with this decision.

However, concerning the component PDF Parser, due to the necessity of fetching the correct links to the correspondent PDF documents containing the legislation documents, a Web Crawler was developed. The next section explains the development of this auxiliary component.

6.2.1 Web Crawler

The PDF Parser has two operating modes in this system. It is responsible for parsing the Republic's Diary PDF documents that are continuously issued. But, it is also responsible for fetching and parsing a certain group of formerly issued documents. Therefore, it is necessary to develop an automatic access to the link of the documents in order to perform both of these operations.

Regarding this situation, a study of the Electronic Republic's Diary was carried out in order to assess this possibility. This study comprehends the analysis of the HTML pages and the communication headers (HTTP). The later was carried out using Firebug[35]. Firebug is an add-on for Mozilla's web browser Firefox[36] which enables the edit; debugging and monitor of CSS, HTML and JavaScript live. It also enables the analysis of HTTP headers of responses to requests carried out. This is crucial for the described study. The following figure presents a print screen of the carried out study environment.

[32] Unicode Transformation Format-8 – http://www.utf-8.com/ [last access: Sep 2012]
[33] ISO-8859-1 – http://www.w3schools.com/tags/ref_entities.asp [last access: Sep 2012]
[34] The SSH command used to execute the PDF Parser component is `java -jar -Dfile.encoding=ISO-8859-1 /legislation/PdfParser.jar`
[35] Firebug 1.10.3 – http://getfirebug.com
[36] Mozilla Firefox – http://www.mozilla.org/en-US/firefox/fx/

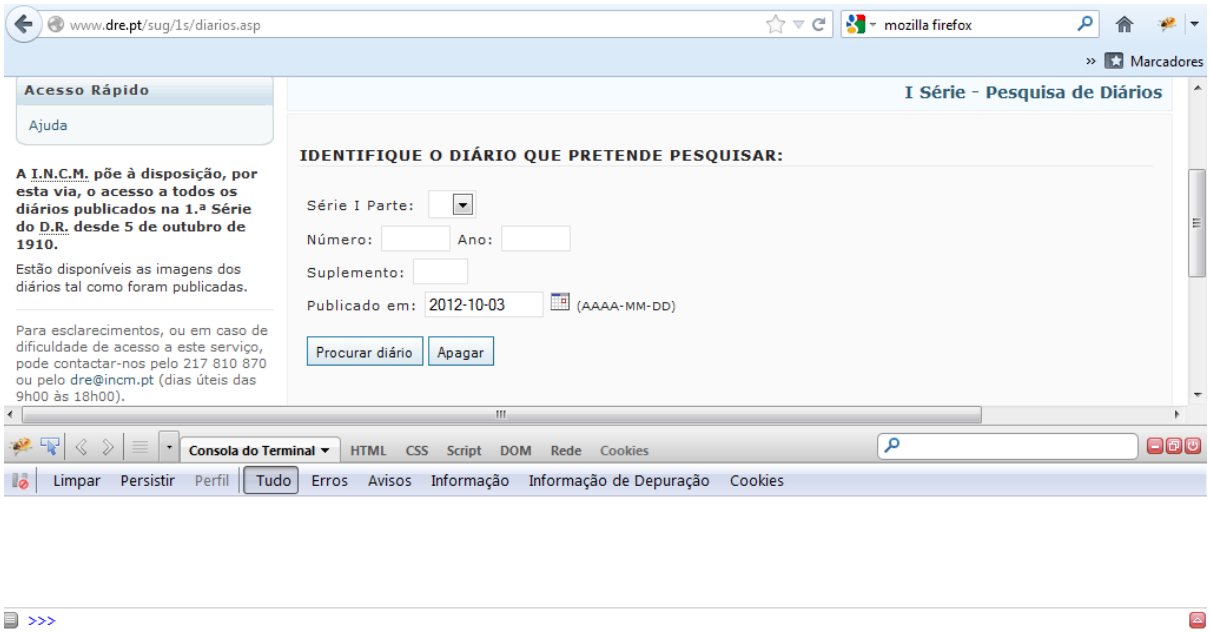


Figure 49 – Web Crawler study

In order to study the HTTP headers, a request for the Republic's Diaries published on the date 3rd of October 2012 was carried out. The result is depicted in the following figure.

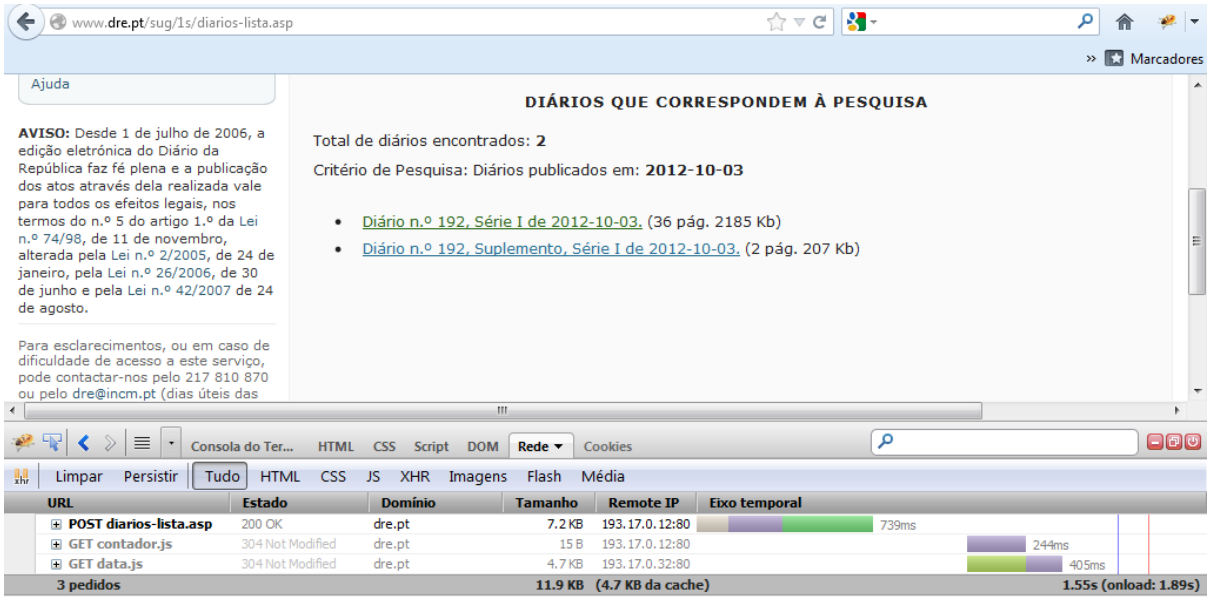


Figure 50 – Firebug results concerning HTTP headers

We find that there are three requests carried out in this process. The `GET` requests concern the fetching of information regarding the online users (`GET contador.js`) and static information to fill others frames on the design of the web page (`GET data.js`). Regarding the development of the Web Crawler, the `POST` request is analysed. The `POST` request sent, according to Firebug monitoring results is presented in the following figure.

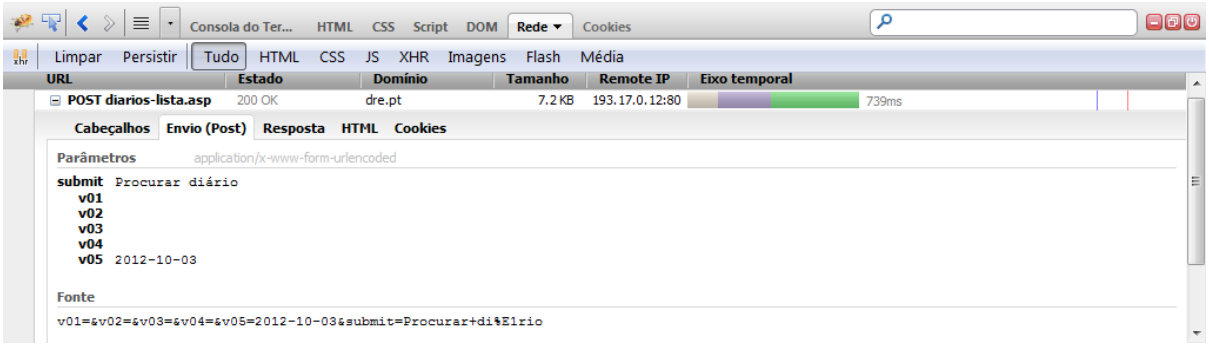


Figure 51 – POST request results from Firebug

These results provide the information necessary for the development of the access part of the Web Crawler. In the former figure the parameters and the composition of the POST request is made explicit and therefore it is now possible to access the information provider[37]. Concerning the requests that the developed system requires, it will only be necessary to post requests using the parameter `v05` which is related to the issue date of the document.

Regarding the retrieval of the documents link, it is necessary to parse the response of the `POST` request. By analysing that response it is discovered that it is a HTML response and that is contains the links to the requested diaries. They are made explicit under a `<li>` tag. The link is selected in the following figure.

