Geographical Linked Data: a Spanish Use Case

Alexander de León
OEG-DIA
Facultad de Informática
Universidad Politécnica de Madrid
aleon@fi.upm.es

Victor Saquicela
OEG-DIA
Facultad de Informática
Universidad Politécnica de Madrid
vsaquicela@fi.upm.es

Luis M. Vilches
OEG-DIA
Facultad de Informática
Universidad Politécnica de Madrid
lmvilches@fi.upm.es

Boris Villazón-Terrazas
OEG-DIA
Facultad de Informática
Universidad Politécnica de Madrid
bvillazon@fi.upm.es

Freddy Priyatna
OEG-DIA
Facultad de Informática
Universidad Politécnica de Madrid
fpriyatna@fi.upm.es

Oscar Corcho
OEG-DIA
Facultad de Informática
Universidad Politécnica de Madrid
ocorcho@fi.upm.es

ABSTRACT

We present the process that has been followed for the development of an application that makes use of several heterogeneous Spanish public datasets that are related to administrative, hydrographic, and statistical domains. Our application aims at analysing existing relations between the Spanish coastal area and different statistical variables such as unemployment, population, dwelling, industry, and building trade. Moreover, we provide an important innovation with respect to other similar processes followed in other initiatives by dealing with the geometrical information of features.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: Miscellaneous; E.2 [**Data storage representations**]: Linked Representations

General Terms

Design, Experimentation

Keywords

linked data, linked government data

1. INTRODUCTION

The rise of the Open Data Movement has contributed significantly to the growth of the Web of Data in the last years. This Web has started to span data sources form a wide range of domains such as people, companies, music, scientific publications, etc. Technically, Linked Data is about employing the RDF language and the HTTP protocol to publish structured data on the Web and to effectively connect data between different data sources through deferenceable URIs [3]. Data published this way is machine-readable and its semantics is explicitly defined [2]. The transformation and publication of the OpenStreetMap [1] and Ordnance Survey [6] data according to the Linked Data principles have added a new dimension to the Web of Data. However, none of them deal with complex geospatial information, they just manage every resource as a point (represented by a coordinate of latitude and longitude), while we deal with these coordinate types and more complex geometry (*LineString*).

GeoLinked Data[1] is an open initiative whose aim is to enrich the Web of Data with Spanish geospatial data. This initiative has started off by publishing diverse information sources belonging to the National Geographic Institute of Spain[2] (IGN-E), and the National Statistic Institute in Spain[3] (INE). Such resources are made available as RDF knowledge bases according to the Linked Data principles.

This paper describes the process that has been followed for the development of an application that combines these diverse Spanish public datasets so that relationships can be inferred amongst these data. The goal of the application is to analyze existing possible relations in the Spanish coastal area and different statistical variables such as unemployment, population, dwelling, industry, and building trade. By this way, Open Government Data should help us to know how seasonal employment changes in these places of Spain, a country where tourism is a very important sector for the economy. Additionally, the application deals with the geometrical information of features, so that spatial data can be extrapolated to the development of similar applications.

The rest of the paper is organized as follows: Section 2 explains the process we followed for the generation of linked data, and Section 3 presents the conclusions and future work.

2. A PROCESS FOR PUBLISHING LINKED GOVERNMENT DATA

In this section we present the process that we have followed for generating and publishing linked data from open government datasets. The process consists of (1) identification of the data sources, (2) generation of the ontology model, (3) generation of the RDF data, (4) alignment of the datasets, and (5) visualization the data. Next, we describe briefly each one of them.

[1] `http://geo.linkeddata.es/`
[2] `http://www.ign.es`
[3] `htpp://www.ine.es`

I-SEMANTICS Triplification Challenge 2010 Graz, Austria

2.1 Identification of the data sources

We have searched for open government information at the two institutions that we have referred to in the introduction: INE and IGN-E. Both INE and IGN-E are providers of Spanish official statistical and geographical information, respectively. Table 2.1 depicts the datasets that we have chosen for this application, together with the format in which they are available. All the datasets correspond to Spain, so their content is available in Spanish or in any of the other official languages in Spain (Basque, Catalan and Galician).

Table 1: Datasets

Data	Provenance	Format
Population	INE	Spreadsheet
Dwelling	INE	Spreadsheet
Industry	INE	Spreadsheet
Building Trade	INE	Spreadsheet
Hydrography (rivers, lake, etc)	IGN-E	Relational database (Oracle)
Beaches	IGN-E	Relational database (MySQL)
Administrative boundaries	IGN-E	Relational database (MySQL)

2.2 Ontology modelling

Our chosen datasets contain information such as time, administrative boundaries, unemployment, etc. For modelling of the information contained in the datasets we have created an ontology network [7]. This network has been developed following the NeOn Methodology [11], by reusing existing ontologies and vocabularies. Next, we describe briefly each one the ontologies that compose this network.

