Huerta, Schade, Granell (Eds): Connecting a Digital Europe through Location and Place. Proceedings of the AGILE'2014 International Conference on Geographic Information Science, Castellón, June, 3-6, 2014. ISBN: 978-90-816960-4-3

Cadastral data integration through Linked Data

Jhonny Saavedra
Universidad Politécnica de Madrid
Madrid, Spain
ja.saavedra@alumnos.upm.es

Luis M. Vilches-Blázquez
Ontology Engineering Group, Dpto. de Inteligencia Artificial, Facultad de Informática.
Universidad Politécnica de Madrid, Madrid, Spain
lmvilches@fi.upm.es

Alberto Boada
Instituto Geográfico Agustín Codazzi
Bogotá, Colombia
aboadar@igac.gov.co

Abstract

Cadastral data is one of the more important types of geospatial data. Taking into account the importance of these data, several international bodies have worked for creating a standardised model for land administration. However, in spite of existing efforts, there are several open issues for the development of a harmonized vision of cadastral data. Taking into account this scenario, Linked Open Data may allow addressing some of these challenges, by proposing best practices for exposing, sharing, and integrating data on the Web.

This paper shows a use case where two cadastral information sources are semantically integrated according to Linked Data principles. These sources belong to different Colombian cadastral producers and are characterized by different heterogeneity issues. Herein, we describe an implementation of Linked Data principles in the cadastral domain using LADM standard (ISO 19152) and GeoSPARQL. Besides, our original data are enriched with different dataset of Linked Data cloud (LinkedGeoData and GeoNames).

Keywords: Cadastre, Linked Data, Land Administration Data Model, GeoSPARQL.

1 Introduction

Cadastral data is one of the more important types of geospatial data. They have been the basis for the land tributes since Roma's times. These data are defined as the geographic extent of the past, current, and future rights and interests in real property including the spatial information necessary to describe that geographic extent [1]. Rights and interests are the benefits or enjoyment in real property that can be conveyed, transferred, or otherwise allocated to another for economic remuneration. Rights and interests are recorded in land record documents. The spatial information necessary to describe rights and interests includes surveys and legal description frameworks such as the Public Land Survey System, as well as parcel-by-parcel surveys and descriptions [1].

Taking into account the importance of these data, several international bodies have worked for creating a standardised model for land administration. Thus, the Federation International of Surveyors (FIG) started to work at 1996 for developing a future cadastral system (Cadastre 2014) [2]. This proposal, which has become the inspiration of the modern cadastres, has allowed increasing associated services of cadastral systems through technology deployment. Likewise, other more recent efforts for developing standardized models in order to facilitate cadastre data exchange are INSPIRE Data Specification on Cadastral Parcels [3] and the ISO standard 19152 Land Administration Domain Model (LADM) [4].

However, in spite of existing efforts, there are several open issues for the development of a harmonized vision of cadastral data. Some of these challenges are related to integration process of different cadastral data and how to connect these data with related information (e.g.: public services, demographic statistics, planning, etc.).

Linked Open Data may allow addressing some of these challenges, by proposing best practices for exposing, sharing, and integrating data on the Web [5]. The principles of Linked Data were first outlined by Berners-Lee in [6], using the following four guidelines: (1) Use URIs as names for things. (2) Use HTTP URIs so that people can look up those names. (3) When someone looks up a URI, provide useful information, using standards, such as: Resource Description Framework (RDF) and SPARQL Query Language for RDF (SPARQL) and (4) Include links to other URIs, so that they can discover more things. Further details about sets of rules for publishing data on the Web are shown in (Berners-Lee 2006).

Several approaches generating and publishing geospatial Linked Data are appearing in the state-of-the-art in order to perform a semantic information integration process. The capabilities of LOD for integration and interoperability into geospatial context have been recognised by many authors like [7], [8], and [9]. This, and fact attaches an increasing the interest for publishing geospatial information as Linked Data over the last years. Nowadays, we find more than 68 geospatial datasets and six billion of triples with location (often, a pair of coordinates) in the Linked Data cloud.

This paper shows a use case where different cadastral information sources are semantically integrated according to Linked Data principles. These sources belong to different Colombian cadastral producers and are characterized by different heterogeneity issues. Herein, we describe one of the first implementations of Linked Data in the cadastral domain using LADM standard (ISO 19152) and GeoSPARQL[1]. Besides, our original data are enriched with different dataset of Linked Data cloud (LinkedGeoData and GeoNames).

