Greek Open Data in the Age of Linked Data: A Demonstration of LOD Internationalization

Charalampos Bratsas
Web Science Program
Mathematical Department
Aristotle University of Thessaloniki
http://webscience.auth.gr
cbratsas@math.auth.gr

Spyros Alexiou
Web Science Program
Mathematical Department
Aristotle University of Thessaloniki
http://webscience.auth.gr
szalexio@math.auth.gr

Dimitris Kontokostas
Web Science Program
Mathematical Department
Aristotle University of Thessaloniki
http://webscience.auth.gr
jimkont@math.auth.gr

Ioannis Parapontis
Web Science Program
Mathematical Department
Aristotle University of Thessaloniki
http://webscience.auth.gr
paraponi@math.auth.gr

Ioannis Antoniou
Web Science Program
Mathematical Department
Aristotle University of Thessaloniki
http://webscience.auth.gr
iantonio@math.auth.gr

George Metakides
Web Science Program
Mathematical Department
Aristotle University of Thessaloniki
http://webscience.auth.gr
george@metakides.net

ABSTRACT

This paper 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. It is a joint effort to become part of the wider, global linked data cloud and aims to contribute in the overall cloud informational value. During the project and while forming and enriching the cloud we actually addressed effectively the wider issue of non-Latin language characters both in resource naming and in SPARQL queries. We propose a method to resolve that issue which is applicable to all Languages with Non-Latin characters

Categories and Subject Descriptors

I.2.4 [**Artificial Intelligence**]: Knowledge Representation Formalisms and Methods; E.2 [**Data**]: Data Storage Representations—*Linked Representations*

General Terms

Management, Languages

Keywords

Linked Open Data, I18n LOD, Internationalization, Semantic Web, International Resource Identifiers

1. INTRODUCTION

Linked Data have been recognized as one of the first success stories of the emerging Semantic Web and the cloud they reside is growing with tremendous rate. Scientists & governments around the world already evaluated the potential and benefits that linked data based applications and services might bring and this drives a common, concerted effort to contribute in this direction [1], [3].

This work contributes towards the internationalization of linked open data and results in a wave of applications and services, with Greek Texts being the first application. The exploited open data are initially made available by public services and local authorities and are then exposed as interlinked datasets of the Greek LOD cloud.

Following the open access to public data initiative, the Greek Government conceptualized and directed its Geodata Gov project (`http://geodata.gov.gr`) into gathering non personal data from all public services and local authorities. Additionally and anticipating the educational value of the Greek Wikipedia, the Government decided to promote article authoring in schools, universities and everyday users with a common goal of doubling the article count within a year.

As far as the Greek linked data are concerned, this work accounts for the first step towards a Greek LOD cloud, as a collection of exposed interlinked datasets and a Greek dbpedia core hub. These steps also include resolving a so far common issue in resource naming and in SPARQL queries when contributing in non-Latin language characters.

Finally, a number of demos are developed to satisfy our main, twofold aim. Firstly to demonstrate the approaches described concerning the formation of the Greek LOD cloud and secondly to present our proposed solution in overcoming the issue of Non-Latin characters so as to open the way for applications in all other languages.