For describing complex statistics, we chose the **Statistical Core Vocabulary (SCOVO)** [8], which provides an expressive modelling framework for statistical information. This vocabulary[4] is currently defined in RDF(S) and terms and labels are provided in English. Regarding geospatial vocabulary we chose diverse ontologies.

- The **FAO Geopolitical Ontology**[5]. This OWL ontology includes information about continents, countries, and so on, in the English language. We have extended it to cover the main characteristics of the Spanish administrative division.
- Regarding the hydrographical phenomena (rivers, lakes, etc.) we chose **hydrOntology** [4], an OWL ontology that attempts to cover most of the concepts of the hydrographical domain.
- With respect to geometrical representation and positioning we reuse the **GML Ontology**[6] and the **WSG84 Vocabulary**[7].

Regarding the time information we chose the **Time Ontology**[8], an ontology for temporal concepts.

Taking into account that the SCOVO and the FAO Geopolitical ontologies were available in the English language, we have used the LabelTranslator system [5] to carry out the task of ontology localization. We use LabelTranslator for translating classes and properties of these ontologies to Spanish.

2.3 Generation of the RDF data

Given the different formats in which the selected datasets were available, we used two different systems for the conversion of data into RDF. Next we describe some details of both of them.

The generation of RDF from spreadsheets was performed using the NOR$_2$O [12] software library. This library performs an Extract, Transform, and Load (ETL) process of the legacy data sources, transforming these non-ontological resources (NORs) [12] into ontology instances.

The transformation of the relational database content into RDF was done using the integrated framework R$_2$O+ and ODEMapster+ [10], which is available as a NeOn Toolkit plugin[9]. This framework allows the formal specification, evaluation, verification and exploitation of semantic mappings between ontologies and relational databases.

2.4 Creation of RDF from geometrical information

Next we describe our approach for transforming geometrical information into RDF (see Figure 1).

We rely on Oracle *STO_UTIL package* for transforming the geometrical data stored in the original databases into GML[10].

The next step is to convert the generated GML into RDF. For this purpose we have developed a software library, GEOMETRYtoRDF, which defines a set of RDF triples for geometrical information. The GML generated in the previous steps is manipulated with GeoTools[11], an open source Java Library that provides tools for geospatial data. Finally, we use Jena[12] to generate the final geospatial RDF.

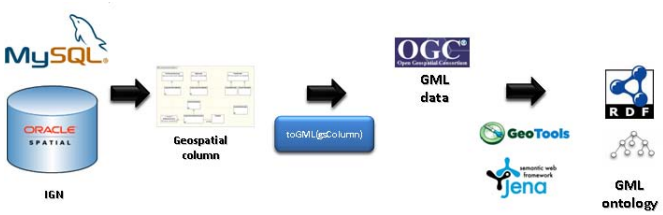

Figure 1: Transformation of the geospatial information into RDF

The RDF generated is compliant with the WSG84 vocabulary and the GML contology. Figure 2 illustrates the resource that describes the Ebro river, together with its related geometrical information.

2.5 Alignment of the datasets

The process of aligning the datasets relied on the correct identification of *owl:sameAs* relationships between administrative units and statistical information. This process is

[4] http://purl.org/NET/scovo
[5] http://www.fao.org/countryprofiles/geoinfo.asp?lang=en
[6] http://loki.cae.drexel.edu/~wbs/ontology/2004/09/ogc-gml.owl
[7] http://www.w3.org/2003/01/geo/wgs84_pos
[8] http://www.w3.org/TR/owl-time/
[9] http://www.neon-toolkit.org
[10] Geography Markup Language http://www.opengeospatial.org/standards/gml
[11] http://www.geotools.org/
[12] http://jena.sourceforge.net/

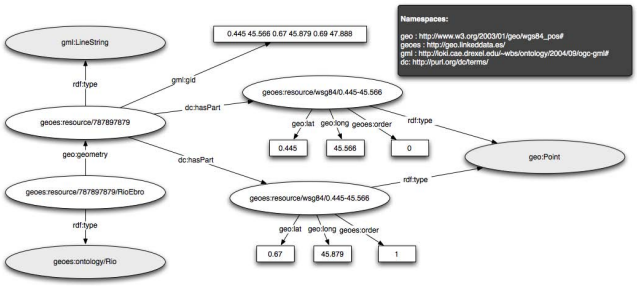

Figure 2: Ebro river and its geospatial information

based on a string matching algorithm for URIs. By this way, we enrich reference information (geometry) with data on population, unemployment, industry, etc.

2.6 Data publication and visualization

For the publication of the RDF data we rely on Virtuoso Universal Server[13]. On top of it, Pubby[14] is used for the visualization and navigation of the raw RDF data. On top of these two systems, we have developed a web based application[15] to enhance the visualization of the aggregated information. This interface combines the faceted browsing paradigm [9] with map-based visualization using the Google Maps API[16]. Thus for instance, the application is able to render on the map distinct geometrical representation such as LineStrings that depict to hydrographical features (reservoirs, beaches, rivers, etc.), or Points that show province capitals (see Figure 3).

