Available online at www.sciencedirect.com

SciVerse ScienceDirect

Procedia
Social and Behavioral Sciences

Procedia - Social and Behavioral Sciences 73 (2013) 479 – 486

The 2nd International Conference on Integrated Information

Applying Linked Data Technologies to Greek Open Government Data: A Case Study

Eleni Galiotou[a],*, Pavlina Fragkou[a]

[a]Dept. of Informatics, Technological Educational Institute of Athens, Ag. Spyridona, Egaleo, GR-122 10 Greece

Abstract

Open government data is a valuable resource of information addressed to a significant number of recipients. However, this information is usually published in raw format, i.e. without following specific guidelines and remains unexploited. Linked data technologies, on the other hand, aim at transforming data published in web sites into a machine readable format (usually RDF using URIs) in order for them to be linked to other external datasets. Regarding Greek Governmental sites, little work has been done towards this direction. An interesting case is the information provided in the ERMIS Greek portal for Public Administration which involves service provision according to the Directive 123/2006/EC. In this paper, we present a case study on the application of linked data technologies on Greek open government data located in the ERMIS Greek Government portal for Public Administration (www.ermis.gov.gr). In particular, we focus on examining how this information can be transformed into linked data in order to be appropriately interconnected to equivalent information found in web sites of other European countries and we propose solutions.

Keywords: *open government data, linked data, Greek one stop shop governmental portal, directive 123/2006/EC.*

1. Introduction

The vast amount of information available on the Web makes the problem of searching related information time consuming. Even though several techniques for making search engines more efficient were developed, for example by providing similar pages, the task still remains problematic. Semantic Web aims at solving this by

* Corresponding author. Tel.: +30-210-5385824; fax: +30-210-5910975.
E-mail address: egali@teiath.gr

doi:10.1016/j.sbspro.2013.02.080

relating information found in different web sites in a unified way. To accomplish this, a number of satellite technologies were developed. The most promising one is perhaps that of the Linked Data paradigm [1].

The Linked Data paradigm [1] involves practices to publish, share, and connect data on the Web, and offers a new way of data integration and interoperability. Briefly, Linked Data is about using the Web to create links between data from different sources. The driving force to implement Linked Data spaces is the RDF technology. The basic principles of the Linked Data paradigm are: (a) use the RDF data model to publish structured data on the Web, and (b) use RDF links to interlink data from different data sources. The aim of the Linked Data technologies is to give rise to the Web of Data which is impelled by the current trend towards an open Web. The Open Data movement is a significant and emerging force towards this direction.

Open public sector data are open data related to useful information to citizens and enterprises for their transactions with the public sector. Obviously, they are publicly available to anyone to analyze and reuse. However, even though public sector information is open, it is organized and published in a chaotic way. This means that, the same information not only can be found in different web sites but a link between those sites is missing. Even worse, links to related information residing in the same website do not exist. The aforementioned situation creates the need to convert public sector data to Linked Open Data (LOD), in order to provide the minimum requirements of data linkage and re-use.

In this work, we present a case study to create Linked Open Data (LOD) by focusing on data hosted in the ERMIS Greek Government portal for Public Administration [2] and most specifically those focusing on service provision according to Directive 123/2006/EC. The directive in question imposes on each member state to provide useful information regarding service provision in two languages, the official of the member state and English. However, no linkage exists between relevant information inside each website, let alone linkage between websites supported by each member state. The latter would potentially provide to the end user the possibility to compare which prerequisite supporting documents are needed in order to exert the same profession or to obtain the same service activity in different countries. In this way, the user would able to choose the country in which such an undertaking is easier.

Our paper is organized as follows: Section 2 presents related work in the field by focusing on work performed on Greek Public Sector websites. Section 3 analyzes our problem in detail, Section 4 provides an overview of the necessary steps to transform data appearing in plain HTML web pages into Linked Open Data, Section 5 provides a description of the proposed methodology to be followed while Section 6 describes future work.

2. Related Work

Linked Open Data is a relatively new research area with a great potential thus deserving special attention. Moreover, as far as Greek open data are concerned, little work appears in the literature regarding their transformation into Linked Data. To our knowledge, three attempts have been made. The first work [3] suggests a classification scheme for open government data (OGD) and analyzes and classifies OGD initiatives based on it. In addition, it presents a case study of linking data between three related public sector websites: a school website (namely that of the Moraitis school in Athens), the 2nd Local Directorate of Secondary Education of Athens website and the Ministry of Education website. However, this case study has limited scope since it is restricted to a small number of web pages.