[1] http://www.opengeospatial.org/standards/geosparql

This paper is structured as follows. We start providing a description of our use case for integrating Colombian cadastral data (section 2). Next, we present a brief overview of the existing related work (section 3). In section 4, we describe the process for generating and publishing cadastral Linked Data. Finally, we summarize some conclusions and identify future work in section 5.

2 Colombian Cadastre: An integration use case

Colombia's cadastre has been one of the most developed in Latin-American Region. Currently, the National Cadastral Authority is the Colombian National Geographic Institute (*Instituto Geographic Agustin Codazzi* – IGAC). Furthermore, this country has four additional cadastral producers, which generate information in a decentralised way, in the municipalities of Medellin, Bogotá, Cali and Antioquia.

These different cadastral producers have different and heterogeneous models, vocabularies, and their own production and management systems. A representative example of this heterogeneity is associated with the National Authority, which has different cadastral model in order to manage generated data.

In order to overcome these barriers, IGAC is working on a Cadastral National System. The goals of this project are to create and consolidate a unique cadastral model for National data, deploy this model in a distributed database system and to create a web for providing these data to final users. However, this project has not taken into account the decentralised cadastral offices, due to the fact that there was not an agreement between several producers for centralizing the cadastral data management.

Driven by this scenario, we present a use case, where main purpose is to support the reusing, exchange and semantic integration of Colombian's cadastral data. Within this use case, which focuses on physical aspect of cadastral information, we deal with heterogeneous datasets, which belong to different producers, and they are reused and semantically integrated in order to connect cadastral data and keep provenance of the different sources and producers. This diversity of producers and datasets entails different issues related to heterogeneity of datasets, which are solved using Linked Data principles. Thus, we take cadastral datasets from the National producer (IGAC) and another dataset from a local producer, concretely from Bogotá cadastre[2].

3 Related work

Currently, there are more than 890 RDF datasets tagged as Linked Data in the Datahub[3]. From these datasets, 61 are tagged as geographic data[4]. These data not only come from geospatial research labs, otherwise they are published by important producers, such as Ordnance Survey[5], National Geographical Institute of Spain[6], U.S. Geological Survey[7], and so on. Besides, there exist important collaborative initiatives (e.g.: OpenStreetMap[8] and GeoNames[9]) that are part of this movement. Within published geospatial Linked Data the most common topics are related to transport, toponyms, administrative boundaries, environmental and statistical data. There are no data associated with cadastral information in this repository (Datahub). However, there exist an initiative related to register information called "The Land Register of UK[10], which is publishing the register and cadastral information according to Linked Data. Besides, in [11] is described an approach for cadastre-register integration using ontologies and Semantic Web.

With respect to cadastral models, Land Administration Domain Model (LADM) - ISO 19152 is the most recent resource in the cadastral topic. This standard provides terminology for land administration and enables the combining of land administration information from different sources in a coherent manner. In spite of the fact that it is a recent standard, LADM has already profiles in different countries, such as: Spain, Portugal, Germany, and Japan among others. Taking into account LADM standard, there is a first approach for developing an OWL ontology of LADM [10].

4 Cadastral data integration

In order to perform this integration process, we use the methodological guidelines for generating, integrating and publishing geospatial data according to Linked Data principles described in [12]. These guidelines propose an iterative incremental life cycle model where data gets continuously improved and extended. It consists of the following steps: (1) specification, (2) modelling, (3) RDF generation, (4) links generation, (5) publication, and (6) exploitation. The detail of each of them is shown in the next items.

4.1 Specification

As aforementioned, we work with datasets from a National producer (IGAC) and a local producer (Bogotá cadaster). Next, we describe main characteristic associated with these datasets: On the one hand, the cadastral data of IGAC are stored in an ESRI geodatabase. Within this geodatabase, we work with a subset of cadastral data, which was provided by IGAC in shapefiles (*shp*). These shapefiles belong to the municipality of *Soacha*. On the other hand, Bogotá cadaster data are available in its website in different formats (e.g.: SHP, KMZ, DWG or GML). In the context of our work, we downloaded and manipulated data in shapefile format. In this

[2] The used dataset in this work belongs to Bogotá Spatial Data Infrastructures (IDECA). http://www.ideca.gov.co/

[3] The Datahub is a data management platform from the Open Knowledge Foundation which collects many of the data sets published as Open Linked Data. http://datahub.io/es/dataset?tags=lod

[4] http://datahub.io/es/owns/dataset?tags=geographic&tags=lod

[5] data.ordnancesurvey.co.uk/

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

[7] http://cegis.usgs.gov/ontology.html

[8] http://linkedgeodata.org/

[9] http://www.geonames.org/ontology/documentation.html

[10] http://landregistry.data.gov.uk/

case, we work with a subset of two localities of the Bogotá city (a subset related to the central area of Bogotá and another, which limits with *Soacha* municipality, called *Bosa*). Both of these datasets have attributes related to their information domain (e.g.: area, name, label, geometry, address, etc.). However, they also have differences in their models and used vocabulary.