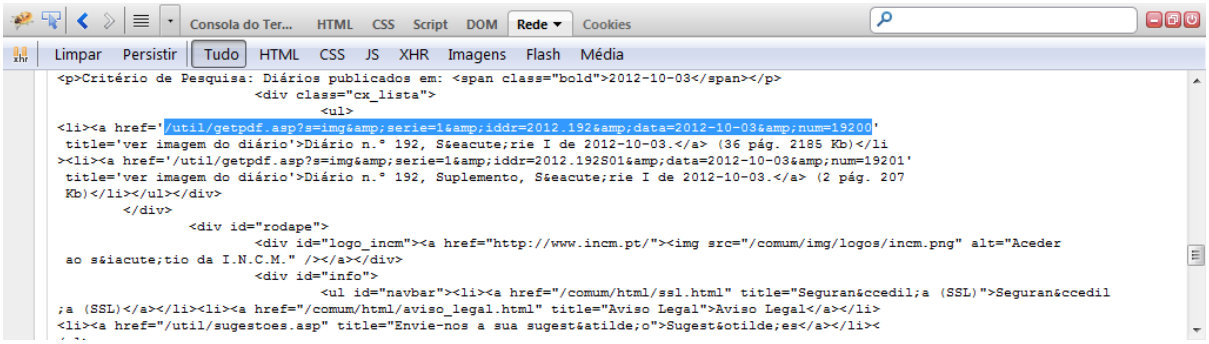


Figure 52 – HTML response to the POST request

Thus, by using a regular expression[38] capable of extracting the URI selected in the former figure, it is possible to obtain the URL of the document. By using these results, and through the study carried out, it is concluded that the URI of the document should be concatenated with the namespace of the Electronic Republic's Diary. In the case depicted in the former figure, the complete link is `http://www.dre.pt/util/getpdf.asp?s=img&serie=1&iddr=2012.192&data=2012-10-03&num=19200`[39].

[37] Information provider refers to the point of information access in the Electronic Republic's Diary site. From the results gathered, that point is http://www.dre.pt/sug/1s/diarios-lista.asp

[38] The regular expression used to capture the URI of the document is "`/util/getpdf.+?num=\\d+`"

[39] The expressions "`amp;`" are removed

Thus, the Web Crawler component is composed by these two processes. The POST request in a first phase, and the analysis and retrieval of the document's URI, in a second phase. Regarding the two operation modes of this component, the outputs may be different. Concerning the implementation of the system, it is required to process a group of Republic's Diary PDF documents, related to the dates decided. In this case, the output is a list of URL; concerning the daily process of the system, it is required to process the latest Republic's Diary PDF document, and therefore the output is a single URL. In both cases, these URLs are transmitted to the PDF Parser component.

7 Results

This chapter presents the results obtained from the execution of the three phases described. Concerning the first phase – Information Extraction – the results presented are extracted from a group of 40 Republic's Diary documents (see Attachment 1). The results of the second phase describe the final result of the organization of knowledge concerning the legislation documents. It is focused on presenting the knowledge created concerning the legislation documents and entities. The third part of the results presents depictions of the Mobile Application developed. In the last section of this chapter an analysis of the results is presented.

7.1 Information Extraction

The information extraction phase was tested with a group of 40 Portuguese Republic's Diary PDF documents (see Attachment 1). The documents were chosen randomly in terms of size and date. For this performance test we did not include the Diaries supplements. Regarding the timeline of the documents, it stands between the 1st of January 2009 and 19th of March 2012. The access to the documents of our sample was done in an online environment – remote access.

For each document in our sample we confirmed if the text extraction was done in a correct and successful manner. The confirmation was based on a manual comparison between the original text in the PDF documents and the XML output. We also confirmed the extraction of entities; it was based on a one-by-one evaluation of each entity extracted. The extraction of entities refers to the ability of the system to recognize a given entity presence in the text.

The documents were graded, in terms of percentage, according to its accuracy in both processes: extracting text and extracting entities. We searched for unsuccessful text extractions and non-entities that were flagged as correct entities.

In our experiments we used the two measures: Text Extraction Accuracy (TEA) and Entity Extraction Accuracy (EEA). The measures were defined as follows:

$$TEA = 1 - (\frac{UTE}{TTE}) \quad (1)$$

$$EEA = 1 - (\frac{UEE}{TEE}) \quad (2)$$

Here, UTE and UEE are defined as Unsuccessful Text Extractions and Unsuccessful Entity Extractions; TTE and TEE are defined as Total of Text Extractions and Total of Entity Extractions. In the following table the results are presented.

Table 16 – Results of first phase evaluation

Period	TEA	EEA
Jan 2009 – Dec 2009	99,82%	93,55%
Jan 2010 – Dec 2010	99,53%	92,55%
Jan 2011 – Dec 2011	99,68%	94,31%
Jan 2009 – Mar 2012	99,73%	93,61%

For both confirmations, partial results were considered as wrong. As for the first confirmation (TEA), the incorrect extractions were promptly pointed by the system. Nonetheless, some results pointed out as incorrect were accepted due to the previous stated expectations: relating to text inside a table, we expect the system to ignore it. As such, these results were considered correct. However, in the second confirmation (EEA), we had to observe and classify one-by-one, each entity. Entities that were incomplete; had incorrect phrasing or minor errors were considered as wrong.

In the development of this evaluation, despite the well-defined layout structure, we found the use of different and unique combinations of fonts. This caused some of the text extraction errors. Most of the text extraction errors were due to minor incompatibilities (a space character misplaced, for example) between the content stream extraction and the layout extraction. At this point we are improving this situation through trial-and-errors. We are also considering different approaches in order to extract the text from the PDF documents, in the correct reading-order using only its content stream.

To complete this performance evaluation we would like to point out some global indicators that were obtained during this process. They are presented in the following table.

Table 17 – Additional first phase evaluation indicators

Indicator	Result
Average PDF size	696,5 Kb
Average Final XML size	101,5 Kb
Average page number per PDF	23
Average processing time per PDF	12 s
Average processing time per PDF page	0,5 s

7.2 Knowledge Organization

Tests were carried out concerning the organization of knowledge according to the designed and implemented SL Ontology, the second phase described. The process included the extraction of information from the Republic's Diary PDF documents from the 1st of September 2009 until the 21st of June 2010, and also the month of May, 2012. Access to the knowledge is done through the use of the Web Service available resources.

The test group of documents related to the referred time interval produced the results presented in the following table.

Table 18 – Results of second phase evaluation

Indicator	Result
Number of triples	~ 167.051
Database size	33.2 MB
Entities	11.018
Legislation documents	5.346
Republic's Diaries (Container)	199
Execution time	~ 3d 2h 44m

Further tests were carried out concerning the correctness of the organization of the knowledge according to the SL Ontology. The following figure depicts a request for the latest legislation documents inserted into the system.

Figure 53 – Second phase evaluation (latest legislation documents request)

In order to evaluate the correctness of the classification and the available information concerning a legislation document and an entity, two requests respective to each of the objectives were made. The following figures present the results.

Figure 54 – Second phase evaluation (classification of a legislation document)

According to the previous figure, the classification is according with the set parameters: a legislation document is classified as Legislation (`lex:Legislation`), Work (`frbr:Work`) and Document (`lex:Document`).

Figure 55 – Second phase evaluation (classification of an entity)

According to the previous figure, the system classifies an entity as Entity (`lex:Entity`) and Agent(`foaf:Agent`). These were the expected classification results.

Figure 56 – Second phase evaluation (legislation properties example)

In the previous figure some properties related to legislation documents are presented. These properties include the object property rights holder (`dc:RightsHolder`), the Republic's Diary that contains the legislation (`dc:isPartOf`) and issuing entity (`foaf:maker`), and data properties title (`dc:title`), link to the original PDF document (`lex:link`) and the date the legislation was issued (`dc:issued`).

Figure 57 – Second phase evaluation (references and referrals example)

The previous figure shows the references of, and referrals to the legislation. The former are described with the predicate `dc:references` and the latter with `dc:isReferencedBy`.

Figure 58 – Second phase evaluation (entity properties example)

The former figure presents object and data properties that describe an entity: the object property `dc:isReferencedBy` to describe the referrals to the given entity, and the data property `foaf:name` for the name of the entity.

The following figure depicts a request for the versions of a given legislation document text.

