An Approach to Publish Spatial Data on the Web: The GeoLinked Data Case

Luis. M. Vilches-Blázquez, Boris Villazón-Terrazas, Alexander De Leon, Freddy Priyatna, and Oscar Corcho
[1] Ontology Engineering Group. Dpto. de Inteligencia Artificial
Facultad de Informática, Universidad Politécnica de Madrid
28660, Boadilla del Monte, Madrid, Spain
{lmvilches, bvillazon,aleon,fpriyatna,ocorcho}@fi.upm.es

Abstract. In this paper we report on an ongoing process aimed at publishing hydrographical data on the Web with a Spanish GeoLinked Data Use Case. Moreover, we discuss the process we followed, and propose methodological guidelines for all the activities involved within the process.

Keywords: RDF, hydrographical information, GeoLinked data

1 Introduction

The rise of the Open Data Movement has led to the Web of Data grow significantly over 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 connect data between different data sources, effectively allowing data in one data source to be linked to data in another data source [8]. This way for publishing data enables that it is machine-readable and meaning is explicitly defined [7]. Further details about sets of rules for publishing data on the Web are shown in [6].

The transformation and publication of the OpenStreetMap [5] and Ordnance Survey [4] data according to the Linked Data principles have added a new dimension to the Web of Data. In this way spatial data can be retrieved and interlinked on an unprecedented level of granularity.

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. Such sources are made available as RDF (Resource Description Framework) knowledge bases according to the Linked Data principles. With this work, the Spanish National Geographic Institute data has joined the Linked Data initiative. In this paper we present the ongoing process of publishing geospatial datasets from the Spanish National provider, specifically data belongs to the hydrographical domain.

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

Next, we discuss the followed process, and propose methodological guidelines for all the activities involved within the process.

2 Publishing GeoLinked Data on the Web

In this section we describe the process we followed for generating GeoLinked Data. This process is based on [2] 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) linking the RDF data with existing other datasets in the cloud. Figure 1 depicts these activities that are described in the following.

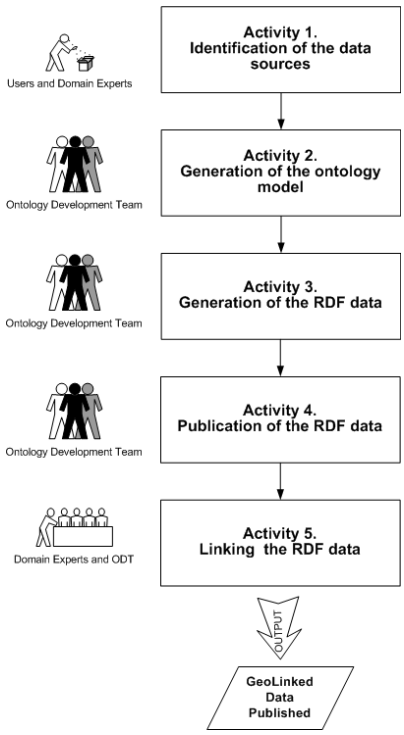

Figure 1. Process for generating GeoLinked Data

2.1 Identification of the data sources

We dealt with different geospatial databases which have information at various levels (European and National) related to Spanish hydrographical features. These databases belong to the National Geographic Institute of Spain; however, these sources are at

different scales (from 1:1 million to 1:25,000), and have multiple information related to Spanish geographical feature instances. An overview of the databases used is shown in the Figure 2.

