Open Geospatial Data, Software and Standards

ORIGINAL ARTICLE Open Access

A sustainable process and toolbox for geographical linked data generation and publication: a case study with BTN100

Paola Espinoza-Arias[1]*, Miguel García-Delgado[1], Oscar Corcho[1], Pedro Vivas-White[2] and Hugo Potti-Manjavacas[2]

Abstract

We describe the process and tools that we have used to generate and publish the BTN100 Linked Dataset, based on the original data from the Spanish Topographic Base (1:100.000 scale) from the Spanish Instituto Geográfico Nacional. We have taken into account the limitations and lessons learned from our initial experience on the generation and publication of Linked Data from a range of geographical sources in Spain, in 2010, and we have now refined the process in order to facilitate: declarative mappings for the transformations from existing open data (shapefiles), automation of transformations whenever there are changes in the original data sources, version control, and alignment with INSPIRE URIs. As a result of this transformation and publication process we have also updated the reference ontology for geographical features and aligned with general ontologies such as GeoSPARQL.

Keywords: Geospatial data, Linked data, Linked dataset, Ontology

Introduction

One of the activities of the Spanish Instituto Geográfico Nacional[1] (IGN) is to produce geographical information for all the territorial entities in Spain. IGN is responsible for maintaining and making accessible cartographic and topographic databases for the representation of the Spanish territory. Their catalogs publish data related to transport networks, geodetic information, administrative units, etc. making it possible for everyone to download them from their data portal[2] under an open data license compatible with CC By 4.0[3].

Governments, via their many agencies and organizations, are constantly producing data that may be highly interrelated, but in practice become isolated data due to lack of interoperability. Cartographic and topographic information from IGN may easily enrich information from other government entities data, e.g. data from the National Institute of Statistics, Institute of Cultural Heritage, General Direction of Cadastre, Geological and Mining Institute, etc.

However, the generalized lack of use of semantics standards in the descriptions of the data elements within the data sources make it difficult to reuse them. Although progress in data availability, there are still plenty issues related to semantic interoperability; this is the ability of information systems to exchange data with unambiguous, shared meaning.

There are several initiatives around the world that have focused on generating and publishing Linked Data from a range of geospatial data sources, and a W3C/OCG Working Group was running between 2015 and 2017 with the title "Spatial Data on the Web" producing recommendations on how to publish different types of geospatial, sensor, and temporal data on the web in a principled manner. The LinkedGeoData initiative[4] aimed to make available the information collected by Open Street Map[5] as RDF and interlinks this data with other knowledges bases. Ordnance Survey Linked Data[6] publishes a number of products from the Great Britain's national mapping agency as Linked Data and provides access to them through a SPARQL endpoint. Swiss Linked Data Service[7] publishes the geospatial datasets from the Swiss Federal Spatial Data Infrastructure via a Link Data Frontend, which provides a search, querying and visualization interface.

*Correspondence: pespinoza@fi.upm.es
[1]Ontology Engineering Group, Universidad Politécnica de Madrid, Boadilla del Monte 28660, Spain
Full list of author information is available at the end of the article

In Spain, there was also some pioneering work on producing geospatial Linked Data from a range of data sources (many of them from IGN), as described in [1, 2]. Additionally, an ontology of administrative units has been created and published at http://vocab.linkeddata.es/datosabiertos/def/sector-publico/territorio and some regions have produced Linked Data about their administrative units, such as Aragón [3]. In 2010, we worked on the GeoLinked Data initiative in order to enrich the web of data with Spanish Linked Data. We used as input several relational data bases and Excel spreadsheets about administrative units, hydrography and statistical domains. Then we modeled an ontology to represent this data. Later, we generated the RDF with the Geomety2RDF plugin[8] to deal with the geometrical transformations. The generated RDF was compliant with the WSG84 vocabulary[9] and the GML ontology[10]. We added some links between terms from each data source, published the resulting RDF in a triplestore and made it available via a visualization tool (Map4RDF[4]). Nevertheless, our process had some limitations. Geometries were not making use of GeoSPARQL, since it was an emerging standard with little tool support. Our transformations were originated from special access to Oracle Spatial databases, instead of already published open data. No automation was included to deal with the evolution of the data sources, what made the Linked Data state quickly. Manual intervention was needed for this update process.