With respect to URI design, we adopt recommendation of [13], that is, we design our URIs to be simple, stable and manageable. Taking into account this, we use the domain http://datos.igac.gov.co/ for publishing our cadastral Linked Data. According to this root domain, we propose the following URI pattern for this work: In order to identify the provenance of data, we create two URI for pointing to different producers. Thus, IGAC resources are identified with *http://datos.igac.gov.co/id/catastro/igacsnc/*+ and Bogotá resources use the following URI *http://datos.igac.gov.co/id/catastro/bogota/*+.

Finally, with respect to the URI pattern for resources (instances), we use the following pattern, where we add to each URI the National cadastral code:

- IGAC:
  http://datos.igac.gov.co/id/catastro/igacsncl/
  25754010100000069000500000000
- Bogotá cadastre:
  http://datos.igac.gov.co/id/catastro/bogota/
  25754010100000045290200010000

4.2 Ontology modelling

For the modelling of the information contained in the aforementioned datasets 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. This network has been developed following the NeOn methodology [14], by reusing existing ontological and non-ontological resources. Next we provide some details about used resources.

Regarding cadastral domain, we reuse a non-ontological resource, which is the core of our ontology network. This resource is the ISO 19152 Land Administration Domain Model (LADM). LADM proposal has not a deep level of detail and is composed of two classes (*SpatialUnitGroup* and *SpatialUnit*).

Taking into account the general viewpoint provided by LADM, we decided to perform an extension of this proposal for considering Colombian cadastral characteristics. Thus, we developed a profile of LADM, called LADM_CO, for modelling these issues in an ontology. A subset of the LADM_CO ontology is shown in the Figure 1, which includes classes as *neighbourhood*, *block*, *construction*, *land*, and so on. Besides, we developed two different ontologies for modelling characteristics of each Colombian cadastral producer (IGAC and Bogotá cadastre). After we developed these ontologies, our work focused on setting mapping between the components of these ontologies.

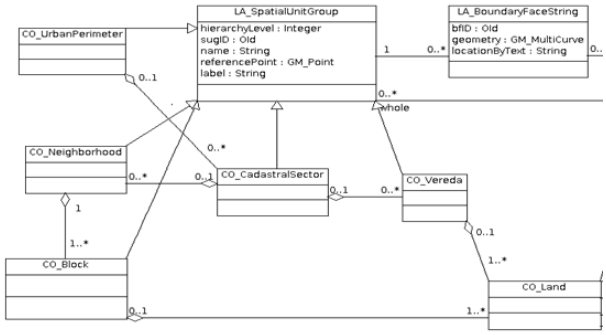

Figure 1. Main classes of the LADM_CO ontology

.

Likewise, we developed an ontology for modelling Colombian administrative boundaries, and reuse GeoSPARQL ontology in order to achieve two essential issues for our work. On the one hand, being able to represent complex geometry and, on the other hand, supporting spatial queries for exploiting cadastral information.

4.3 RDF generation

According to the followed methodology, in this activity we have to take the data sources selected in the specification activity, and transform them to RDF according to the vocabulary mentioned in the modelling activity. We use different systems to transform cadastral features from aforementioned datasets into RDF.

For dealing with geospatial information, we develop an extension of Geometry2RDF[11], called shp2GeoSPARQL[12], in order to transform geometrical information from datasets to RDF. This extension parses shapefiles in order to retrieve the associated geometric data, and generate the geospatial RDF according to GeoSPARQL ontology. For dealing with thematic data, we use Google Refine[13] and its RDF extension[14].