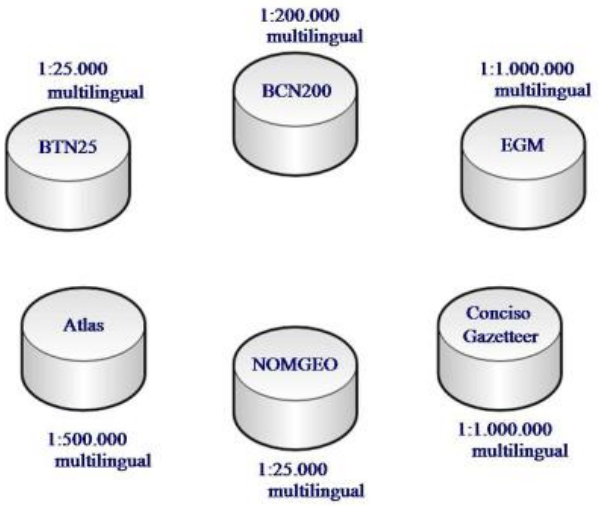

Figure 2. Geospatial databases context

With regard to the European level, we focused on the EuroGlobalMap project, which is supervised by EuroGeographics[2]. EuroGlobalMap (EGM) is a topographic dataset that covers the whole of Europe at scale 1:1 million. This dataset is produced in cooperation with the National Mapping Agencies (NMAs) of Europe, using official national databases. This project provides the first European Geospatial Information infrastructure that is maintained at the source level by the NMAs. In addition, this data source offers smooth access conditions to geographical information. It contains 3,500 Spanish hydrographical toponyms.

Within the national level, we worked with five databases. Next, we describe briefly the main characteristics of them.

- The Numerical Cartographic Database, called BCN200 (scale 1:200,000), is a data source that complies with the data specifications required to be exploited in Geographic Information Systems (GIS) environments. This data source includes 60,000 toponyms related to hydrographical instances.
- The Numerical Cartographic Database, called BTN25, (scale 1:25,000) was built to obtain cartographic information at the abovementioned scale because this scale complies with the data specifications required to be exploited in Geographic Information Systems (GIS) environments. This database includes more than 190,000 entries related to hydrographical toponyms.

[2] EuroGeographics represents nearly all European National Mapping and Cadastral Agencies.

- The National Geographic Gazetteer data source (scale 1:50,000) also called Georeferenced DataBase or NOMGEO, has more than 490,000 toponyms, of which more than 74,000 are hydrographical toponyms belonging to 44 different features.
- The Conciso Gazetteer data source is a basic corpus of standardized toponyms created by the Spanish Geographical Name Commission. This data source has more than 3,600 toponyms and its information is compiled at a scale 1:1 million. This gazetteer agrees with the United Nations Conference Recommendations on Geographic Names Normalization.
- The National Atlas data source (scale 1: 1:500,000) provides an overview of Spain's human and physical environment and integrates different thematic data. This data source has more than 1,100 hydrographical instances.

All these data sources contain multilingual information in the official languages of Spain (Castilian, Galician, Catalan, Basque, and Aranese).

2.2 hydrOntology as ontology model

hydrOntology [9] is an ontology in OWL that follows a top-down development approach. Its main goal is to harmonize heterogeneous information sources coming from several cartographic agencies and other international resources. Initially, this ontology was created as a local ontology that established mappings between different data sources (feature catalogues, gazetteers, etc.) of the Spanish National Geographic Institute (IGN-E). Its purpose was to serve as a harmonization framework among Spanish cartographic producers. Later, the ontology has evolved into a global domain ontology and it attempts to cover most of the concepts of the hydrographical domain.

hydrOntology has been developed according to the ontology design principles proposed by [12] and [13]. Some of its most important characteristics are that the concept names (classes) are sufficiently explanatory and are correctly written. Thus each class tries to group only one concept and, therefore, classes in brackets and/or with links (“and”, “or”) are avoided. According to certain naming conventions, each class is written with a capital letter at the beginning of each word, while object and data properties are written with lower case letters.

Regarding methodological issues, the approach adopted is METHONTOLOGY, a widely-used ontology building methodology. This methodology emphasises the reuse of existing domain and upper-level ontologies and proposes using, for formalisation purposes, a set of intermediate representations that can be later transformed automatically into different formal languages. A detailed description of the methodology for building this ontology can be found in [11].

In order to develop this ontology following a top-down approach, different knowledge models (feature catalogues of the IGN-E, the Water Framework European Directive, the Alexandria Digital Library, the UNESCO Thesaurus, Getty Thesaurus, GeoNames, FACC codes, EuroGlobalMap, EuroRegionalMap, EuroGeonames, several Spanish Gazetteers and many others) have been consulted; additionally, some integration issues related to geographic information and several structuring criteria

[10] have been considered. The aim was to cover most of the existing geospatial information sources and build an exhaustive global domain ontology. For this reason, the ontology contains one hundred and fifty (150) relevant concepts related to hydrography (e.g. river, reservoir, lake, channel, and others), 34 object properties, 66 data properties and 256 axioms.