The second and closest one to our problem is the work presented in [4]. In this work, a publicly accessible Web point was constructed aiming to promote clarity and enhance citizen awareness regarding public spending in Greece through easily consumed visualization diagrams. Information provision is based on semantic processing of real-time open data provided by the Greek government through the “Diavgeia” program [5] and the Greek Taxation Information System. Web pages are downloaded from those sites, checked for their validity and are semantically enriched with concepts (resource and property URIs) using a manually constructed ontology for this purpose. The final objective of this work is data interconnection to existing ontological and data schemes derived from other similar initiatives worldwide and core vocabularies.

The third attempt aiming at constructing Greek - Linked Open Data and their internationalization, is the one appearing in [6]. This work presents the first steps towards a Greek Linked Open Data (LOD) cloud, initially as a collection of exposed interlinked datasets and a Greek DBpedia core hub. In the course of the project and while forming and enriching the cloud, the authors addressed the wider issue of non-latin language characters both in resource naming and in SPARQL queries. They proposed a method for resolving that issue which is applicable to all languages with a non-latin alphabet.

3. The Problem

The problem examined here, involves public data hosted in the ERMIS Greek portal for Public Administration [2]) and most specifically information related to the Directive 123/2006/EC [7]. The directive in question focuses on simplifying the procedure of practicing a profession by a European citizen in another member state. The implementation of this Directive requires from each member state the provision of a dedicated portal named Point of Single Contact (PSC). This portal must contain information regarding the required supporting documents for each service activity (for example Operation License to Tour Guide's) as well as all related information in two languages i.e. the one of the member state and English. A central portal [8] provides a pointer to the specific ones supported by each member state.

Each service activity is categorized into one or more categories which follow the NACE basis [10]. The European Classification of Economics Activities (NACE) is the European reference framework for the production and the dissemination of statistics related to economic activities. NACE is an important tool for comparing statistical data related to economic activities at a world level. Table 1 provides a list of service activities appearing in the Greek PSC [9] classified according to the NACE basis.

Each service activity contains the following information:

- The service activity's Name
- The service activity's description
- The Public Sector responsible for the legal framework of the service activity.
- The Public Sector responsible for receiving and expediting the authorization of the service activity.
- The service activity Type, i.e. permit, license, certificate, etc.
- The Life Event, such as Starting a Business, Getting Insured, Studying etc.
- The Legal Rule, i.e. all related laws.
- All prerequisites that may restrict some categories of applicants.
- The service activity Cost
- The delivery Time i.e. the required time for the provision of the authorization decision.
- Comments, which may include any information which may prove to be useful for the applicant such as a telephone and/or an e-mail for help or complaints.
- A list of related keywords
- The investment environment, e.g. whether this service activity is financed by any investment program.
- Whether physical presence is required for the submission of the application.
- Whether physical presence is required when receiving the result of the application.
- The description of how the application is expedited by the involved public sector(s) by paying special attention to the factors which may lead to its rejection.
- All required supporting documents.
- Bank account, in case where a service activity cost is required and can be paid in a bank.
- NACE code(s) i.e. the thematic category in which the service activity can be categorized.

Table 1. List of service activities in the Greek PSC

Category	Service activities in English	Service activities in Greek
A- Agriculture, forestry and fishing	4	19
B- Mining and quarrying	0	15
C – Manufacturing	0	72
E - Water supply; sewerage; waste management and remediation activities	1	6
F - Construction	5	94
G- Wholesale and retail trade; repair of motor vehicles and motorcycles	22	63
H- Transporting and storage	5	47
I- Accommodation and food service activities	2	10
J- Information and communication	0	1
K - Financial and insurance activities	0	0
L- Real estate activities	1	3
M- Professional, scientific and technical activities	29	72
N- Administrative and support service activities	16	40
P - Education	33	66
R - Arts, entertainment and recreation	15	29
S - Other services activities	1	22
Total	**134**	**659**

The aforementioned content that was chosen in order to describe each service activity, was based on the specific e-GIF ontology. The e-GIF ontology is part of the Greek e-Government Interoperability Framework [11], a survey containing a list of rules for the provision of e-Government services to public bodies, businesses and citizens in a unified manner, addressed to all Greek Public Administration. The Electronic Government framework aims at supporting effectively e-Government at a central, regional and local level and contributes to achieving interoperability at the level of information systems, procedures and data.

However, the way that service activity information is organized in the Greek PSC provides no interconnection between related service activities. For example, no link exists between two service activities A and B where A is a prerequisite for B. Additionally, no link exists between service activity A and its translation in English. This is due to the fact that, no *semantic* linkage between related service activities exists. Therefore, a need for transforming this case of Open Government Data (OGD) to Linked Open Data (LOD) has emerged.