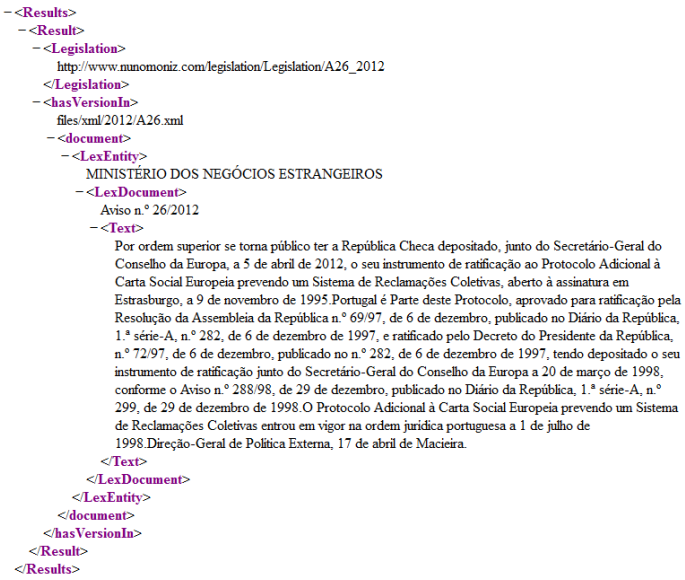


Figure 59 – Second phase evaluation (request for text version example)

To finish the evaluation of this phase, a depiction of the entity Jaime Gama is presented in the following figure, illustrating the retrieval of information from the DBPedia SPARQL endpoint and its integration in the system.

Figure 60 – Second phase evaluation (DBPedia information)

7.3 Information Access

This section is focused on the outcome of the visualizations of the Mobile Application. The results of the Web Service are presented in the previous section according to the requests made in order to obtain the necessary information.

A set of requests were carried out in the Mobile Application in order to access the information and produce the respective visualizations.

The main visualization presents the latest legislation documents in the system. The following figure presents a depiction of the visualization. This layout also contains the tab menu.

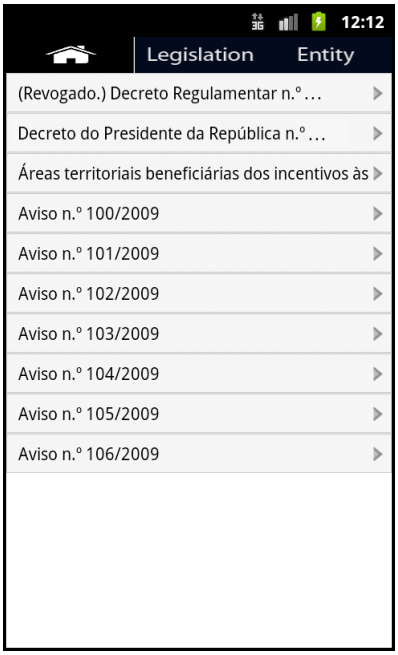


Figure 61 – Mobile Application evaluation (main visualization)

The following figure presents the depiction of the legislation tab. It provides the complete list of legislation documents in the system.

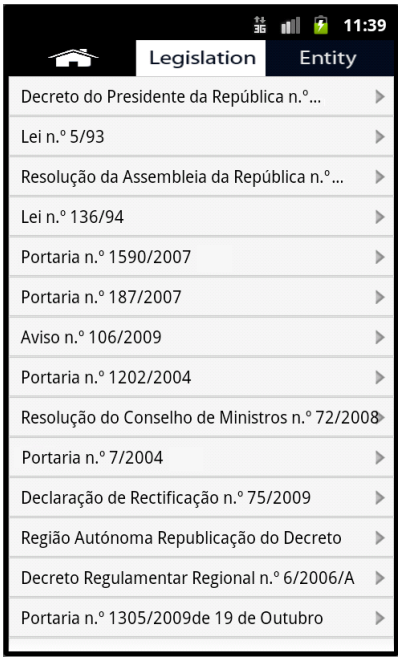


Figure 62 – Mobile Application evaluation (legislation documents list)

Concerning the list of entities, the following figure depicts the list retrieved by accessing the Entity tab.

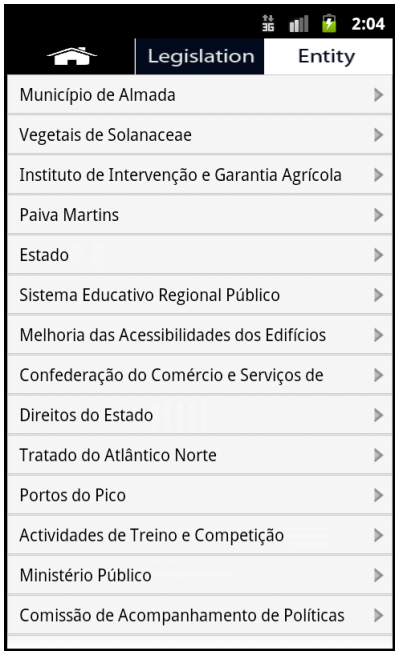


Figure 63 – Mobile Application evaluation (entities list)

The following figures present depictions of a request for details of a given legislation document and an entity.

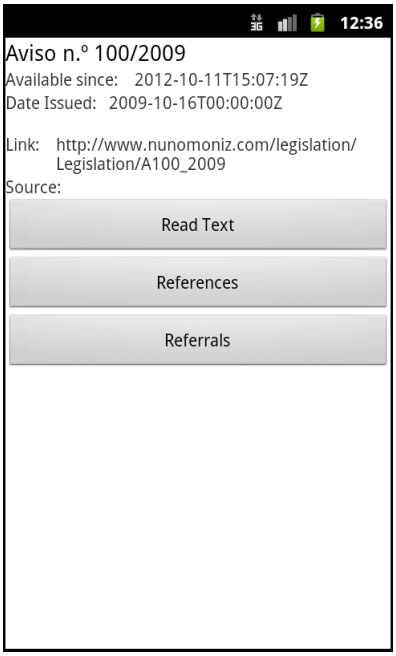


Figure 64 – Mobile Application evaluation (legislation document details)

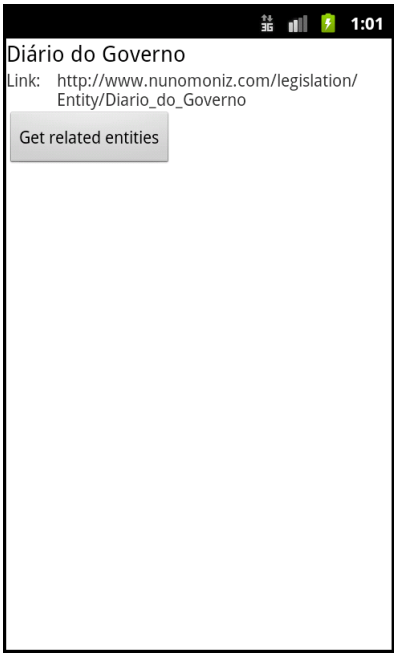


Figure 65 – Mobile Application evaluation (entity details)

In the legislation details layout it is possible to obtain the most recent version of the legislation document text, the entities referred and the referrals to the given legislation. The following figures present depictions of the entities referred by the legislation Aviso n.º 100/2009 and the referrals to the legislation Portaria n.º 1202/2004.

Figure 66 – Mobile Application evaluation (entities referred by legislation Aviso n.º 100/2009)

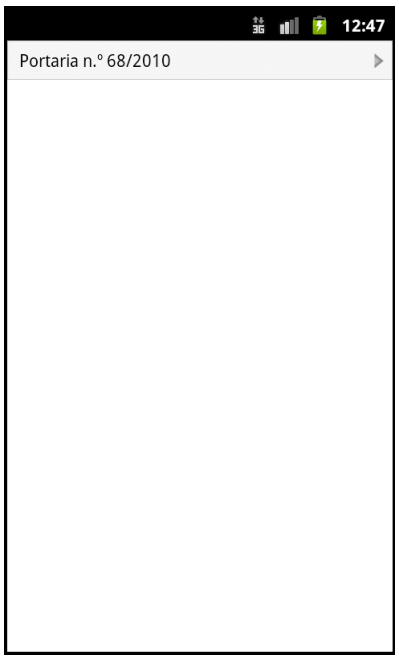


Figure 67 – Mobile Application evaluation (referrals to legislation Portaria n.º 1202/2004)

Regarding the entity details layout, it is possible to obtain the entities that are referred with a given entity, therefore, its related entities. The following figure depicts an example for the entity Diário do Governo.

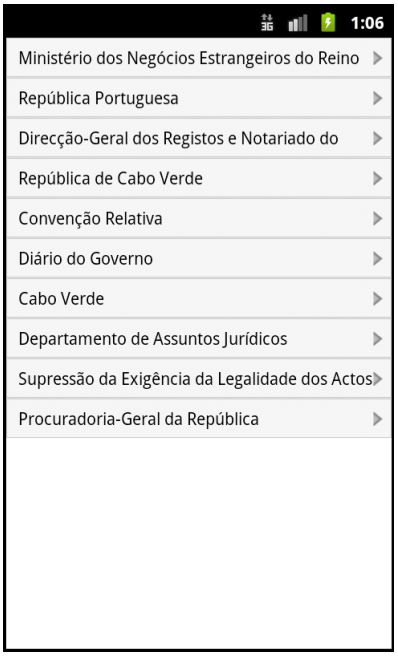


Figure 68 – Mobile Application evaluation (related entities of entity Cabo Verde)

7.4 Results Analysis

Results from the three phases are presented. This section presents an analysis of the results. According to the results it is possible to say that the system does comply with the set objectives for this project.