In this paper we describe the transformation process of Spanish Topographic Base in scale 1:100.000 (BTN100) catalog into Linked Data. We used Github[11] for version control and as archival to store all the RDF transformations. The process includes defining the semantic model for this geospatial data, generating the data transformations, publishing them as a SPARQL endpoint, and maintaining the Linked Dataset.

Our work represents a forward step to improve semantic interoperability in the geospatial domain. We make use of the open BTN100 dataset and define semantics for its data. Through the semantic model we represents complex geometrical shapes, e.g. multi-lines, polygons, multi-polygons, etc. This makes it possible to visualize these geometric entities and to infer assertions such as those related to any geographical entity being embedded within another entity. In addition, we provide an automatic way to deal with changes in the data source in order to provide an always up-to-date Linked Dataset.

The paper is organized as follows: "Methodology and results" section describes all the followed steps for generate and maintain the Linked Dataset. "Conclusions and future directions" section shows the conclusions and future work.

Methodology and results

Generating Linked Data is a process that involves several activities and decisions in order to obtain a high-quality Linked Dataset. We followed the Methodological Guidelines for Publishing Government Linked Data [5]. These guidelines cover all the steps and details that are necessary for the activities involved. The activities described by the guidelines are: specification, modelling, generation, publication and exploitation. Each activity involves one or more tasks and some techniques for carrying out them.

Specification

The first task of this activity is focused on the identification of the data sources, formats, information within the datasets and general requirements for the resulting Linked Dataset. In our case, we used the open BTN100 catalog as our data source. This data source is available from the Spanish National Center for Geographic Information[12] (CNIG) as shapefiles in the ETRS89 and REGCAN95 Coordinate Reference System (CRS). The BTN100 catalog contains geographic information about topographic and thematic data; it was designed following the INSPIRE Directives[13]. It clusters the data in the following themes: administrative units, protected zones, buildings and population entities, transport networks, energy and conduction, geodetic vertices, altimetry and hydrography.

A URI design task is involved in this activity. In our case, we defined the persistent URIs for our features according to the Spatial Data on the Web best practices described in [6] and the Technical Interoperability Standard [7]. The base URI structure for all elements is https://datos.ign.es. We followed an upper camel case strategy to name classes and a lower camel case strategy for object and data properties and resources. In Table 1 we present our URIs design.

The final task of this activity is the definition of the license of the Linked Dataset. We decided to reuse the IGN license[14] for the BTN100 Linked Dataset. It is a Creative Common Attribution 4.0 International (CC BY 4.0) license.

Table 1 Defining the URIs

Element	URI
Ontology	https://datos.ign.es/def/{ontology_name}
Class	https://datos.ign.es/def/{ontology_name}#{Class}
Object property	https://datos.ign.es/def/{ontology_name}#{objectProperty}
Data property	https://datos.ign.es/def/{ontology_name}#{dataProperty}
skos	https://datos.ign.es/kos/{theme}/{SKOS_name}
Resource	https://datos.ign.es/recurso/{linked_dataset_name}/{resource_type}/{resource_identifier}

Modelling

In order to represent all themes of the dataset, we generated an ontology, which replaces the former http://geo.linkeddata.es ontology. The ontology development was made by following the LOT Methodology described online[15] (originally proposed in [8]) and used for example in [9]. Reusing ontologies was important through our development process. We focused our analysis on the common spatial vocabularies recommended by the W3C Working Group Note [6]. We decided to reuse the GeoSPARQL vocabulary[16] to represent geospatial data, since it makes it possible to use specialized functions for geometries.

The GeoSPARQL vocabulary does not allow representing elements such as identifiers for resources, labels for geographical objects, altitude, etc. In order to address these shortcomings, we developed the *btn100* ontology[17]. This model represents all the geographical objects from our dataset. The *btn100* has links to SKOS thesauri that were developed to represent some categories of elements in our dataset; for instance, type of highway access, type of roadway, etc. These thesauri will be linked in the future to these maintained in the INSPIRE registry. All files generated during the ontology development, including the requirements, ontologies, thesauri and documentation are available into a Github repository at https://github.com/oeg-upm/ontology-BTN100.

In Fig. 1 we show the general ontology model. The esam[18] and escj[19] ontologies are included in our model because they are reused in order to represent Administrative Units and Streets respectively.

In Fig. 2a we show an extract from the classes of the *btn100* ontology. As is depicted, all classes are a subclass of *ObjetoGeografico* which is equivalent to *Feature* class from the GeoSPARQL vocabulary. This equivalence is defined in *geo_core* ontology[20] which reuses the GeoSPARQL vocabulary and defines several data properties that are common for all concepts of the BTN100 catalog. Finally, on the right side of Fig. 2a we present the main metrics of the *btn100* which includes a summary of the total number of axioms, classes, properties, etc.