3.1. Open Government Data (OGD)

OGD initiatives emerged only recently and as a result, there is a lack of classification schemes to analyze them. There is, however, an increasing number of practical guidelines suggested by various stakeholders. The World Wide Web Consortium (W3C) e-Government Interest Group suggest three steps for public administrations to open and share their data [12]:

- firstly, publish data in raw form by means of files in well-known and non-proprietary formats such as CSV and XML
- next, create online catalogues of the raw data
- finally, make the data machine-readable.

3.2. Linked Open Data (LOD)

Linked data seem to play a prevalent role in the future of OGD initiatives. The term "linked data" refers to *"data published on the web in such a way that it is machine-readable, its meaning is explicitly defined, it is linked to other external datasets, and can in turn be linked to from external datasets"* [13]. The Linked Data initiative is based on the philosophy and technologies of the Semantic Web. Yet, in contrast to the full-fledged semantic web vision, it is mainly about publishing structured data in RDF using URIs rather than focusing on the ontological level or inferencing [14]. It promises the creation of the "Web of Data" as data from decentralised and heterogeneous sources can be interlinked through typed links. Web of Data aims at replacing isolated data islands by a giant distributed dataset built on top of the web architecture.

The Linked Data approach requires the identification of resources with URI references that can be de-referenced over the HTTP protocol into RDF data describing the identified resource. In addition, Linked Data include the creation of typed links between URI references, so that one can discover more data. More specifically, the four linked data principles are the following [1]:

- all items should be identified using URIs (instead of blank nodes, to which it is not possible to create RDF links because they are limited to the document in which they appear).
- all URIs should be de-referenceable, that is, using HTTP URIs should allow looking up the item identified through the URI
- when looking up a URI it leads to more data, which is usually referred to as the "follow your nose" principle
- links to other URIs should be included in order to enable the discovery of more data.

Linked data distinguish between informational and non-informational resources [15]. The former refers to all the resources we find on the traditional document web such as documents, images etc, while the latter refers to real world things such as people, schools, laws, public agencies etc. The adoption of identifiers ensures the unique identification of information resources in the web but not of the real world things to which the information resources refer. Hence a central issue in the web of data is finding identifiers that refer to similar real world things. These identifiers became known as URI aliases [15]. A very useful guide to publishing Linked Data on the Web can be found in [15].

After identifying common resources inside data sources to connect data contained in each of them, applications that can rely on shared reuse of URIs (which identify a resource directly) and on an ontology's inverse functional properties (which uniquely identify a resource indirectly) can be developed. However, performing the required ontology reasoning to find the inverse functional properties in a web site can be computationally expensive and prevent efficient indexing. To address both challenges in discovering decentralized Semantic Web data, an architectural element is required which (a) creates the missing links and (b) allows Semantic Web clients to find and integrate independent sub-graphs each of which belonging to a unique web site, into a single, virtual, large graph.

4. Proposed methodology

Having taken under consideration the analysis of the problem as this was presented in Section 2, we propose the following components, each of which is described in a separated subsection.

4.1 Web page retrieval and storage

As it was presented in Section 2, web pages are hosted in a dedicated site of the ERMIS Greek Government portal for Public Administration [9]. Even though information hosted in the portal is publicly available, no interventions, i.e. addition of RDF tags and annotations, are permitted. This leads us to the necessity to download a copy for each page, as well as a mechanism that will periodically look for changes in the content of each page. A database will also be created which will host all information related to each service activity. This information will be hosted in a dedicated site that will be provided by the Technological Educational Institute (TEI) of Athens.

4.2 Ontology usage- Transformation to semantic concepts

Since our purpose is to enrich information describing each service activity with semantic information in the form of RDF links, an ontology is necessary to depict the basic concepts with their attributes as well as the relations between them. In section 2 we mentioned that, the e-GIF ontology [11] was used by ERMIS as a basis for organizing information. However, a detailed examination of this ontology led us to the conclusion that a number of omissions exist. Among those, the most important ones are the lack of relations referring to our case as well as the lack of some concept properties. This obliges us to enrich the ontology with all necessary concepts, properties and relations in order to depict the current status of semantic interconnections. The updated ontology will serve us as guideline in order to create all links via RDF annotations. Examples of such links are: links between a service activity in Greek and its translation in English, links involving the Public Sector responsible for the legal framework of the service activity, the Public Sector responsible for reception and expedition of the authorization of the service activity, the service activity type, each type of supporting document as well as the NACE code(s) i.e. the thematic category in which the service activity can be categorized. Once the appropriate RDF annotation is performed in each web page, linkage will be achieved i.e. the user will be able to navigate to relevant information.