Additionally within this activity we devoted some effort to choosing cool URIs[3] [14] for our resources. For concepts and properties we follow the pattern: `http://geo.linkeddata.org/ontology` and for instances we follow the pattern: `http://geo.linkeddata.org/resource`.

2.3 Generation of the RDF data

For generating the RDF data we rely on the integrated framework: R_2O and ODEMapster [1]. This framework allows the formal specification, evaluation, verification and exploitation of the semantic mappings between ontologies and relational databases. This integrated framework consists of:

- R_2O is a declarative, XML-based language that permits describing arbitrarily complex mapping expressions between ontology elements (concepts, attributes and relations) and relational elements (relations and attributes).
- ODEMapster is a processor that generates Semantic Web instances from relational instances based on the mapping description expressed in an R_2O document. ODEMapster offers two modes of execution: (1) query driven upgrade (on-the-fly query translation) and (2) a massive upgrade batch process that generates all possible Semantic Web individuals from the data repository.

In the context of the NeOn Project[4], we have developed the ODEMapster plug-in that is included in the NeOn Toolkit[5]. This plug-in offers the user a graphical user interface to create, execute, or query the R_2O mappings.

As can be seen in Figure 3, an ontology defines terms in a particular domain (territory, area, economic group, etc) and a database does the same in another (geographical group, economical unions, etc). The level of overlap of these two domains allows the definition of correspondences between the terms of one and the other.

A query like the one described in Figure 3 "Give me the names of all geographical regions and economic regions" would have its corresponding SQL over the database and, similarly, the tuples returned by the database would have their equivalent in a set of instances of the ontology answering the initial question.

[3] http://www.w3.org/TR/cooluris/
[4] http://www.neon-project.org
[5] http://www.neon-toolkit.org

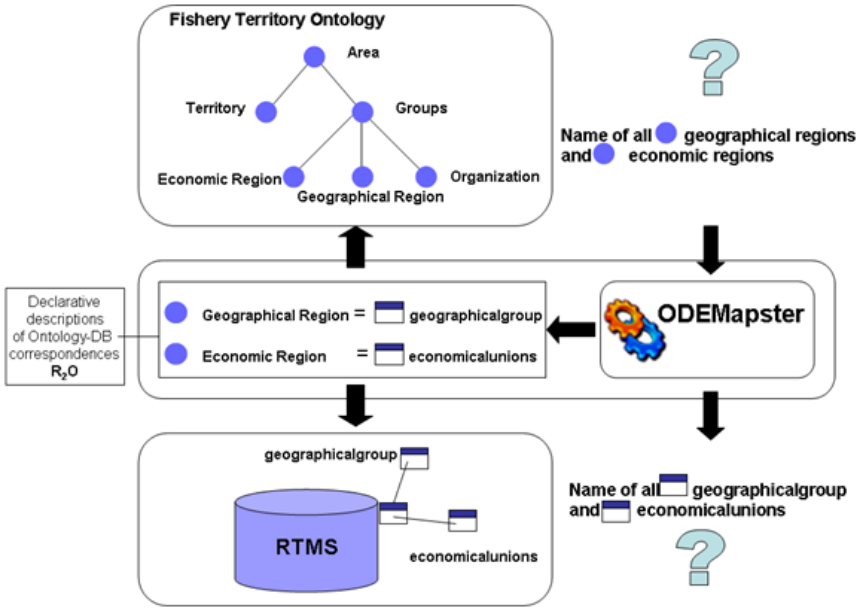

Figure 3. Schematic description of the R_2O and ODEMapster

The question is expressed in ODEMQL query language [1], which is specifically designed for ODEMapster processor; the ontology must be represented in OWL or RDF(S); and the database must be stored in ORACLE or MySQL database.

