GeoLinked Data and INSPIRE through an Application Case

Luis M. Vilches-Blázquez
Ontology Engineering Group
DIA - Facultad de Informática
Universidad Politécnica de Madrid
lmvilches@fi.upm.es

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

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

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

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

Asunción Gómez-Pérez
Ontology Engineering Group
DIA - Facultad de Informática
Universidad Politécnica de Madrid
asun@fi.upm.es

ABSTRACT

In this paper 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 three themes of INSPIRE Directive, specifically Administrative Units, Hydrography, and Statistical Units. Our application aims at analysing existing relations between the Spanish coastal area and different statistical variables such as population, unemployment, dwelling, industry, and building trade. Besides providing methodological guidelines for the generation, publishing and exploitation of Linked Data from such datasets, 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.4 [**Information Systems Applications**]: Miscellaneous

General Terms

Experimentation, Standardization, Languages.

Keywords

Geospatial and statistical information, INSPIRE, RDF, Linked Data.

1. INTRODUCTION

Governments and government agencies worldwide are increasingly publishing data on the Internet, especially in the context of the Open Data Movement. Sharing this data enables greater transparency, delivers more efficient public services, and encourages greater public and commercial use and re-use of governmental information. Some governments have even created catalogs or portals (such as the United States –http://www.data.gov– and the United Kingdom –http://data.gov.uk– governments) to make it easy for the public to find and use this data [10], which are available in a range of formats (e.g., spreadsheets, relational database dumps, RDF – Resource Description Framework–) and span through a wide range of domains (e.g., geospatial, statistics, transport).

Linked Data has been recently suggested as one of the best alternatives for creating these shared information spaces [9]. The notion of Linked Data refers to the recommended best practices for exposing, sharing, and connecting RDF data via dereferenceable URIs on the Semantic Web [1]. These best practices have been adopted by an increasing number of data providers, leading to the creation of a global data space containing billions of assertions - the Web of Data [9].

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

This paper describes the results of developing an application that combines these diverse Spanish public datasets so that relationships can be inferred amongst these data. Moreover, we discuss the process followed, and propose methodological guidelines for all the activities involved within the process, which could be extrapolated to the development of similar applications.

ACM SIGSPATIAL GIS 2010 Conference, November 2-5 2010 - San Jose, California, USA.

[1] http://geo.linkeddata.es/

[2] The INSPIRE Directive addresses 34 spatial data themes needed for environmental applications. http://inspire.jrc.ec.europa.eu/

[3] http://www.w3.org/DesignIssues/LinkedData.html

This paper is structured as follows. In Section 2 we provide a general overview of the process that we propose for publishing GeoLinked Data on the Web. The details of the activity of identification and selection of data sources are provided in Section 3. In Section 4 we present the ontology network modelling process. The generation of the RDF data is introduced in Section 5, and alignment of the datasets is detailed in Section 6. In Section 7 we describe data publication and visualization. A comparison of our approach with other geospatial Linked Data approaches is presented in Section 8. Finally, Section 9 presents some conclusions of this paper and identifies our future work.

2. A PROCESS FOR PUBLISHING GEOLINKED DATA ON THE WEB

The followed process for generating and publishing GeoLinked Data from a set of data sources is a generalization of the one described in [7], and consists of the following activities: (1) identification of the data sources, (2) generation of the ontology model, (3) generation of the RDF data, (4) publication of the RDF data, and (5) linkage of the RDF data with other existing datasets in the Web of Data.

In the following sections we will be describing each of the steps in this process, in the context of the development of an application that makes use of Spanish public datasets that are related with three themes of INSPIRE, specifically with Administrative Units, Hydrography, and Statistical Units.

The goal of this application is to analyze existing relations, in the context of the Spanish coastal area, between different statistical variables such as unemployment, population, dwelling, industry, and building trade. In this way, Open Government Data should help us to know how seasonal employment changes in these areas of Spain, where the tourism sector is very relevant for their economy.