However, in the first phase, data extraction, the influence of the misleading recognition of entities does have an effect on the posterior phases. The use of unexpected fonts and the geometric extraction are the main reasons responsible for this issue. The first issue corresponds to unexpected behaviour by the system, translated in inability to recognize some parts of the structure of the documents or to simply join two different blocks of the structure. The second issue is related directly with the recognition of entities. Due to the geometric extraction some phrases are joined and spaces are ignored. This issue is not tackled with the rules used and therefore, a small group of non-entities is recognized as such.

The second phase, knowledge organization, does comply with all the objectives that were set. The classification and the description of resources are according to the conceptualization of the SL Ontology and the vocabulary is properly used. The performance of this process denotes an issue concerning the time necessary to accomplish the organization of big sets of data extracted from the first phase. However, this was expected due to the size of information being extracted and organized.

The third phase, information access, also complies with the objectives that were set. The Mobile Application presents visualization for a group of resources of the Web Service and the request of

other resources are presented in the evaluation of the second phase. However, in these visualizations it is also possible to see the effect of the misrecognition of entities from the first phase.

8 Conclusions

A system capable of extracting the information from the Portuguese Republic's Diary PDF documents and transforming that information into knowledge is proposed. This system enables the access and re-use of such knowledge.

The proposed system contains three phases corresponding to the coarse-grain processes necessary to accomplish its goal. It is composed by a PDF Parser component, responsible for the extraction of information from the PDF documents; the Ontology component responsible for the storage of that information according to a given ontology; the Index component responsible for the indexing process of the documents text; the Web Service component, responsible for the processing of requests from third-parties and the retrieval of information; and the Mobile Application component, a test-case application, that enables a visualization of the knowledge contained in the system's Database.

The developed system provides an automated information extraction from available documents as well as the storage of open data copies of the documents text. It also provides new and more versatile manners of searching the Portuguese legislation: it allows the search of entities that are present in the legislative texts, references to entities and referrals to legislation documents, the relations between entities, amongst others.

This system can also be classified as a five-star system according to the rating system for Linked Open Data of Tim Berners-Lee. The data is available on the web, in machine-readable structured data. The data is published in non-proprietary format and uses RDF and SPARQL to identify the resources. Also, the system uses data from other silos namely DBPedia.

8.1 Achievements

- Comprehensive investigation regarding information extraction from PDF documents
- An approach proposal for extraction of text, structure, and entities in PDF documents
- Design and implementation of a system that enables the translation of PDF collections into non-proprietary formats (XML, RDF)
- Comparative study of available solutions regarding ontologies in the legal domain
- Design and development of a new ontology for the legal domain, SL Ontology
- Development of a Linked Data database of Portuguese Legislation with free access and information reutilization
- Creation of a legislative text index of the Portuguese Legislation
- Development of a Web Service capable of sharing the knowledge that the developed system extracted
- Implementation of a SPARQL endpoint to allow direct SPARQL queries

- Development of an Android Mobile Application to provide visualizations for the responses of information requests

8.2 Limitations and Future Work

Concerning the proposed and implemented approach of information extraction, the main objective was to achieve a structure, text and entities extraction system from PDF documents that would be simple, fast and able to receive inputs from the user. Simple because we still need a solution that is flexible; fast because the volume of PDF documents used requires a system with the ability to process a large number of documents; and a user-guided system, because this is directed for cases where there is more specific knowledge than general knowledge [Klink and Kieneger, 2001], and that specific knowledge is static throughout every document of that type.

There are some immediate subjects to improve or develop in order to achieve a more enthusiastic result.

Tests have shown that due to the often use of unexpected fonts in the text, results can be misleading. However, it showed that although it reduces the ability for classification of the text through a rule based approach, the system still generally recognizes it as valid text strings.

The use of an ontology based component instead of the developed rule based was not contemplated. Nonetheless, this presents an inevitable question for the future, due to the present growth of Semantic Web [Hendler et al., 2002].

However, with the use of the developed ontology, SL Ontology (A Simple Ontology for Legislation), the extracted information is in fact translated into triples and stored in a triple-store (Database).

The developed ontology is defined as a minimal set of core services and axioms that are required to organize and describe the knowledge extracted from the Republic's Diary documents. The main characteristics incorporated are that it should be small, interoperable and shareable. The small characteristic is incorporated due to the classification of the ontology (coarse-grained ontology), in order to pursue other objectives such as interoperability and share-ability.

Other ontologies were studied, with a special focus on the Legislation.gov.uk project ontology and the Greek Public Administration Ontology. Proposed XML standards were also studied, such as MetaLex and Akoma Ntoso. The outcome of the study presents the requirements for the achievement of the interoperability characteristic, namely, the use of top-level ontologies such as FOAF, FRBR and Dublin Core Metadata Initiative (DCMI).

Furthermore, the used vocabulary is important in order to achieve the shareable characteristic. It is mainly focused on the independencies of the vocabulary. Therefore, in order to provide a share-ability characteristic the ontology was developed as being country, language and jurisdiction-independent.

In the development phase of the ontology it was stated that the alternatives in order to pursue the objectives set out for the system were not sufficient. However, in July 2012, Barabucci and others proposed an ontology (ALLOT: A Light Legal Ontology on TLC [Barabucci et al., 2012]) based on the Akoma Ntoso non-ontology that seemingly complies with the proposed objectives. It is meant to be used to expressing all the non-documental semantic data found in Akoma Ntoso documents and its surrounding systems. It is particularly useful to connect Akoma Ntoso documents to the external entities they refer to, entities that are stored in legacy knowledge-bases, newly created Linked Data silos and relational databases.

The ALLOT ontology is very similar to the proposed SL Ontology, although it is focused on the Akoma Ntoso non-ontology. It provides an alike basis of interoperability in terms of documents and entities, through the of FRBR and FOAF ontologies. Regarding future work, it would be necessary to implement and test the ALLOT ontology in order to compare the results. Nonetheless, the interoperable characteristic of the SL Ontology enables the ontology alignment with this ontology.

The results of the SL Ontology use in this context were enthusiastic. Regarding either the classification of documents and entities or the description of information using data and object properties the results were in accordance with the set objectives. However, it is necessary to pursue the reuse of other silos of information, enforcing the Linked Data characteristic of this system. For example, in terms of future work, it would be interesting to develop and use a system capable of classifying the entities with more precision, for example, classifying people, locations, countries or even government posts. The classes of the FOAF ontology are used in the SL Ontology, but no reuse of ontologies capable of describing these other entities was incorporated. They are classified as entities according to the SL Ontology class Entity. This could be a future update. Also, the tests and use of the ontology were limited to the Portuguese Legislation. Therefore, it is necessary to test the SL Ontology with knowledge from other legislative systems such as the Legislation.gov.uk project regarding the UK legislation and the GovTrack.us project regarding the USA legislation.

Finally, it should be considered to incorporate some lessons from the Greek Public Administration Ontology, such as the ability to describe the organization of the Public Administration. This could be very useful to provide a description and the relation of the entities that are in fact position-holders in the Government or Public Administration.

Also, concerning the choice of the triple store, other solutions should be tested in order to produce a complete benchmark and adopt the best solution. Considering the work Ma and others [Ma et al., 2006] and Rohloff and others [Rohloff et al., 2007], the solution used in this system would not be the optimal choice.

Regarding the linkage of the stored information with other data silos, there is still future work to accomplish. This system uses information from DBPedia. But, DBPedia contains more information than the information extracted in this implementation of the system. It could be interesting to use DBPedia and other silos available to increment the SL Ontology and provide a more specific classification of the resources on-the-fly.

Concerning the Web Service component, it mediates the information requests from third-parties providing responses that result from the access of the Database and the Index, depending on the nature of the request. The Web Service shares a limited set of resources, described formerly. However, the knowledge stored in the system provides more options in terms of user queries than the available ones. For example the search of legislation documents given an issuing entity, search of legislation documents given a data (day and/or month and/or year), and more.

8.3 Contributions

The main contribution of this project and dissertation is the creation of a five-star system according to the rating system for Linked Open Data of Tim Berners-Lee, concerning the legal domain, and specifically the Portuguese Legislation. This database was designed, implemented and tested with thousands of documents and entities and has proven to be capable to achieve the objectives set out. Furthermore, the access and reuse ability of the system is very important considering the rise of the Semantic Web, because it provides more information that is machine-readable and thus provides the ability to interlink information regarding this domain.

Concerning specific contributions, they are the proposed approach for extraction of structure, text and entities from PDF documents and the proposed ontology for the legal domain.

Regarding the former, it provides an answer to the issue of previous documentation in specific cases where the documents are stored. This approach provides the ability to extract information from these silos of usually proprietary format documents in order to pursue its translation to open data (eg., XML). For example, it can provide the ability to recover historical information stored in PDF collections.