The model depicted in Fig. 2b presents the classes defined in order to represent all concepts from the administrative units and protected zones themes from those mentioned before in Specification subsection. The *btn100* documentation in HTML format, including further details of the representation of the other themes and their diagrams, is available at https://datos.ign.es/def/btn100.

In Fig. 3 we present an example about the URIs definition for *btn100*. This illustration represents that: La Autopista "AP2-E50" tiene acceso de tipo peaje (The "AP2-E50" highway has toll access).

Generation

In this activity we followed the process depicted in Fig. 4 in order to generate the Linked Dataset. At the beginning, we extracted all shapefiles from the BTN100 data source and unzip them in order to obtain the shapefiles for each theme. Then we made some data transformations and obtained as result a RDF file. Finally, we linked the resulting dataset to other resources and obtained the Linked Dataset. As introduced before, we stored all files

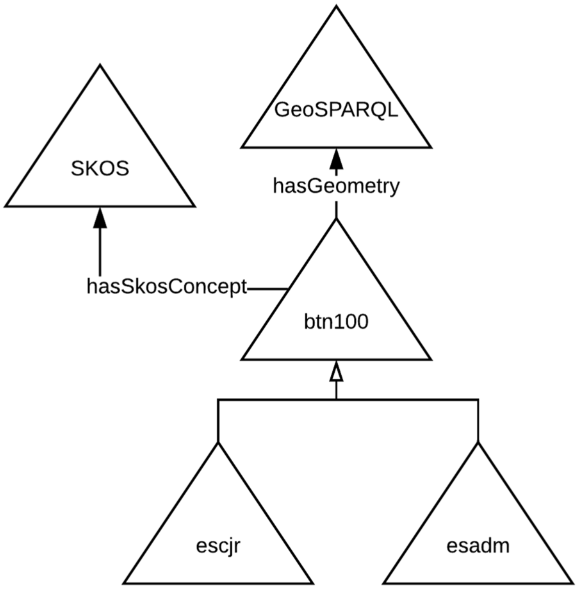
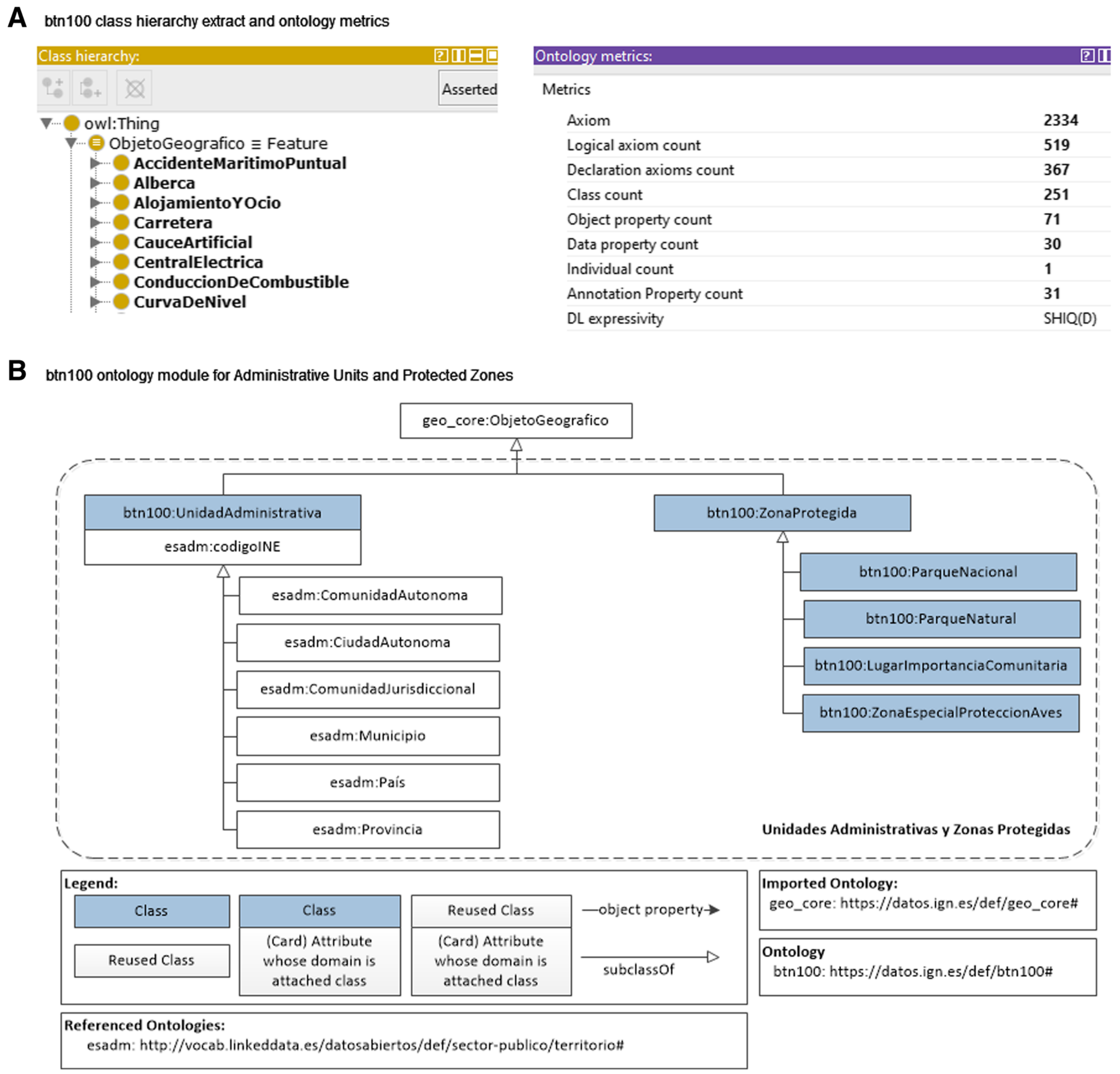