3. IDENTIFICATION AND SELECTION OF 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. 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. Used datasets

Data	Provider	Format
Population	INE	Excel Spreadsheet
Unemployment	INE	Excel Spreadsheet
Building Trade	INE	Excel Spreadsheet
Dwelling	INE	Excel Spreadsheet
Industry	INE	Excel Spreadsheet
Hydrography	IGN-E	Relational Database (Oracle)
Beaches	IGN-E	Relational Database (MySQL)
Adm. boundaries	IGN-E	Relational Database (MySQL)

Table 1 depicts the datasets that we have chosen for this application, together with the format in which they are available.

4. ONTOLOGY MODELLING

For the modelling of the information contained in the datasets (time, administrative boundaries, unemployment, etc.) we have created an ontology network, which is a collection of ontologies joined together through a variety of different relationships such as mapping, modularization, version, and dependency relationships [4]. This network has been developed following the NeOn methodology [5], by reusing existing ontologies and vocabularies. Next, we describe briefly each one of ontologies that compose this network.

- For representing complex statistics, we chose **Statistical Core Vocabulary (SCOVO)**[4], which provides an expressive modelling framework for statistical information.
- Regarding geospatial vocabulary we chose diverse ontologies.

  The **FAO geopolitical ontology**[5]. This OWL ontology includes information about continents, regions, 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**[6], an OWL ontology build following a top-down development approach, and which attempts to cover most of the concepts of the hydrographical domain.

  With respect to geometrical representation and positioning we reuse the **GML Ontology**[7] and the **WSG84 Vocabulary**[8].
- Regarding the time information we chose the **Time Ontology**[9], an ontology of temporal concepts developed into the context of World Wide Web Consortium (W3C).

Taking into account that the SCOVO and the FAO geopolitical ontologies were available in the English language, and it was important for our application to have labels in Spanish, we have used the LabelTranslator system [6] to carry out the task of ontology localization. This way we use LabelTranslator for translating components of these ontologies to Spanish.

5. GENERATION OF THE RDF DATA

We decided to use RDF as the normal form for the datasets to be published, due to the fact that we are pursuing a Linked Data approach. RDF is one of the standard languages in which

[4] http://vocab.deri.ie/scovo

[5] http://www.fao.org/countryprofiles/geoinfo.asp?lang =en

[6] http://mayor2.dia.fi.upm.es/index.php/en/ontologies/107-hydrontology

[7] http://loki.cae.drexel.edu/~wbs/ontology/2004/09/ogc-gml.owl

[8] http://www.w3.org/2003/01/geo/wgs84_pos

[9] http://www.w3.org/TR/owl-time/

information has to be made available, according to the Linked Data principles. The reason for this is that it offers several advantages, such as provision of an extensible schema, de-referenceable URIs, and as RDF links are typed, safe merging (linking) of different datasets.

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 NOR2O [2] software library. This library performs an Extract, Transform, and Load (ETL) process of the legacy data sources, transforming these non-ontological resources (NORs) [2] into ontology instances.

The transformation of the relational database (Oracle and MySQL) content into RDF was done using the integrated framework R2O+ and ODEMapster+ [3], which is available as a NeOn Toolkit plugin[10]. This framework allows the formal specification, evaluation, verification and exploitation of semantic mappings between ontologies and relational databases.

5.1 Creation of RDF of geometrical information

A very important aspect to be considered in this transformation of geographical information, which is also a distinctive feature of our approach, as it will be described later, is the definition of geometrical information in RDF. Next we describe our approach for transforming geometrical information into RDF.

GML and WKT. We rely on the Oracle STO UTIL package for the transformation of the geometrical data stored in the original databases into GML (*Geography Markup Language*)[11]. The generation of GML is applied to the *GEOMETRY* column, where different rows of a table have geometrical information of each feature.

For MySQL spatial databases, we work with the WKT[12] format for extracting information of the *GEOMETRY* column, where different rows of a table have geometrical information of each feature. In this case, we work directly with WKT, so there is no function in order to generate GML in these databases.