[11] http://oeg-upm.net/index.php/en/technologies/151-geometry2rdf
[12] https://github.com/jasaavedra/shp2geosparql/
[13] https://code.google.com/p/google-refine/
[14] http://refine.deri.ie/

Listing 1. Example RDF of building IGAC

```
<http://datos.igac.gov.co/id/catastro/igacsnc/257540000000000110039000000000> a snc:RConstruccion ;
        snc:codigoConstruccion "257540000000000110039000000000" ;
        snc:tipoConstruccion "CONVENCIONAL" ;
        snc:tipoDominio "PRIVADO" ;
        snc:numeroPisos "2"^^xsd:int ;
        snc:numeroSotanos "0"^^xsd:int ;
        snc:numeroMezanines "0"^^xsd:int ;
        snc:numeroSemisotanos "0"^^xsd:int ;
        snc:identificador "A" ;
        snc:codigoEdificacion "1"^^xsd:int ;
        snc:codigoAnteriorConstruccion "25754000000110039000" ;
        snc:area "48.1300000037695"^^xsd:double ;
        snc:perteneceA <http://datos.igac.gov.co/id/catastro/igacsnc/257540000000000110039000000000> .
        snc:hasgeometry <http://datos.igac.gov.co/id/catastro/igacsnc/Geometry/237f9dc77f73be36d9371c2420c01e37cac362ac>

<rdf:Description rdf:about="http://datos.igac.gov.co/id/catastro/igacsnc/Geometry/237f9dc77f73be36d9371c2420c01e37cac362ac">
    <rdf:type rdf:resource="http://www.opengis.net/ont/sf#MultiPolygon"/>
    <geo:asWKT rdf:datatype="http://www.opengis.net/ont/sf#wktLiteral">MULTIPOLYGON (((984419.3488769521 998229.4562988281,
</rdf:Description>
```

This RDF generation process allows transforming data of dealt datasets into structured and non-proprietary data. Besides, RDF data are generated according to a common and shared vocabulary (developed ontology network) and, therefore, they are semantically interoperable. An example of the generated RDF is shown in the Listing 1. This RDF is associated with a land parcel of Bogotá. In this RDF we include information about points of interest from LinkedGeoData and geometrical information serialized in WKT. Besides, we include topological relationships in the aforementioned RDF, due to the fact that we add information associated with block and adjoining parcels.

4.4 Links generation

We cadastral data was linked with two useful datasets, such as: LinkedGeoData and GeoNames.

Taking into account that LinkedGeoData and GeoNames data are represented by points and these datasets have no cadastral attributes, location information is the only one that we can use in the linking process. This fact does not allow using traditional linking tools, for instance, SILK framework[15]. Therefore, we perform this process using a GIS tool and spatial analysis to determine that points of GeoNames and LinkedGeodata were contained in each land parcel of the Colombian cadastral data. This analysis was performed previously RDF generation process. Thus, we generate a file with our URIs and use Google Refine and its RDF extension for setting links with the aforementioned datasets. This means that the RDF and linking generation was performed at the same time in the context of our work.

4.5 Publication

We have published our datasets using Parliament[16], which was chosen as triple store because it deals with GeoSPARQL standard and can be used to make spatial query in order to exploit our data.

When the data was loaded in our triple store, we performed some tests for checking the consistency of transformation process and quality of results. By this way, we solved minor mistakes related to RDF data and our ontology network. Some examples of the executed queries executed are shown in the Listing 2, and the Figure 2. In Listing 2 we check the consistency of the transformation process in each dataset using the developed ontology network. In this example we ask about number of buildings with more than of 5 floors within a specific area. This query allows us checking that we integrate both datasets in a right way.

Listing 2: SPARQL query for checking transformation process

```
SELECT DISTINCT
count(?id)
WHERE {
   ?id rdf:type   snc:UConstruccion.
   ?id ladm:namespace "IGAC".
   ?id snc:numeroPisos ?NoPisos.
   filter (?NoPisos > 5) .
} 73
```

```
SELECT DISTINCT
count(?id)
WHERE {
   ?id rdf:type   bog:Construccion
   ?id ladm:namespace "UAECD".
   ?id bog:numeroPisos ?NoPisos.
   filter (?NoPisos > 5) .
} 1414
```

```
SELECT DISTINCT
count (?id)
WHERE {
   ?id rdf:type ladmco:Building.
   ?id ladmco:floorNumber ?NoPisos.
   filter (?NoPisos > 5)
} 1487
```

Furthermore, due to the fact that linking process performed, we enrich our original data. This enrichment of our data is exploited through an SPARQL query (Figure 3). With this query we add to our data addresses associated with different points of interest, which are gathered from GeoNames and LinkedGeodata.