Regarding the latter, the problem of the non-existence of a legal domain ontology standard is still an obstacle for further interoperability between different legislative processes and databases. The partial recognition of MetaLex as a standard is a step forward in the response to this issue, but still others questions remain, namely, the recognition and relation of entities in these legislative processes and produced documents. The SL Ontology provides a coarse-grained response to this issue by developing an ontology capable of linking documents and entities and also describing their respective information.

8.4 Endnotes

The development of this project started in August, 2011 and throughout the development phase and the writing of this document, submissions to international conferences were made. A submission to the World Wide Web 2012 Conference was made, which provided some reviews in terms of further work. This feedback was incorporated and presented in a submission to the Knowledge Discovery and Information Retrieval Conference (KDIR 2012) which was accepted and published in its

proceedings. Also, related to the latter, a communication was presented in Barcelona, in October 2012, in the referred conference.

Also, regarding future work of this project, two submissions are being prepared. The first concerning the system proposal and the second concerning the proposed ontology. Concerning the first submission, it will be focused on the proposal of a system that enables the translation of proprietary format document silos into open format documents as well as providing means to extract and store the documents information. Concerning the second submission, it will be focused on the use of the SL Ontology to gather information from various countries legislation documents and to prove its ability to interlink various documents and entities independently of country, language or jurisdiction.

These submissions will be prepared in the following months and should be finished by the end of 2012.

References

[Adobe, 2008] Adobe System Incorporated. Document management – Portable document format – Part 1: 1.7, 2008. http://www.adobe.com/content/dam/Adobe/en/devnet/acrobat/pdfs/PDF32000_2008.pdf [last access: Sep 2012]

[Ajani et al., 2007] Gianmaria Ajani, Leonardo Lesmo, Guido Boella et al., "Terminological and ontological analysis of european directives: multilinguism in law", in *Proceedings of the 11th international conference on Artificial Intelligence and law*, ICAIL '07, New York, NY, USA, pp. 43-48, ACM, 2007

[Antonacopoulos and Coenen, 1999] A. Antonacopoulos and F.P. Coenen, "Region Description and Comparative Analysis Using a Tesseral Representation", 5th International Conference on Document Analysis and Recognition, IEEE Computer Society Press, pp. 193-196, 1999

[Barabucci et al., 2009] Gioele Barabucci, Luca Cervone, Monica Parlmirani et al., "Multi-layer markup and ontological structures in Akoma Ntoso", in *Proceedings of the 2009 international conference on AI approaches to the complexity of legal systems: complex systems, the semantic web, ontologies, argumentation, and dialogue*, AICOL-I/IVR-XXIV'09, pp. 133-149, Berlin, Heidelberg, Springer-Verlag, 2010.

[Barabucci et al., 2012] Gioele Barabucci, Angelo Di Iorio, Francesco Poggi, "Bridging legal documents, external entities and heterogeneous KBs: from meta-model to implementation". In Semantic Web Journal, IOS Press, 2012 [awaiting decision]. http://www.semantic-web-journal.net/content/bridging-legal-documents-external-entities-and-heterogeneous-kbs-meta-model-implementation [last access: Sep 2012]

[Bechhofer and Miles, 2009] Sean Bechhofer and Alistair Miles, "SKOS Simples Knowledge Organization System Reference", Recommendation, W3C, 2009. http://www.w3.org/TR/2009/REC-skos-reference-20090818/. Latest version available at http://www.w3.org/TR/skos-reference [last access: Sep, 2012]

[Berners-Lee et al., 2001] Berners-Lee, T., Hendler J., Lassila O. The Semantic Web. In Scientific American Magazine, May 2001. http://campus.fsu.edu/bbcswebdav/users/bstvilia/lis5916metadata/readings/scientific-american_0.pdf [last access: Aug 2012]

[Berners-Lee, 2009] Tim Berners-Lee, Design Issues: Linked Data, 2009, W3C. http://www.w3.org/DesignIssues/LinkedData.html [last access: Aug 2012]

[Biagioli et al., 2003] C. Biagioli, E. Francesconi, P. Spinosa, and M. Taddei, "The NIR project: Standards and tools for legislative drafting and legal document web publication", in *Proceedings of ICAIL Workshop on e-Government: Modelling Norms and Concepts as Key Issues*, pp. 69-78, 2003

[Biasiotti et al., 2008] Mariangela Biasiotti, Enrico Francesconi, Monica Parlmirani et al., "Legal Informatics and Management of Legislative Documents", Giovanni Sartor ed., Global Centre for ICT in Parliament Working Paper no. 2, 2008

[Boer et al., 2002] Alexander Boer, R. J. Hoekstra and Radboud Winkels, "MetaLex: Legislation in XML", in *Proceedings of JURIX 2002: Legal Knowledge and Information System*, pp. 1–10, 2002

[Boer et al., 2007] Alexander Boer, Radboud Winkels and Fabio Vitali, "XML Standards for Law: MetaLex and LKIF", in *Proceedings of JURIX 2007*. Amsterdam, The Netherlands, IOS, 2007

[Boer et al., 2010] Alexander Boer, R. J. Hoekstra, E. De Maat et al., "Metalex (open XML interchange format for legal and legislative resources)", *Management Center*, 2010.

[Breuker et al., 2002] Joost Breuker, Abdullatif Elhag, Emil Petkov and Radboud Winkels, "Ontologies for the legal information serving and knowledge management", in *Legal Knowledge and Information Systems, Jurix 2002: The Fifteenth Annual Conference*, IOS Press, pp. 73-82, 2002

[Breuker and Hoekstra, 2004] Joost Breuker and Rinke Hoekstra, "Epistemology and ontology in core ontologies: FOLaw and LRI-CORE, two core ontologies for law", 2004

[Brickley and Miller, 2010] Dan Brickley and Libby Miller, "FOAF Vocabulary Specification 0.98", Namespace document, Marco Polo (ed.), 2010. http://xmlns.com/foaf/spec/ [last access: Sep 2012]

[Cappelli et al., 2007] Amedeo Cappelli, Valentina Bartalesi Lenzi, Rachele Sprugnoli and Carlo Biagioli, "Modelization of domain concepts extracted from the italian privacy legislation", in *Proceedings of Seventh International Workshop on Computational Semantics*, IWCS-7, pp. 305-308, 2007

[CGOV, 2012] CGOV: Central government ontology, an ontology of UK central government, 2012. http://lov.okfn.org/dataset/lov/details/vocabulary_cgov.html [last access: Sep 2012]

[CIW, 2012] China Internet Statistics Whitepaper, China Internet Watch, 2011. Available at http://www.chinainternetwatch.com/whitepaper/china-internet-statistics/ [last access: Sep 2012]

[DCMI, 2012] DCMI Usage Board, "DCMI Metadata Terms", DCMI Recommendation, 2012. http://dublincore.org/documents/dcmi-terms/ [last access: Sep 2012]

[Despress and Szulman, 2007] Sylvie Despress and Sylvie Szulman, "Merging of legal micro-ontologies from European directive", *Artifitial Intelligence Law*, 15(2), pp. 187-200, 2007

[EU, 2003] European Parliament and Council, DIRECTIVE 2003/98/EC on the re-use of public sector information. http://www.ec-gis.org/document.cfm?id=486&db=document [last access: Aug 2012]

[Fielding, 2000] Roy Thomas Fielding, "Architectural Styles and the Design of Network-based Software Architectures", Dissertation, University of California, Irvine, 2000

[Gangemi et al., 2003] Aldo Gangemi, Maria-Teresa Sagri and Daniela Tiscornia, "A constructive framework for legal ontologies", in Law and the Semantic Web, pp. 97-124, 2003

[Gruber, 1993] Thomas R. Gruber, "A translation approach to portable ontology specifications", Knowledge Acquisition, 5: pp. 199-220, 1993

[Gruber, 1995] Thomas R. Gruber, "Toward Principles for the Design of Ontologies User for Knowledge Sharing", International Journal of Human and Computer Studies, 43(5/6): pp. 907-928, 1995

[Guarino, 1997] Nicola Guarino, "Semantic Matching: Formal Ontological Distinctions for Information Organization, Extraction, and Integration", Information Extraction: A Multidisciplinary Approach to an Emerging Information Technology. Springer Verlag: pp. 139-170, 1997

[Guarino, 1998] Nicola Guarino, “Formal Ontology and Information Systems”, Proceedings of FOIS’98, Trento, Italy, IOS Press, pp. 3-15, 1998

[Guarino and Giaretta, 1995] Nicola Guarino and Pierdaniele Giaretta, “Ontologies and Knowledge Bases”, Towards Very Large Knowledge Bases, N.J.I. Mars (ed.), IOS Press, Amsterdam, 1995

[Giuffrida et al., 2000] Giovanni Giuffrida, Eddie C. Shek and Jihoon Yang, “Knowledge-Based Metadata Extraction from PostScript Files”, DL ’00, 2000

[Hassan, 2010] Tamir Hassan, “User-guided Information Extraction from Print-Oriented documents”, Vienna University of Technology, 2010