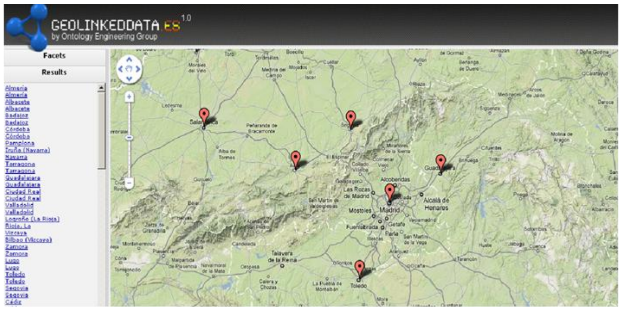
The next step is to convert the generated GML into RDF. For this purpose we have developed one software library, called GEOMETRYtoRDF, which defines a set of RDF triples for geometrical information (which could be available in GML or WKT). The GML and WKT generated in the previous steps is manipulated with GeoTools[13], an open source Java library that provides tools for geospatial data, in order to retrieve geometry, and can be also used to perform coordinate transformation if necessary. Finally, we use Jena[14], a Semantic Web Framework for Java, to generate the final geospatial RDF. The RDF generated is compliant with the WSG84 vocabulary and the GML ontology.

6. ALIGNMENT OF THE DATASETS

The process of aligning the datasets relied on the correct identification of *owl:sameAs* relation between administrative units and statistical information. This process is based on an algorithm for matching of similarity strings between URIs. By this way, we enrich reference information (geometry) with data on population, unemployment, industry, etc.

7. DATA PUBLICATION AND VISUALISATION

For the publication of the RDF data we rely on Virtuoso Universal Server[15], which is a middleware and database engine that combines the functionalities of traditional DBMS, virtual databases, RDF triple stores, XML stores, web application servers and file servers. On top of Virtuoso, Pubby[16] is used for the visualization and navigation of the raw RDF data.



Almería
Almería
Albacete
Albacete
Badajoz
Badajoz
Córdoba
Córdoba
Pamplona
Iruña (Navarra)
Navarra
Tarragona
Tarragona
Guadalajara
Guadalajara
Ciudad Real
Ciudad Real
Valladolid
Valladolid
Logroño (La Rioja)
Rioja, La
Vizcaya
Bilbao (Vizcaya)
Zamora
Zamora
Lugo
Lugo
Toledo
Toledo
Segovia
Segovia
Cádiz

Figure 1. Visualization of province capitals as *Points*

On top of these two systems, we have developed a web based application[17] to enhance the visualization of the aggregated information. This interface combines the facet browsing paradigm [8] with map-based visualization using the Google Maps API. 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.) (see Figure 2), or *Points* that show province capitals (see Figure 1).

Finally, we have implemented a faceted browser for GeoLinked Data (see Figure 2). This browser is an example of an exploratory search interface, whose design has been investigated in some recent Human-Computer Interaction (HCI) research for supporting users who have less clear or more complex needs. The application is able to render on the map the distinct geometrical shapes of the geographical features published as RDF. Statistical data is also displayed over the map so that the user can observe and compare the relative magnitudes of the statistics, which are represented by different graphs, occurring on the distinct

[10] http://www.neon-toolkit.org

[11] http://www.opengeospatial.org/standards/gml

[12] *Well-Know Text* is a text markup language for representing vector geometry objects on a map, spatial reference systems of spatial objects and transformations between spatial reference systems.

[13] http://www.geotools.org/

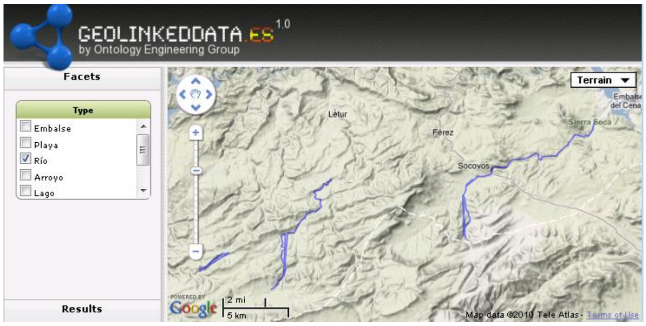
[14] http://jena.sourceforge.net/

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

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

[17] http://geo.linkeddata.es/web/guest/visualizacion-beta

geographical regions, and specifically on the coastal area of the country.