Fig. 1 btn100 general model. Ontology model overview which includes the ontologies reused by the btn100 ontology and the relations between them


A
btn100 class hierarchy extract and ontology metrics
Class hierarchy:
Asserted
owl:Thing
ObjetoGeografico ≡ Feature
AccidenteMaritimoPuntual
Alberca
AlojamientoYOcio
Carretera
CauceArtificial
CentralElectrica
ConduccionDeCombustible
CurvaDeNivel
Ontology metrics:
Metrics

Metrics	
Axiom	2334
Logical axiom count	519
Declaration axioms count	367
Class count	251
Object property count	71
Data property count	30
Individual count	1
Annotation Property count	31
DL expressivity	SHIQ(D)

B
btn100 ontology module for Administrative Units and Protected Zones
geo_core:ObjetoGeografico
btn100:UnidadAdministrativa
esadm:codigoINE
esadm:ComunidadAutonoma
esadm:CiudadAutonoma
esadm:ComunidadJurisdiccional
esadm:Municipio
esadm:País
esadm:Provincia
btn100:ZonaProtegida
btn100:ParqueNacional
btn100:ParqueNatural
btn100:LugarImportanciaComunitaria
btn100:ZonaEspecialProteccionAves
Unidades Administrativas y Zonas Protegidas
Legend:
Class
Reused Class
Class
(Card) Attribute whose domain is attached class
Reused Class
(Card) Attribute whose domain is attached class
object property
subclassOf
Imported Ontology:
geo_core: https://datos.ign.es/def/geo_core#
Ontology
btn100: https://datos.ign.es/def/btn100#
Referenced Ontologies:
esadm: http://vocab.linkeddata.es/datosabiertos/def/sector-publico/territorio#

Fig. 2 a Class hierarchy extract and metrics, and **b** administrative units and protected zones module of the btn100 ontology. *btn100* ontology details about class hierarchy, metrics, and the model of the module about two themes from the BTN100 catalog