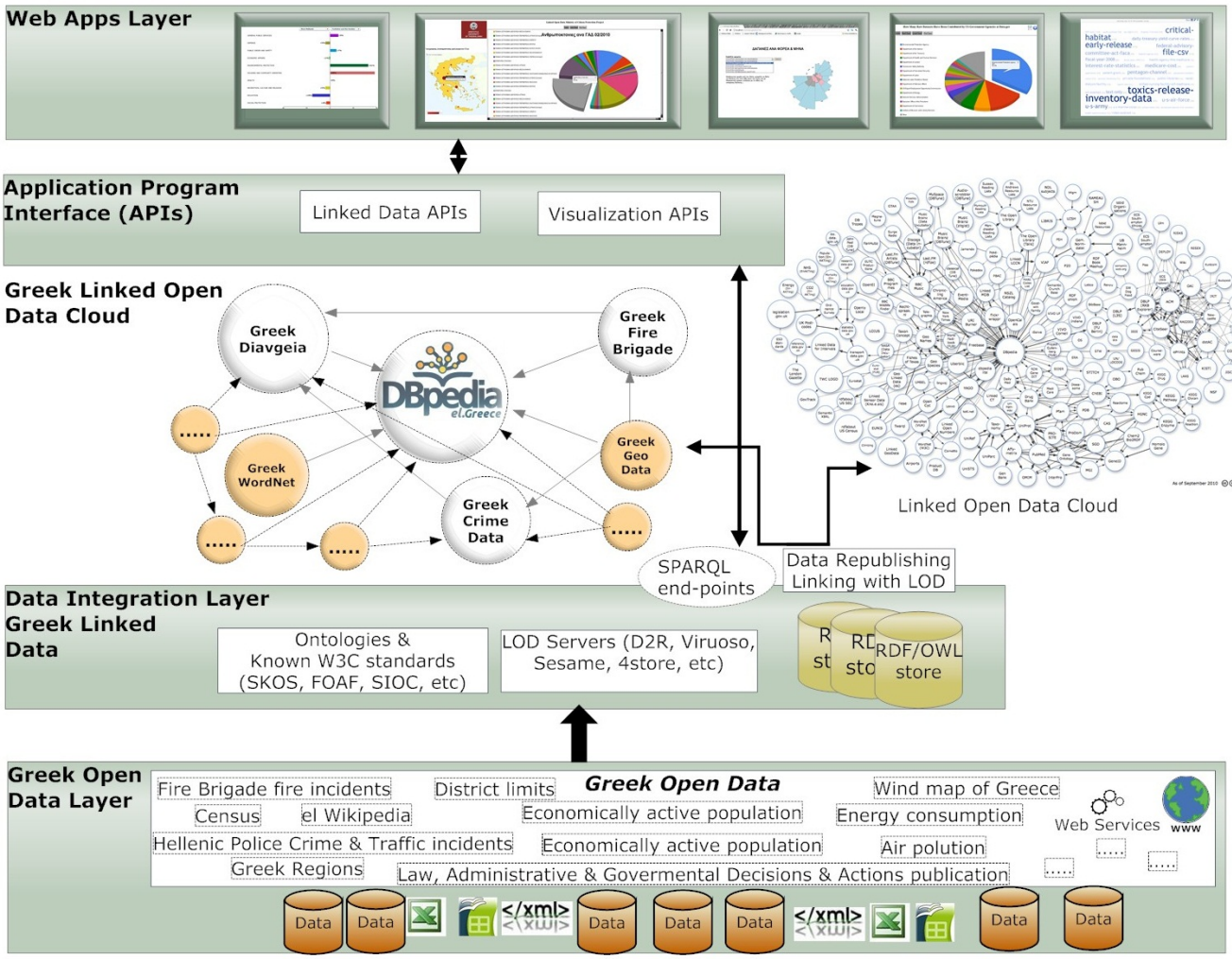

Figure 1: The Greek LOD architecture framework.

2. TOWARDS INTERNATIONALIZATION OF THE LINKED OPEN DATA: GREEK LOD

One main issue for internationalization of the LOD that needs to be solved in non-Latin languages - including Greek - is the adoption of the International Resource Identifier (IRI) [4] form, instead of Uniform Resource Identifier (URI) [2]. According to the URI rfc, resources can only contain Latin characters and everything else is percent-encoded, resulting in long and unreadable resources. The IRI format is not fully supported in current triple store implementations and this causes problems for non-Latin string SPARQL queries. To overcome such issues, custom solutions were applied (e.g. in D2R's Sparql endpoints [2] non-Latin characters have been converted to utf-8 hexadecimal codes).