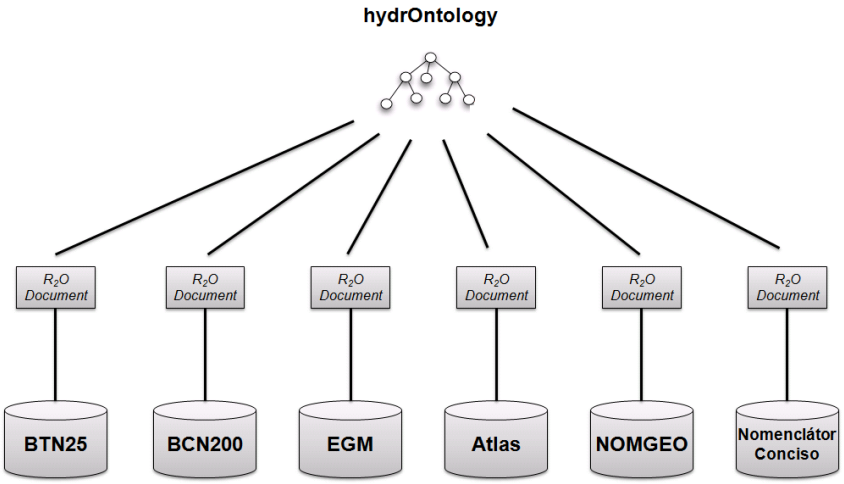

Figure 4. Geographical datasets mapped to the hydrOntology

In this particular case, we have created an R_2O document for every dataset. Then we ran the ODEMapster processor, for generating the RDF instances (see Figure 4).

To keep up-to-date the RDF datasets we just need to run the processor whenever the data sources are updated.

Finally, for the process of aligning the resources that come from different data sources we relied on the generation of the same URIs for these resources.

2.4 Publication of the RDF data

For the publication of the RDF data we rely on Virtuoso Universal Server[6], which is a middleware and database engine hybrid that combines the functionality of a traditional DBMS, virtual database, RDF, XML, free-text, web application server and file server functionality. Pubby[7] was also installed to provide visualization and navigation of the raw data.

Additionally we developed a web based application[8] to enhance visualization of the aggregated information. This interface combines the facet browsing paradigm [8] with map based visualization (see Figure 3). The application is able to render on the map distinct hydrographical features (for instance, reservoirs, beaches, lakes, etc.) published as RDF.

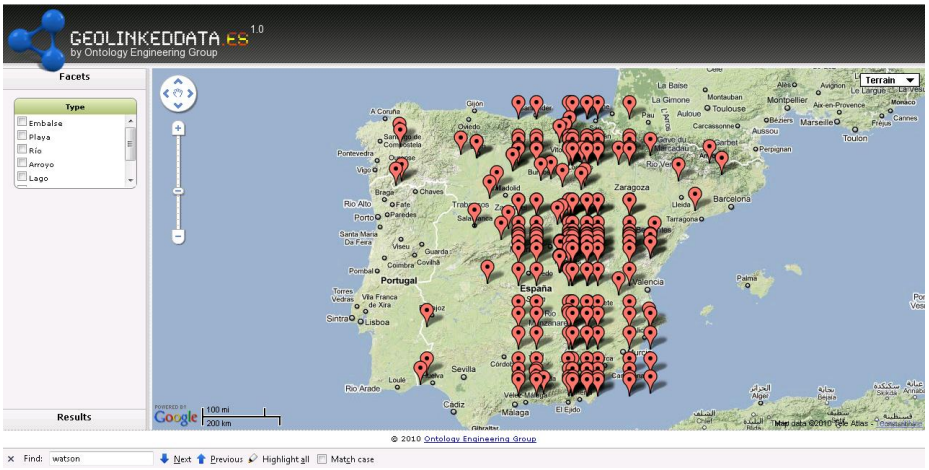

Figure 5 Screenshot of our Web application

Within this activity we have to link our resources with existing resources in the cloud. This is an ongoing work and we are relying on DBpedia[9] and GeoNames[10] for identifying and linking our resources into the cloud.

---

[6] http://virtuoso.openlinksw.com/
[7] http://www4.wiwiss.fu-berlin.de/pubby/
[8] http://geo.linkeddata.es/browser
[9] http://dbpedia.org/
[10] http://www.geonames.org/

As many geospatial providers, including licensing and provenance information in the graph is very important. We have included some elements of Dublin Core[11] and voID[12] metadata concerning provenance and licensing as additional graph triples.