[Hassan and Baumgartner, 2005] Tamir Hassan and Robert Baumgartner, “Intelligent Text Extraction from PDF”, Database & Artificial Intelligence Group, Vienna University of Technology, Austria, 2005

[Hendler et al., 2002] James Hendler, Tim Berners-Lee and Eric Miller, "Integrating Applications on the Semantic Web" in Journal of the Institute of Electrical Engineers of Japan, Vol 122(10), pp. 676-680, 2002

[Hoekstra et al., 2009] Rinke Hoekstra, Joost Breuker, Marcello Di Bello and Alexander Boer, “LKIF-Core: Principled Ontology Development for the Legal Domain”, in Proceedings of the 2009 conference on Law, Ontologies and the Semantic Web: Channelling the Legal Information Flood, pp. 21-52, Amsterdam, The Netherlands, 2009, IOS Press, 2009

[Hollingsworth et al., 2005] Bill Hollingsworth, Ian Lewin and Dan Tidhar, “Retrieving Hierarchical Text Structure from Typeset Scientific Articles – a Prerequisite for E-science Text Mining”, University of Cambridge Computer Laboratory, 2005

[Hu et al., 2005] Yunhua Hu, Hang Li, Yunbo Cao et al., “Automatic Extraction of Titles from General Documents using Machine Learning”, Published in JCDL ’05, Denver, Colorado, USA, 2005

[Klink and Kieneger, 2001] Stefan Klink and Thomas Kieneger, “Rule-based Document Structure Understanding with a Fuzzy Combination of Layout and Textual Features”, German Research Center for Artificial Intelligence, 2001

[Klink et al., 2000] Stefan Klink, Andreas Dengel, Thomas Kieneger, “Document Structure Analysis Based on Layout and Textual Features”, DAS 2000: Proceedings of the International Workshop of Document Analysis Systems, 2000

[Laukyte et al., 2008] Migle Laukyte, Regis Riveret, Claudia Cevenini et al., “Development of the ALIS IP Ontology: Merging Legal and Technical Perspectives”, in *IFIP 20th World Computer Congress, Proceedings of the Second Topical Session on Computer-Aided Innovation*, volume 277 of *Computer-Aided Innovation (CAI)*, pp. 169-180, 2008

[Ma et al., 2006] Li Ma, Yang Yang, Zhaoming Qiu et al., “Towards a Complete OWL Ontology Benchmark”, in "The Semantic Web: Research and Applications", Lecture Notes in Computer Science, Springer Berlin/Heidelberg, pp. 125-139, Vol 4011, 2006

[Madison, 2000] Olivia M.A. Madison, “The IFLA Functional Requirements for Bibliographic Records: International standards for universal bibliographic control”,*Library Resources & Technical Services*, 44(3), pp. 153-159, 2000

[Masolo et al., 2004] Claudio Masolo, Laure Vieu, Emanuele Bottazzi et al., “Social roles and their descriptions”, pp. 267-277, AAAI Press, 2004

[mGCI, 2011] Mobile Government Consortium International, Introducing Mobile Government, Mar 2011. http://www.mgovernment.org/2011/03/introducing-mobile-government/ [last access: Aug 2012]

[Moniz and Rodrigues, 2012] Nuno Moniz, Fátima Rodrigues, “Extracting structure, text and entities from the Portuguese Legislation”, Knowledge Discovery and Information Retrieval 2012, Barcelona, Spain, 2012

[Mori et al., 1999] Shunji Mori, Hirobumi Nishida and Hiromitsu Yamada, "Optical Character Recognition" (1st edition), John Wiley & Sons, Inc. New York, NY, USA, 1999

[Muehlen et al., 2005] Michael zur Muehlen, Jeffrey V. Nickerson and Keith D. Swenson, “Developing Web Services Choreography Standards – The Case of REST vs. SOAP”, Decision Support Systems 37, Elsevier, North Holland, 2005

[Niyogi, 1994] Debashish Niyogi, “A Knowledge-Based Approach to Deriving Logical Structure from Document Images”, PhD thesis, State University of New York at Buffalo, 1994

[Noy and Musen, 1999] N. Fridman Noy and M. A. Musen, “An algorithm for merging and aligning ontologies: Automation and tool support”, in *Proceedings of the Workshop on Ontology Management at the Sixteenth National Conference on Artificial Intelligence* (AAAI-99), AAI Press, Orlando, USA, 1999

[OCD, 2011] Ontologia Camera dei Deputati, an ontology for the Italian Chamber of Deputies, 2011. http://lov.okfn.org/dataset/lov/details/vocabulary_ocd.html [last access: Sep 2012]

[ODa, 2005] Open Definition – Open Data, http://opendefinition.org/okd/ [last access: Aug 2012]

[ODb, 2005] Open Definition – Open Government, http://opendefinition.org/government/ [last access: Aug 2012]

[ONS, 2011] Internet Access – Households and Individuals, UK Office for National Statistics, 2011. http://www.ons.gov.uk/ons/dcp171778_227158.pdf [last access: Sep 2012]

[PARL, 2010] PARL: Parliament ontology, an ontology of UK parliament, 2010. http://lov.okfn.org/dataset/lov/details/vocabulary_parl.html [last access: Sep 2012]

[Powell et al., 2005] A. Powell, M. Nilsson, A. Naeve and P. Johnston, “Dublin core metadata initiative: Abstract model”, White Paper, 2005

[Rohloff et al., 2007] Kurt Rohloff, Mike Dean, Ian Emmons et al., “An Evaluation of Triple-Store Technologies for Large Data Stores”, in OTM’07 Proceedings of the 2007 OTM Confederated International Conference on “On the move to meaningful internet systems” (Vol 2), pp. 1105-1114, Springer-Verlag Berlin/Heidelberg, 2007

[Rosenfeld et al., 2002] Binyamin Rosenfeld, Ronen Feldman and Yonatan Aumann, “Structural Extraction from Visual Layout of Documents”, CIKM ’02, pp. 203-210, 2008

[Rubino et al., 2006] Rossella Rubino, Antonino Rotolo and Giovanni Sartor, “An OWL ontology of fundamental legal concepts”, in Proceedings of the 2006 conference on Legal Knowledge and Information Systems: JURIX 2006: The Nineteenth Annual Conference, pp. 101-110, Amsterdam, The Netherlands, IOS Press, 2006

[Russell, 2012] Mobile Internet to exceed PC access in India by the end of this year, Jon Russell, 2012. http://thenextweb.com/in/2012/05/03/mobile-internet-to-exceed-pc-access-in-india-by-the-end-of-this-year/ [last access: Sep 2012]

[Sagri and Tiscornia, 2003] Maria-Teresa Sagri and Daniela Tiscornia, “Metadata for content description in legal information”, in Proceedings of the 14th International Workshop on database and Expert Systems Applications, DEXA ’03, pp. 745-, Washington, DC, USA, IEEE Computer Society, 2003

[Sartor, 2007] Giovanni Sartor, “Legislation in the semantic web”, Law Department, European University Institute, Florence, University of Bologna, 2007 [PPT, last access: Oct 2011]

[Savvas and Bassiliades, 2009] Ioannis Savvas and Nick Bassiliades, “A process-oriented ontology-based knowledge management system for facilitating operational procedures in public administration”, Expert Syst. Appl., 36(3), pp. 4467-4478, 2009

[Shadbolt et al., 2006] Nigel Shadbolt, Wendy Hall, Tim Berners-Lee, "The Semantic Web Revisited", IEEE Intelligent Systems Journal, May/June 2006, 96-101.

[Shen et al., 2008] Yueting Shen, Robert Steele and John Murphy, “Building a Semantically Rich Legal Case Repository in OWL”, in *Proceedings of AusWeb08, the Fourteenth Australasian World Wide Web Conference*, AusWeb08, 2008

[Siefkes, 2003] Christian Siefkes, “Learning to Extract Information for the Semantic Web”, Berlin-Brandenburg Graduate School in Distributed Information System, Database and Information Systems Group, Freie Universität Berlin

[Sirin et al., 2007] Evren Sirin, Bijan Parsia, Bernardo Cuenca Grau et al., “Pellet: A Pratical OWL-DL Reasoner”, Web Semantic Science Services and Agents on the World Wide Web, 5(2), pp. 51-53, 2007

[Sowa, 2000] John F. Sowa, “Knowledge Representation: Logical, Philosophical and Computational Foundations”, Brooks/Cole, 2000

[Taylor, 1999] Arlene G. Taylor, “The Organization of Information”, Libraries Unlimited, 1999, 241-257

[Taylor et al., 1994] Suzanne L. Taylor, Deborah A. Dahl, Mark Lipshutz et al., “Integrated Text and Image Understanding for Document Understanding”, Unisys Corporation, 1994.