A subsequent step will involve the semantic enrichment of the higher elements of web pages with concepts (resource and property URIs) of the updated e-GIF ontology. This can be accomplished by using tools such as the Jena Framework [16]. The next phase involves the instantiation of all RDF related triples (i.e. the actual data) by implementing appropriate tools or using the D2RQ declarative mapping language [18] that captures mappings between database schemas and RDFS/OWL schemas. The produced RDF triples will be stored in an LOD Server component which can be thought as a special kind of database, capable of storing, inferencing and querying RDF or OWL data. In order to publish the content of our database on the Semantic Web we intend to use well known tools such as Open Link Virtuoso [17], or D2R [18]. The D2R SPARQL end point facilities or the corresponding ones of the Open Link Virtuoso Environment will also be incorporated in order to enable users to formulate SPARQL queries for data retrieval. A graphical representation of our methodology appears in Figure 1.

4.3 Examination of potential interconnections with other datasets.

Once we have ensured that interconnection via linkage is achieved for all web pages of the Greek Government portal for Public Administration related to Directive 123/2006/EC [7], we plan to examine their interconnection with information found in the dedicated sites of other countries. Recall that, this can be relatively easily accomplished since each country is obliged to provide service activity information in English thus, linkage between different sites can be performed via their pages written in English. Once this is succeed, a European citizen aiming to exert a profession i.e. a service activity and wants to choose the country that requires the least supporting documents will be able to easily navigate from a country's website to another in order to find the desired target country.

The aforementioned task requires a common ontology to be used as a bridge. A candidate ontology is DBpedia [19]. According to its description [19], "*DBpedia is a community effort to extract structured information from Wikipedia and to make this information available on the Web. DBpedia allows you to ask sophisticated queries against Wikipedia, and to link other data sets on the Web to Wikipedia data*". A graphical representation of DBpedia ontology can be found in [20]. As it can be easily seen, DBpedia ontology does not contain concepts and relations concerning public sector, service activities and supporting documents. However, DBpedia benefits from the fact that it allows the user to enrich it with concepts and classes. Thus, an in depth examination of all candidate websites of all involved countries is required in order to analyze their content. As a sequel, a taxonomy of concepts and relations will be created in order to depict the fusion of all semantic information contained in those websites. This taxonomy can be provided to DBpedia as a compement not only to serve our purposes but to be used as a starting point for other purposes as well. Another potential connection that will be examined is with the Greek DBPedia [21] the implementation of which is on-going.

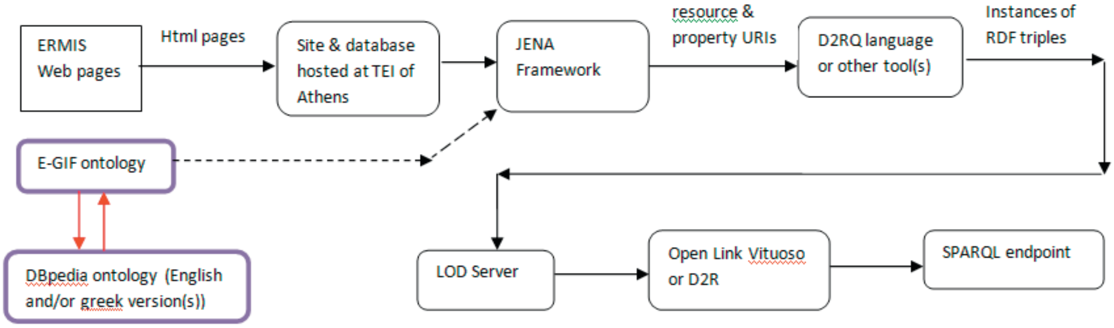

Fig. 1. Proposed Methodology Schema

All the aforementioned work involves the manual identification of interlinking with datasets via the manual addition of the links (such as OWL:sameAs, RDFS:seeAlso). The Silk framework [22] which can automatically discover and propose such links will also be examined.

4.4 Statistical Results.

RDF links resulting from actions described in Section 4.2 will also be used for other purposes. Among those, is the extraction of statistical results like: a) how may service activities are put through by Public Sector X b) in which service activities a certain supporting document A is required c) in how many service activities both supporting documents A and B are required d) which service activity requires the minimum/maximum number of supporting documents e) to how many service activities, service activity A is a prerequisite. Those statistical results can be given as input to public authorities in order for them to obtain an overview of the difficulty of exerting certain service activities thus bringing to light the need to perform administrative simplification actions. Public authorities may also benefit from information taken from websites from other countries once appropriate implementation of actions described in Section 4.3 are performed. Undoubtedly, administrative simplification actions will be beneficial for making our country more appealing for business reinforcement.