Finally, we are implementing a faceted browser[13] for GeoLinked Data (see Figure 5). This browser is an example of Exploratory Search Interfaces, whose design has been investigated in some recent Human Computer Interaction (HCI) research for supporting users who have less clear or more complex needs [3].

3 Conclusions and future work

In this paper we report on an ongoing process aimed at publishing spatial data on the Web with a Spanish GeoLinked Data Use Case. Moreover, we described the process we followed, and proposed methodological guidelines for all the activities involved within the process.

Future work will focus on identifying and interlinking with other knowledge bases belonging to the Linking Open Data 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 like data, i.e. not only point like data, we will also treat information representation through lines and polygons.

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 Angél García and Raúl Alcázar.

References

1. Barrasa, J. (2007) Modelo para la definición automática de correspondencias semánticas entre ontologías y modelos relacionales. Ph.D. Thesis, Universidad Politécnica de Madrid, Madrid, 2007.
2. Bizer, C., Cyganiak, R., Heath, T. (2007) How to publish Linked Data on the Web. Retrieved June 14, 2009, http://www4.wiwiss.fu-berlin.de/bizer/pub/LinkedDataTutorial/
3. Wilson, M. L., Schraefel, M. C. (2007) Bridging the Gap: Using IR Models for Evaluating Exploratory Search Interfaces. In: SIGCHI 2007 Workshop on Exploratory Search and HCI, 28th April 2007, San Jose, CA, USA.
4. 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

[11] http://dublincore.org/
[12] http://vocab.deri.ie/void/guide#sec_1_6_SPARQL_endpoint_and_Examp
[13] http://geo.linkeddata.es:8181/GeoLinkedDataBrowser/

5. Auer, S., Lehmann,, J., Hellmann, S. (2009) LinkedGeoData – Adding a spatial Dimension to the Web of Data, ISWC 2009.
6. Berners-Lee, T. (2006). Linked Data - Design Issues. Retrieved August 05, http://www.w3.org/DesignIssues/LinkedData.html
7. Bizer, C., Heath, T., Berners-Lee, T. (2009) Linked Data - The Story So Far. Heath, T., Hepp, M., Bizer, C. (Eds.) International Journal on Semantic Web and Information Systems, pp 1-22.
8. Bizer, C., Heath, T., Idehen, K., Berners-Lee, T. (2008) - LDOW2008. In WWW'08 pp. 1265-1266, Beijing, China.
9. Vilches-Blázquez, L. M., Ramos, J. A., López-Pellicer, F. J., Corcho, O., Nogueras-Iso, J. (2009). "An approach to comparing different ontologies in the context of hydrographical information". In Popovich et al., (eds.), IF&GIS'09. LNG&C Springer: St. Petersburg, 193-207.
10. Vilches-Blázquez L.M., Bernabé-Poveda M.A., Suárez-Figueroa M.C., Gómez-Pérez A., Rodríguez-Pascual A.F. "Towntology & hydrOntology: Relationship between Urban and Hydrographic Features in the Geographic Information Domain". In Ontologies for Urban Development. Studies in Computational Intelligence, Springer. 2007, vol.61, pp73–84
11. Gómez-Pérez, Asunción, M. Fernández-López, Oscar Corcho, (2003). Ontological Engineering. Springer-Verlag: London.
12. Gruber T.R. (1995). "Toward principles for the design of ontologies used for knowledge sharing". International Journal of Human-Computer Studies, 1995, v.43 n.5-6.
13. Arpírez JC, Gómez-Pérez A, Lozano A, Pinto HS (1998). "(ONTO)2Agent: An ontology-based WWW broker to select ontologies". In: Gómez-Pérez A, Benjamins RV (eds) ECAI'98 Workshop on Applications of Ontologies and Problem-Solving Methods: Brighton, 16-24.
14. Ayers, D., Vlkel, M. (2008) Cool URIs for the semantic web. Interest Group Note 20080331, W3C, 2008. URL http://www.w3.org/TR/2008/NOTE-cooluris-20080331/. W3C Interest Group Note 31 March 2008.