[Todoran et al., 2001] Leon Todoran, Marcel Worring, Marco Aiello and Christof Monz, “Document Understanding for a Broad Class of Documents”, ISIS technical report series, Vol. 2001-15, 2001

[UK Legislation, 2012] The official home of UK legislation. http://www.legislation.gov.uk/developer/formats/rdf [last access: Sep 2012]

[Vitali and Zeni, 2007] Fabio Vitali and Flavio Zeni, “Towards a country-independent data format: the Akoma Ntoso experience”, in *Proceedings of the V legislative XML Workshop*. http://www.europeanpress.eu/dlib/9788883980466/art5.pdf [last access: Sep 2012]

[Winkels et al., 2002] Radboud Winkels, Alexander Boer and Rinke Hoekstra, “CLIME: Lessons Learned in Legal Information Serving”, in Frank van Harmelen (ed.), *ECAI*, pp. 230-234, IOS Press, 2002

[W3C, 2001] World Wide Web Consortium, Web Services Description Language (WSDL) 1.1. http://www.w3.org/TR/wsdl [last access: Sep 2012]

[W3Ca, 2004] World Wide Web Consortium, RDF Test Cases. http://www.w3.org/TR/rdf-testcases/#ntriples [last access: Sep 2012]

[W3Cb, 2004] World Wide Web Consortium, Web Services Glossary. http://www.w3.org/TR/2004/NOTE-ws-gloss-20040211/ [last access: Sep 2012]

[W3Cc, 2004] World Wide Web Consortium, Web Services Architecture. http://www.w3.org/TR/ws-arch/#relwwwrest [last access: Sep 2012]

[W3C, 2008] World Wide Web Consortium, SPARQL Query Language for RDF. http://www.w3.org/TR/rdf-sparql-query/ [last access: Aug 2012]

[W3C, 2009] World Wide Web Consortium, Publishing Open Government Data. http://www.w3.org/TR/2009/WD-gov-data-20090908/ [last access: Aug 2012]

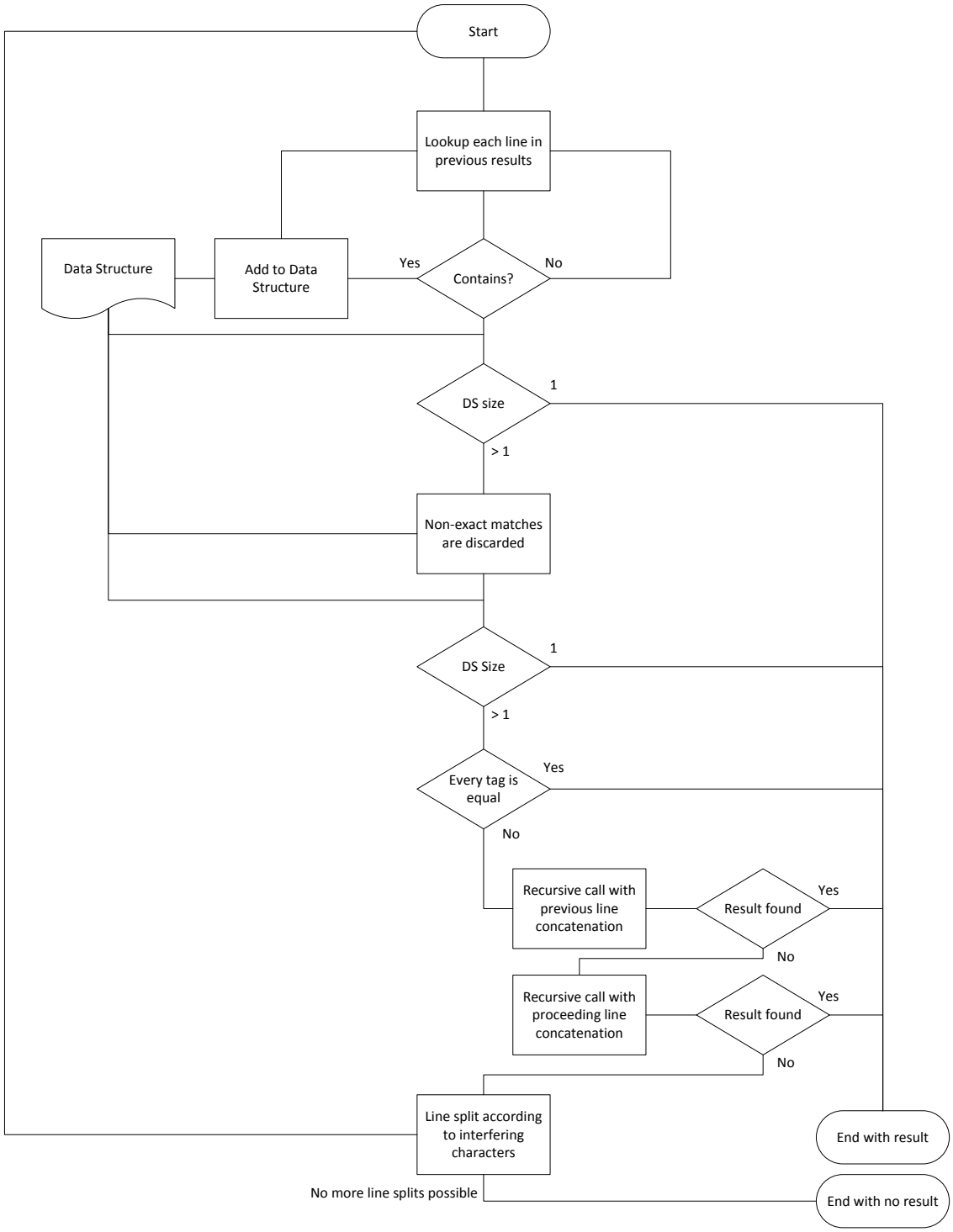
Attachment 1

Group of 40 first series Republic's Diary documents used in the study of Document Structure and Information Hierarchy. The date of the documents ranges from the 2nd of January 2009 to the 19th of March 2012.

Link	Date
http://dre.pt/pdfgratis/2009/01/00100.pdf	02/01/2009
http://dre.pt/pdfgratis/2009/02/02500.pdf	05/02/2009
http://dre.pt/pdfgratis/2009/03/04500.pdf	05/03/2009
http://dre.pt/pdfgratis/2009/04/06600.pdf	03/04/2009
http://dre.pt/pdfgratis/2009/05/08500.pdf	04/05/2009
http://dre.pt/pdfgratis/2009/06/10700.pdf	03/06/2009
http://dre.pt/pdfgratis/2009/07/12700.pdf	03/07/2009
http://dre.pt/pdfgratis/2009/08/14800.pdf	03/08/2009
http://dre.pt/pdfgratis/2009/09/17100.pdf	03/09/2009
http://dre.pt/pdfgratis/2009/10/19200.pdf	02/10/2009
http://dre.pt/pdfgratis/2009/11/21200.pdf	02/11/2009
http://dre.pt/pdfgratis/2009/12/23300.pdf	02/12/2009
http://dre.pt/pdfgratis/2010/01/00500.pdf	08/01/2010
http://dre.pt/pdfgratis/2010/02/02100.pdf	01/02/2010
http://dre.pt/pdfgratis/2010/03/04100.pdf	01/03/2010
http://dre.pt/pdfgratis/2010/04/06400.pdf	01/04/2010
http://dre.pt/pdfgratis/2010/05/08500.pdf	01/05/2010
http://dre.pt/pdfgratis/2010/06/10800.pdf	04/06/2010
http://dre.pt/pdfgratis/2010/07/12900.pdf	06/07/2010
http://dre.pt/pdfgratis/2010/08/15300.pdf	09/08/2010
http://dre.pt/pdfgratis/2010/09/17600.pdf	09/09/2010
http://dre.pt/pdfgratis/2010/10/19600.pdf	08/10/2010
http://dre.pt/pdfgratis/2010/11/21600.pdf	08/11/2010
http://dre.pt/pdfgratis/2010/12/23500.pdf	06/12/2010
http://dre.pt/pdfgratis/2011/01/00300.pdf	05/01/2011
http://dre.pt/pdfgratis/2011/02/02300.pdf	02/02/2011
http://dre.pt/pdfgratis/2011/03/04300.pdf	02/03/2011
http://dre.pt/pdfgratis/2011/04/06600.pdf	04/04/2011
http://dre.pt/pdfgratis/2011/05/08400.pdf	02/05/2011
http://dre.pt/pdfgratis/2011/06/10700.pdf	02/06/2011
http://dre.pt/pdfgratis/2011/07/13200.pdf	12/07/2011

http://dre.pt/pdfgratis/2011/08/15200.pdf	09/08/2011
http://dre.pt/pdfgratis/2011/09/17400.pdf	09/09/2011
http://dre.pt/pdfgratis/2011/10/20400.pdf	24/10/2011
http://dre.pt/pdfgratis/2011/11/22600.pdf	24/11/2011
http://dre.pt/pdfgratis/2011/12/24400.pdf	22/12/2011
http://dre.pt/pdfgratis/2012/01/01300.pdf	18/01/2012
http://dre.pt/pdfgratis/2012/02/03300.pdf	15/02/2012
http://dre.pt/pdfgratis/2012/03/05400.pdf	15/03/2012
http://dre.pt/pdfgratis/2012/03/05600.pdf	19/03/2012