5. Conclusions – Future Work

In this paper we have presented a case study concerning the transformation of publicly available information regarding service activity provision enforced by Directive 123/2006/EC [7] to Linked Open Data. The problem studied resulted in the projection of a number of issues - needs: a) download web pages into a local copy b) enrichment of the e-GIF ontology in order to cover all types of concepts and relations to accurately describe web pages c) transform web pages into RDF using the e-GIF ontology d) choice of appropriate tools to perform all the aforementioned efficiently in order to produce Linked Open Data.

Our vision consists of two steps. The first step aims to create appropriate links i.e. LOD between pages belonging to our core dataset, taking under consideration that for a number of pages their translated version in English web page exists. The second step involves the interconnection of our dataset with the equivalent ones found in the dedicated sites of other European countries i.e. the creation of interconnection links between sites focusing on the English content of each site. In order to perform this, the DBpedia ontology [20] can be considered as common ontology to use. To this end, a potential enrichment of the DBpedia ontology with appropriate classes and relations will also be investigated.

Acknowledgements

This research has been co-funded by the European Union (Social Fund) and Greek national resources under the framework of the “Archimedes III: Funding of Research Groups in the TEI of Athens” project of the “Education & Lifelong Learning” Operational Programme.

References

[1] Linked Data Paradigm: http://linkeddata.org/.
[2] Ermis portal: www.ermis.gov.gr
[3] Kalampokis, E., Tambouris, E., & Tarabanis, K. (2011) A Classification Scheme for Open Government Data: Towards Linking Decentralized Data, International Journal of Web Engineering and Technology, 6(3), 266-285.
[4] Vafopoulos, M., Meimaris, M., Papantoniou, A., Anagnostopoulos,I., Alexiou, G., Avraam, I., Xidias, I., Vafeiadis, G. & Loumos, V. (2012). Publicspending.gr: interconnecting and visualizing Greek public expenditure following Linked Open Data directives. USING OPEN DATA: policy modeling, citizen empowerment, data journalism, The European Commission's Albert Borschette Conference Center, Brussels.
[5]"Diavgeia" http://opendata.diavgeia.gov.gr
[6] Bratsas, C., Alexiou, S., Kontokostas, D., Parapontis, I., Antoniou, I. & Metakides G. (2011). Greek Open Data in the Age of Linked Data: A Demonstration of LOD Internationalization. In Proceedings of the ACM WebSci'11, 1-4.
[7] Directive 2006/123/EC. http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2006:114:0009:0021:en:PDF
[8] EUGO, http://ec.europa.eu/internal_market/eu-go/index_en.htm
[9] The Greek PSC: http://gis.ermis.gov.gr/sdportal/index.jsp
[10] Nace codes: http://www.geodirectory.ie/Downloads-(1)/NACE-Rev-2.aspx
[11] http://www.e-gif.gov.gr/portal/page/portal/egif
[12] http://www.w3.org/TR/gov-data/
[13] Bizer, C., Heath, T. & Berners-Lee, T.(2009).Linked Data - The Story So Far. edited by: T. Heath, M. Hepp, C. Bizer International Journal on Semantic Web and Information Systems (IJSWIS), 5 (3), 1-22.
[14] Hausenblas, M. (2009). Exploiting Linked Data to Build Web Applications. IEEE Internet Computing, 13(4), 68-73.
[15]Bizer, C., Cyganiak, R. & Heath, T. How to Publish Linked Data on the Web.
http://www4.wiwiss.fu- berlin.de/bizer/pub/linkeddatatutorial/
[16] Jena Framework: http://jena.apache.org/
[17] http://virtuoso.openlinksw.com
[1] http://www4.wiwiss.fu-berlin.de/bizer/d2r-server/
[19]DBpedia: http://dbpedia.org/About
[20] DBpedia ontology: http://www4.wiwiss.fu-berlin.de/dbpedia/dev/ontology.htm.
[21] Kontokostas, D., Bratsas, C., Auer, S., Hellmann, S., Antoniou, I. & Metakides, G. (2011). Towards Linked Data Internationalization - Realizing the Greek DBpedia. In Proceedings of the ACM WebSci'11, 1-4.
[22] Jentzsch, A., Isele, R. & Bizer, C. (2010) Silk - Generating RDF Links while publishing or consuming Linked Data. In Poster at the International Semantic Web Conference (ISWC2010).