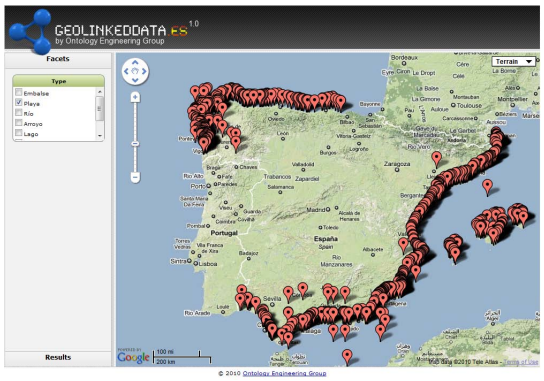

Figure 3: Screenshot of the web application

3. CONCLUSIONS AND FUTURE WORK

In this paper we have presented an application that makes use of several Spanish public datasets, specifically datasets related with administrative, hydrographic and statistical information. The application allows to analyse possible existing relations in the Spanish coastal area and different statistical variables such as unemployment, population, dwelling, industry, and building trade. Additionally, the application deals with the different geometrical information of features and alings of statistical and geometrical information.

[13] http://virtuoso.openlinksw.com/

[14] http://www4.wiwiss.fu-berlin.de/pubby/

[15] http://geo.linkeddata.es/browser

[16] http://code.google.com/apis/maps/index.html

Future work will focus on identifying and interlinking with other knowledge bases belonging to the Linking Open Initiative, mainly DBpedia and GeoNames. Moreover, we will also continue publishing GeoLinked Data on the Web for other domains and providers, and improve our faceted browser. Finally, we plan to cover complex geometrical information, i.e. not only *Point* and *LineString*-like data; but also polygons and other geometrical representation types.

4. ACKNOWLEDGMENTS

This work has been supported by the R&D project España Virtual, funded by Centro Nacional de Información Geográfica and CDTI under the R&D programme Ingenio 2010, as well as by an R+D grant from the UPM. We would like to kindly thanks Miguel Angel García and Raúl Alcázar.

5. ADDITIONAL AUTHORS

Additional authors: Carlos Buil (OEG-DIA, F.I., U.P.M., email: `cbuil@fi.upm.es`), José Mora (OEG-DIA, F.I., U.P.M., email: `jmora@fi.upm.es`) and Jean Paul Calbimonte (OEG-DIA, F. I., U.P.M., email: `jpcalbimonte@fi.upm.es`)

6. REFERENCES

[1] S. Auer, J. Lehmann, and S. Hellmann. LinkedGeoData - adding a spatial dimension to the web of data. In *ISWC*, 2009.

[2] 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.

[3] C. Bizer, T. Heath, K. Idehen, and T. Berners-Lee. Linked data on the web (ldow2008). In *WWW '08*, pages 1265–1266, New York, NY, USA, 2008. ACM.

[4] L. M. V. Blázquez, J. A. R. Gargantilla, F. J. López-Pellicer, O. Corcho, and J. Nogueras-Iso. An Approach to Comparing Different Ontologies in the Context of Hydrographical Information. In *IF&GIS*, pages 193–207, 2009.

[5] M. Espinoza, A. Gómez-Pérez, and E. Mena. LabelTranslator - A Tool to Automatically Localize an Ontology. In *ESWC*, pages 792–796, 2008.

[6] J. Goodwin, C. Dolbear, and G. Hart. Geographical linked data: The administrative geography of great britain on the semantic web. *Transaction in GIS*, 12(1):19–30, February 2009.

[7] P. Haase. D1.1.1: Networked ontology model. In *NeOn Deliverable*. 2006-12-21.

[8] M. Hausenblas, W. Halb, Y. Raimond, L. Feigenbaum, and D. Ayers. SCOVO: Using Statistics on the Web of Data. In *ESWC*, volume 5554 of *LNCS*, pages 708–722. Springer, 2009.

[9] E. Oren, R. Delbru, and S. Decker. Extending faceted navigation for RDF data. In *ISWC*, pages 559–572, 2006.

[10] F. Priyatna. RDF-based Access To Multiple Relational Data Sources. Master's thesis, Universidad Politécnica de Madrid, 2009.

[11] M. C. Suarez-Figueroa and A. Gómez-Pérez. NeOn Methodology for Building Ontology Networks: a Scenario-based Methodology. In *(S3T 2009)*, 2009.

[12] B. Villazón-Terrazas, A. Gómez-Pérez, and J. P. Calbimonte. NOR_2O: a Library for Transforming Non-Ontological Resources to Ontologies. In *ESWC*, volume 5554 of *LNCS*. Springer, 2010.