Attachment 2



Attachment 3

- Document *lex:Document / frbr:Work*
    - Object Properties
        - lex:hasVersion / dc:hasVersion (lex:File)
        - lex:isPartOf /dc:isPartOf (lex:Container)
        - lex:hasPart / dc:hasPart (lex:Legislation)
        - dc:rightsHolder (lex:Entity)
        - lex:references / dc:references (lex:Entity, lex:Document)
        - lex:isReferencedBy / dc:isReferencedBy (lex:Legislation, lex:Container)
    - Data Properties
        - dc:identifier (literal)
        - dc:title (literal)
        - dc:alternative (literal)
        - dc:issued (date)
        - dc:available (date)
        - dc:link (literal)

- Container *lex:Container* (sub-class of Document)
    - Object Properties
        - lex:hasPart / dc:hasPart (lex:Legislation)
        - dc:rightsHolder (lex:Entity)
    - Data Properties
        - dc:identifier (literal)
        - dc:title (literal)
        - dc:alternative (literal)
        - dc:issued (date)
        - dc:available (date)
        - dc:link (literal)

- Legislation lex:Legislation (sub-class of Document)
    - Object Properties
        - lex:hasVersion / dc:hasVersion (lex:File)
        - dc:author (Lex:Entity)
        - lex:isPartOf / dc:isPartOf (lex:Container)
        - dc:rightsHolder (lex:Entity)
        - lex:references / dc:references (lex:Entity, lex:Document)
        - lex:isReferencedBy / dc:isReferencedBy (lex:Legislation, lex:Container)
    - Data Properties
        - dc:identifier (literal)
        - dc:title (literal)
        - dc:alternative (literal)

        - dc:issued (date)
        - dc:available (date)
        - dc:link (literal)

- File *lex:File* / *frbr:Expression*
    - Object Properties
        - lex:isVersionOf (lex:Document)
    - Data Properties
        - dc:issued (date)
        - dc:available (date)
        - dc:source (literal)
        - lex:link (literal)

- Entity *lex:Entity* / *foaf:Agent*
    - Object Properties
        - lex:isReferencedBy / dc: isReferencedBy (lex:Document)
        - foaf:made (lex:Legislation)
    - Data Properties
        - foaf:name (literal)
        - dc:identifier (literal)

Attachment 4

Republic Diary	PDFSize (kb)	TotalTime (s)	TextLinesFound	Accuracy (%)	EntitiesFound	EntitiesAccuracy (%)	XMLSize (kb)	Pages	Link
Jan-09	700	11	1536	99,09%	256	96,88%	74	16	http://dre.pt/pdfgratis/2009/01/00100.pdf
Fev-09	889	17	2489	99,84%	90	93,33%	105	26	http://dre.pt/pdfgratis/2009/02/02500.pdf
Mar-09	400	6	427	98,13%	64	90,63%	17	6	http://dre.pt/pdfgratis/2009/03/04500.pdf
Abr-09	734	25	3782	99,95%	240	93,33%	164	48	http://dre.pt/pdfgratis/2009/04/06600.pdf
Mai-09	370	8	1463	99,11%	125	91,20%	78	14	http://dre.pt/pdfgratis/2009/05/08500.pdf
Jun-09	1352	16	2630	99,66%	203	93,60%	162	28	http://dre.pt/pdfgratis/2009/06/10700.pdf
Jul-09	1583	10	1569	99,87%	76	97,37%	74	24	http://dre.pt/pdfgratis/2009/07/12700.pdf
Ago-09	794	41	5550	99,59%	247	95,14%	257	52	http://dre.pt/pdfgratis/2009/08/14800.pdf
Set-09	495	18	3392	99,82%	138	89,86%	157	32	http://dre.pt/pdfgratis/2009/09/17100.pdf
Out-09	1722	293	16704	99,85%	431	93,50%	874	152	http://dre.pt/pdfgratis/2009/10/19200.pdf
Nov-09	346	7	1151	99,83%	80	95,00%	53	12	http://dre.pt/pdfgratis/2009/11/21200.pdf
Dez-09	411	15	2232	99,96%	40	97,50%	109	22	http://dre.pt/pdfgratis/2009/12/23300.pdf
Jan-10	1036	127	10167	98,80%	1331	65,74%	290	104	http://dre.pt/pdfgratis/2010/01/00500.pdf
Fev-10	517	4	215	99,53%	56	92,86%	10	4	http://dre.pt/pdfgratis/2010/02/02100.pdf
Mar-10	587	11	2278	99,39%	129	92,25%	166	24	http://dre.pt/pdfgratis/2010/03/04100.pdf
Abr-10	500	12	2201	99,86%	134	82,84%	114	20	http://dre.pt/pdfgratis/2010/04/06400.pdf
Mai-10	782	28	4846	99,07%	128	89,06%	182	78	http://dre.pt/pdfgratis/2010/05/08500.pdf
Jun-10	1443	10	1738	99,48%	106	90,57%	138	20	http://dre.pt/pdfgratis/2010/06/10800.pdf
Jul-10	1302	5	507	99,21%	77	90,91%	31	8	http://dre.pt/pdfgratis/2010/07/12900.pdf
Ago-10	3990	13	2406	99,79%	147	96,60%	96	48	http://dre.pt/pdfgratis/2010/08/15300.pdf
Set-10	693	9	1708	99,94%	128	95,31%	90	20	http://dre.pt/pdfgratis/2010/09/17600.pdf
Out-10	425	5	619	99,52%	60	96,67%	31	8	http://dre.pt/pdfgratis/2010/10/19600.pdf
Nov-10	585	36	4919	99,84%	204	93,63%	260	42	http://dre.pt/pdfgratis/2010/11/21600.pdf
Dez-10	1029	13	2416	99,88%	161	93,79%	116	26	http://dre.pt/pdfgratis/2010/12/23500.pdf
Jan-11	963	138	9748	99,57%	1137	91,91%	307	86	http://dre.pt/pdfgratis/2011/01/00300.pdf
Fev-11	639	9	1331	99,77%	98	94,90%	85	16	http://dre.pt/pdfgratis/2011/02/02300.pdf
Mar-11	9889	17	2470	99,88%	220	89,09%	102	32	http://dre.pt/pdfgratis/2011/03/04300.pdf
Abr-11	1461	25	4182	99,67%	196	92,35%	203	42	http://dre.pt/pdfgratis/2011/04/06600.pdf
Mai-11	2558	12	2134	99,67%	163	92,64%	97	26	http://dre.pt/pdfgratis/2011/05/08400.pdf
Jun-11	1580	7	1205	99,92%	85	95,29%	54	14	http://dre.pt/pdfgratis/2011/06/10700.pdf
Jul-11	1125	9	683	99,56%	84	95,24%	34	12	http://dre.pt/pdfgratis/2011/07/13200.pdf
Ago-11	332	5	626	99,68%	52	92,31%	26	8	http://dre.pt/pdfgratis/2011/08/15200.pdf
Set-11	14534	9	131	93,13%	19	94,74%	4	6	http://dre.pt/pdfgratis/2011/09/17400.pdf
Out-11	456	15	2117	99,67%	93	96,77%	101	20	http://dre.pt/pdfgratis/2011/10/20400.pdf
Nov-11	368	11	1879	100,00%	33	93,94%	93	18	http://dre.pt/pdfgratis/2011/11/22600.pdf
Dez-11	471	13	2529	99,84%	113	94,69%	211	24	http://dre.pt/pdfgratis/2011/12/24400.pdf
Jan-12	403	19	2592	99,92%	83	95,18%	125	24	http://dre.pt/pdfgratis/2012/01/01300.pdf
Fev-12	407	7	1303	100,00%	115	88,70%	65	14	http://dre.pt/pdfgratis/2012/02/03300.pdf
Mar-12	545	27	3214	99,19%	128	94,53%	133	50	http://dre.pt/pdfgratis/2012/03/05400.pdf
Mar-12	314	5	368	100,00%	35	97,14%	17	6	http://dre.pt/pdfgratis/2012/03/05600.pdf

Attachment 5

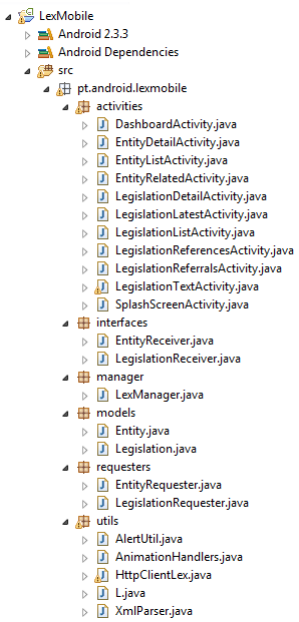


Figure 69 – Mobile Application Project Tree