Fig 2: SPARQL query for checking links

```
SELECT
?nombre ?direccion ?nbarrio

WHERE {
   ?s bog:puntoDeInteres ?PI .
   ?s rdf:type <http://linkedgeodata.org/ontology/Shop> .
   ?s rdfs:label ?nombre .
   ?s bog:tiene ?tiene .
   ?tiene bog:direccion ?direccion.
}
 limit 20
```

nombre	direccion
"Crayon"	"CL 18A 1 11"
"Seneka"	"CL 18A 1 83"
"Terraza Pasteur"	"CL 24 6A 19"

15 http://wifo5-03.informatik.uni-mannheim.de/bizer/silk/

16 http://parliament.semwebcentral.org/

5 Conclusions and future work

In this paper we described the process followed to generate, integrate and publish cadastral Linked Data from two Colombian producers. The main goal of this work was to allow combining different sources using Linked Data principles, for overcoming current problems of information integration associated with National producers of this information type.

For achieving our goal, we have developed an ontology network, which is reusing LADM standard and GeoSPARQL ontology, and have generated an extension of Geometry2RDF tool for dealing with characteristics of our datasets. Besides, we have integrated and enriched Colombian cadastral data with two different datasets of the Linked Data cloud (GeoNames and LinkedGeodata). It demonstrates that interaction with other kinds of data is possible too.

Future work will continue integrating more datasets from other cadastral producers through Linked Data principles. We will also focus on identifying and interlinking other cadastral features with knowledge bases belonging to the Linked Open Data Initiative. Furthermore, we will work on exploitation process in order to show our cadastral Linked Data in a friendly way for final users. Finally, we will focus on improving existing metrics for linking process in order to deal with characteristics of cadastral information and increase the accuracy of this process and, therefore, existing tools.

Acknowledgements

We would like to kindly thank all members involved in this work, especially people from IGAC and Bogotá Spatial Data Infrastructures (IDECA) for their interest, help and support.

6 References

[1] Federal Geographic Data Committee. (2008). Geographic Information Framework Data Content Standard – Part 1: Cadastral. https://www.fgdc.gov/standards/projects/FGDC-standards-projects/framework-data-standard/GI_FrameworkDataStandard_Part1_Cadastral.pdf

[2] Federation International of Surveyors. (1998). Cadastre 2014: A vision for future cadastral systems. http://www.fig.net/cadastre2014/translation/c2014-english.pdf

[3] INSPIRE. (2009). Data Specification on Cadastral Parcels – Guidelines. http://inspire.jrc.ec.europa.eu/documents/Data_Specifications/INSPIRE_DataSpecification_CP_v3.0.pdf

[4] ISO. (2012) ISO 19152, Geographic information — Land Administration Domain Model (LADM). http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=51206

[5] Heath, Tom and Bizer, Christian. (2011). Linked Data: Evolving the Web into a Global Data Space. http://linkeddatabook.com/editions/1.0/

[6] Berners-Lee, T. (2006) Linked data. World Wide Web design issues. http://www.w3.org/DesignIssues/LinkedData.html.

[7] Sheth, A. P., (1998) Changing focus on interoperability in information systems: From system, syntax, structure to semantics. In Interoperating Geographic Information Systems, Goodchild, M. F., Egenhofer, M. J., Fegeas, R., Kottman, C. A. (eds), pp. 5–30, Kluwer.

[8] Goodchild, M., Egenhofer, M. J., Fegeas, R. Kottman, C. Eds. (1999) Interoperating Geographic Information Systems. The International Series in Engineering and Computer Science, Kluwer.

[9] Kuhn, W. (2005) Geospatial Semantics: Why, of What, and How? In Spaccapietra, S. Zimányi, E. (Eds.): Journal on Data Semantics III. Lecture Notes in Computer Science, 3534 (3), 1-24.

[10] Kean Huat Soon (2013) International FIG workshop on the Land Administration Domain Model. http://wiki.tudelft.nl/pub/Research/ISO19152/ImplementationMaterial/LADMOntology.owl

[11] Piña, Nelcy et. al., (2011). Ontología web semántica del registro catastral venezolano. http://cesimo.ing.ula.ve/~jacinto/websemantica/Art%C3%ADculo%20Ontolog%C3%ADa%20web%20sem%C3%A1ntica%20del%20Registro%20Catastral%20Venezolano.pdf

[12] Vilches-Blázquez, L.M., Villazón-Terrazas, B., Corcho, O., Gómez-Pérez, A.: Integrating geographical information in the Linked Digital Earth. International Journal of Digital Earth. ISSN 1753-8947. (2013).

[13] Sauermann et al. (2006). Cool URIs for the Semantic Web. http://www.dfki.unikl.de/~sauermann/2007/01/semweburisdraft/uricrisis.pdf

[14] Suarez, Maria, et al, (2010). NeOn Methodology for Building Ontology Networks. http://oa.upm.es/3879