The Greek LOD architecture framework is illustrated in Figure 1. The first layer demonstrates the isolated resources of Greek Open Data (published in various formats). The 2nd layer involves publishing of Greek Open Data as Linked Data. Knowledge extracted from these data is mapped and represented in a semantic manner by building new ontologies in RDF/OWL and/or reusing existing ontologies (like SKOS[1], FOAF[2], etc). These RDF or OWL triples are stored in an LOD Server component which can be thought of as a special kind of database, capable of storing, inferencing and querying RDF or OWL data. A few of the most popular ones are Open Link Virtuoso[3], D2R[4], Sesame and 4Store[5]. The Data Published LOD Linking element inter-linked and connected to the LOD cloud (e.g. via special tools like Silk [6]). In the upper layers, Linked Data APIs and the Visualization APIs (like Google Viz API, Protovis, jqplot etc) were used for the development of applications. A snapshot of the Greek LOD cloud is depicted in the middle of Figure 1. Each node represents a distinct dataset published as Linked Data. The arcs represent the links between items in the two connected datasets. The heavier white nodes represent the published datasets, following the LOD principles [3], while the orange nodes indicate the near future Greek LOD datasets. The LOD cloud in its current form is described

[1] http://www.w3.org/TR/skos-reference/
[2] http://xmlns.com/foaf/spec/
[3] http://virtuoso.openlinksw.com/
[4] http://www4.wiwiss.fu-berlin.de/bizer/d2r-server/
[5] http://4store.org/

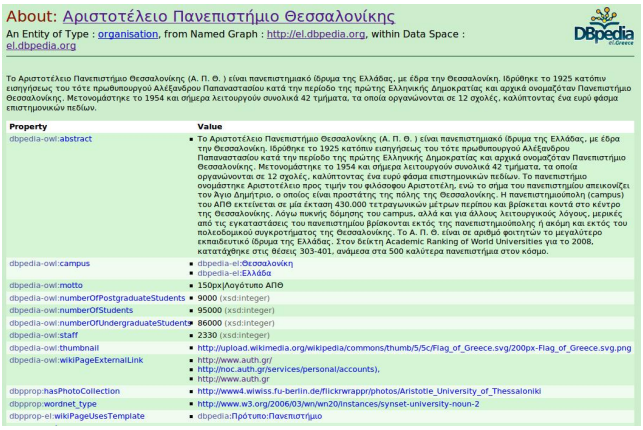

Figure 2: HTML representation using TCN rules

below.

Based on Wikipedia's, free, web-based, collaborative and multilingual context, DBpedia becomes an effort to extract Wikipedia's knowledge and to render it back enriched according to the linked data guidelines [8]. In a very short time the project has managed to transform the English DBpedia into a Linked Data hub [7]. Following the same path and although a work still in progress, the Greek DBpedia has managed to provide the same quality information and at the same time to be the first international project to provide TCN rules [5] (cf. Figure 2)

The triples were extracted using a modified DBpedia extraction framework and hosted on the virtuoso server (open-source edition). In order to address internationalization issues the Greek DBpedia project differentiated from the English in the resource domain name scheme and the resource name format. DBpedia has a global resource naming scheme for all languages. The participating multilingual articles are only the ones that provide an English translation link, using the English resource name. Using the present approach a lot of Greek articles - without an English translation link, would be unpublished. The naming scheme that fitted best the project's need is similar to the Wikipedia's naming strategy: for example http://el.dbpedia.org/page/Αριστοτέλειο_Πανεπιστήμιο_Θεσσαλονίκης (Aristotle University of Thessaloniki, Figure 2). This approach not only differentiates the default name-space (dbpedia.org) but also provides a uniform resource for all articles.

The "DIAVGEIA" project, initiated by the Greek Government (`http://diavgeia.gov.gr`) establishes the obligation to publish laws, administrative and other acts of governmental and other administrative institutions on the Web, and provides the necessary tools for easy access. Publishing acts of public institutions on line is an important step towards the integration into an "open service", allowing re-use and further processing of public sector data, as well as the query of specific legal acts, using a wide range of search criteria.

In the frame of "DIAVGEIA" project, the Diavgeia ontology was created by respectively assigning classes and properties for every table and relation described by the XML Schema. This way, the taxonomy described by the XML Schema were mapped to respective classes (keeping the same names used in the XML schema) whilst new properties were introduced to semantically describe the relations between these classes. The Diavgeia's data are mapped to Diavgeia ontology and then semantically enriched using RDFS, OWL & SKOS where possible.