Type
- □ Embalse
- □ Playa
- ☑ Río
- □ Arroyo
- □ Lago

Figure 2. Visualization of rivers as *StringLines*

8. RELATED WORK

The transformation and publication of the OpenStreetMap [12] and Ordnance Survey [11] data according to the Linked Data principles have added a new dimension to the Web of Data. Likewise, various geodata sources are starting to appear in knowledge bases of the Linked Open Data initiative such as:

- Ordnance Survey (Great Britain’s national mapping agency). It provides Linked Data of the administrative and voting regions in Great Britain. This data includes the names, census code, and area of the regions of Great Britain.
- LinkedGeoData. It provides Linked Data of the OpenStreetMap project, interlinking this data with other knowledge bases in the Linking Open Data initiative.
- GeoNames[18]. It integrates geographical data, such as place names, population, etc. from various sources. GeoNames gives access to users to manually edit, correct, and add new names.
- DBpedia[19]. It extracts structured information from Wikipedia, linking this information to other datasets, and making it available as Linked Data.

However, neither the Ordnance Survey nor the OpenStreetMap initiatives, as the main geospatial Linked Data providers to date, deal with complex geospatial information as we do in our approach. For the time being, they just manage every resource as a point, while we deal with this coordinates types and more complex geometry (*LineString*).

9. CONCLUSIONS AND FUTURE WORK

In this paper we have presented an application that makes use of several Spanish public datasets, specifically datasets related with three INSPIRE themes (Administrative units, Hydrography, and Statistical units). The goal of this application case is to analyze 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 establishes alignments of statistical and geometrical information. Moreover, we described the process we followed, and proposed methodological guidelines for all the activities involved.

Future work will focus on identifying and interlinking with other knowledge bases belonging to the Linking Open Data Initiative. 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.

10. 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 UPM.

11. REFERENCES

[1] Bizer, C., Heath, T., Idehen, K., and Berners-Lee, T. 2008. Linked data on the web (LDOW2008). In Proceeding of the 17th international conference on World Wide Web, pages 1265-1266, Beijing, China.

[2] Villazón-Terrazas, B. Gómez-Pérez, A. and Calbimonte, J. P. 2010. NOR2O: a Library for Transforming Non-Ontological Resources to Ontologies. In ESWC, volume 5554 of Lecture Notes in Computer Science. Springer.

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

[4] Haase, P., Rudolph, S., Wang, Y., Brockmans, S., Palma, R., Euzenat, J., d'Aquin, M. 2006. NeOn Deliverable D1.1.1. Networked Ontology Model. Available at: http://www.neon project.org/.

[5] Suárez-Figueroa, M.C. 2010. NeOn Methodology for Building Ontology Networks: Specification, Scheduling and Reuse. PhD Thesis. Universidad Politécnica de Madrid.

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

[7] Bizer, C., Cyganiak, R., Heath, T. (2007) How to publish Linked Data on the Web. http://www4.wiwiss.fu-berlin.de/ bizer/pub/LinkedDataTutorial/

[8] Oren, E., Delbru, R., and Decker, S. 2006. Extending faceted navigation for RDF data. In International Semantic Web Conference, pages 559-572.

[9] Bizer, C., Heath, T. and Berners-Lee, T. 2009. Linked data - the story so far. International Journal on Semantic Web and Information Systems (IJSWIS). Vol. 5(3), Pages 1-22.

[10] W3C. 2009. Publishing Open Government Data. W3C Working Draft. http://www.w3.org/TR/gov-data/

[11] Goodwin, J., Dolbear, C., Hart, G. (2009) Geographical Linked Data: The Administrative Geography of Great Britain on the Semantic Web. Transactions in GIS, Volume 12 Issue s1, Pages 19 – 30

[12] Auer, S., Lehmann, J., Hellmann, S. (2009) LinkedGeoData – Adding a spatial Dimension to the Web of Data. In Proc. of 7th International Semantic Web Conference (ISWC).

[18] http://www.geonames.org/

[19] http://dbpedia.org/