In accordance to the principles of open government, transparency and unobstructed access to public data, the Ministry of Citizen Protection encouraged its departments to publish their non personal data through their portals. Part of this project involved the exploitation of open data originating from two such departments and specifically from the Hellenic Police and the Fire Brigade. The selected data involved records of crime and fire incidents and in each separate case, these were first analyzed to create an ontology suited for the occasion. The ontologies were then used to semantically describe the data by translating database records into triples stored in RDF files.

Referring to the Hellenic Police and Fire Brigade ontologies, the two main classes of Police and Fire Brigade departments connected via appropriate properties to other classes (e.g. Crimes, Traffic Incidents & Fire Incidents). In addition and to ensure that the resulting data would be more easily discovered and consumed by Semantic Web client applications, RDF/OWL links were used as bridges to connect to a number of external datasets.

The primary objective of our effort was to form the Greek LOD cloud. Driven by the vision of expanding and adding pragmatic value to this newly formed knowledge space, it was decided to seek possible inter-relations with other datasets to assist in the process. This involved both manual and automatic identification of similarity between data published in such spaces. The Silk framework [6] was employed to automatically discover and propose such links, mainly between the Police and Fire Brigade datasets with Diaygeia. It actually returned the first indications to be checked before the manual addition of the links (owl:sameAs, rdfs:seeAlso) took place. The manual identification involved interlinking with datasets, such as DBpedia, (English & Greek descriptions), Freebase, (Related info), Eurostat, (Regions & Related info), Diaygeia, (Laws, Decisions & Administrative expenses) and Geonames (Location info).

3. DEMONSTRATION

Apart from the SPAQRL endpoints, great focus was put on bringing Linked Data closer to the everyday user. For this, a number of data visualizations were created and enriched with user interaction, to allow users to filter the data according to their needs. Among the presentation options were selections of the timeframe of a chart, the records and how many of them to be included as well as the charts themselves. Interactive maps were also offered as a visualization tool enabling users to summarize and regionally compare data from interlinked datasets. Finally, a Graphical User Interface for SPARQL query construction was used, so as for the user to select from a list of predefined queries and even to modify them by adding/removing or combining data from the interlinked datasets and filtering the output.

4. CONCLUSION

We have demonstrated that the first attempt of a Greek LOD cloud exhibits potential to contribute beyond the usual cloud enrichment. Besides adding overall informational value to the global LOD cloud, we provided a framework to overcome issues arising with all non-Latin character Languages, facing restrictions in resource naming and SPARQL queries.

5. ACKNOWLEDGMENTS

This project would not have been completed without the continuous support of the DBpedia team. The administrative and financial support of the municipality of Veria is also gratefully acknowledged.

6. REFERENCES

[1] T. Berners-Lee. Linked Data - Design Issues, 2006.

[2] T. Berners-Lee, R. Fielding, and L. Masinter. Rfc 3986, uniform resource identifier (uri): Generic syntax. Request For Comments (RFC), 2005.

[3] C. Bizer, T. Heath, and T. Berners-Lee. Linked data - the story so far. *Int. J. Semantic Web Inf. Syst*, 5(3):1–22, 2009.

[4] M. Duerst and M. Suignard. Internationalized Resource Identifiers (IRIs). RFC 3987 (Proposed Standard), January 2005.

[5] K. Holtman and A. Mutz. Transparent Content Negotiation in HTTP. RFC 2295 (Experimental), March 1998.

[6] A. Jentzsch, R. Isele, and C. Bizer. Silk - Generating RDF Links while publishing or consuming Linked Data. In *Poster at the International Semantic Web Conference (ISWC2010), Shanghai*, 2010.

[7] G. Kobilarov, C. Bizer, S. Auer, and J. Lehmann. DBpedia - a linked data hub and data source for web applications and enterprises. In *Proceedings of Developers Track of 18th International World Wide Web Conference (WWW 2009), April 20th-24th, Madrid, Spain*, April 2009.

[8] J. Lehmann, C. Bizer, G. Kobilarov, S. Auer, C. Becker, R. Cyganiak, and S. Hellmann. DBpedia - a crystallization point for the web of data. *Journal of Web Semantics*, 7(3):154–165, 2009.