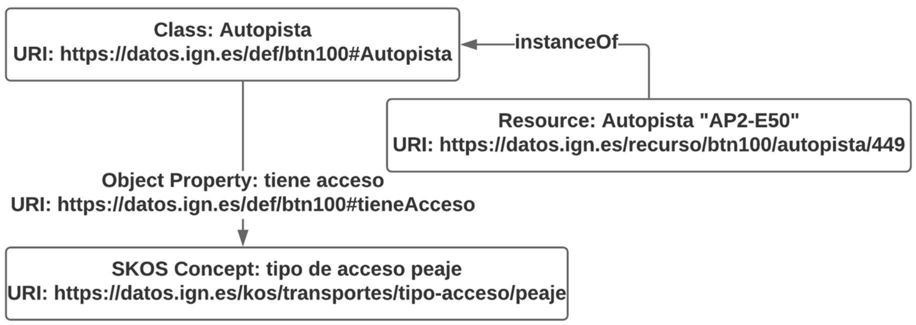

Class: Autopista
URI: https://datos.ign.es/def/btn100#Autopista
instanceOf
Resource: Autopista "AP2-E50"
URI: https://datos.ign.es/recurso/btn100/autopista/449
Object Property: tiene acceso
URI: https://datos.ign.es/def/btn100#tieneAcceso
SKOS Concept: tipo de acceso peaje
URI: https://datos.ign.es/kos/transportes/tipo-acceso/peaje

Fig. 3 URIs definition example. The example provides the URIs structure in order to represent the entities and relations of a highway

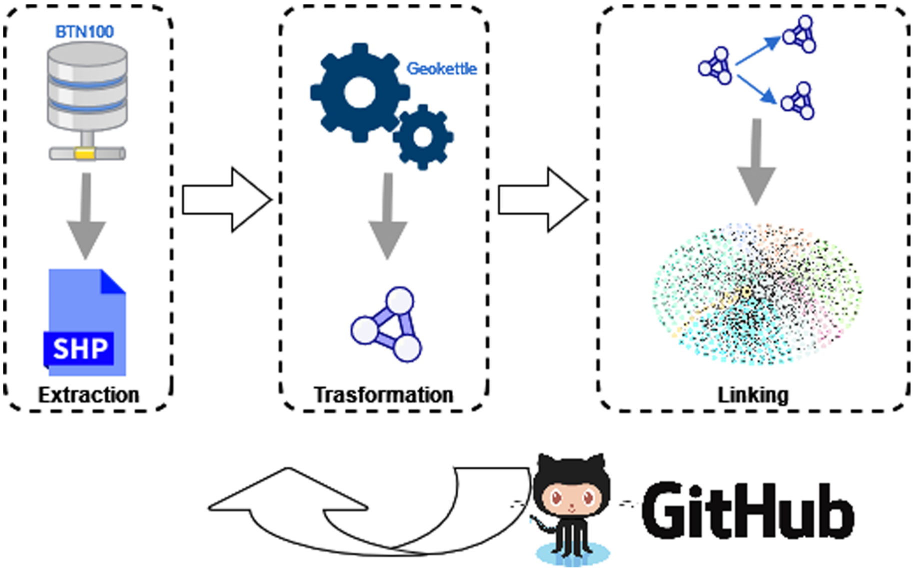


Fig. 4 Geospatial Linked Data generation process. The three step process followed in order to generate the Linked Data

involved in this process into a Github repository available at https://github.com/oeg-upm/btn100.

In order to deal with the transformation tasks we used GeoKettle[21]. With this tool we created a transformation file for each shapefile and configured a workflow to perform the activities described as follows. First, we cleaned the data, e.g. correcting malformed/incompatible datatypes. Then, we mapped the data to their corresponding equivalents in the SKOS concepts. After, we converted ETRS89 and REGCAN95 into WGS84 CRS in order to represent data in the GeoSPARQL standard. Last, we transformed the data, via TripleGeo plugin[22], into triples according to the model defined in the *btn100* ontology.

TripleGeo converts the geospatial features into a RDF serialization, in our case into Turtle[23] files. We enable TripleGeo as a Geokettle plugin[24] in order to provide an accurate generation of the semantic information; for instance, a correct URI definition. In Fig. 5 we depict an example of the TripleGeo structure. At the top of the TripleGeo window we can set the type and URI for a resource. At the bottom we can set prefixes and URIs for the fields available in the shape file. Further details about TripleGeo configuration parameters are available in their wiki[25].

Finally, for the linking task we used the *owl:sameAs* relationship to align our resources with DBpedia[26] and other resources from the Spanish government open data portal[27]. All files generated during the linking task are also available at the Github repository.

Publication

This step aims to provide access to the resulting dataset. We stored the RDF files into a Virtuoso triplestore[28]. Virtuoso provides a SPARQL endpoint, available at https://datos.ign.es/sparql. Some use cases with their SPARQL queries are available at https://datos.ign.es/casos-de-uso.html.

We also provided a web interface to the SPARQL endpoint via Pubby[29]. A web portal about this work is available at https://datos.ign.es; it delivers a single entry point to all the resources (e.g queries, ontologies, skos, etc.).

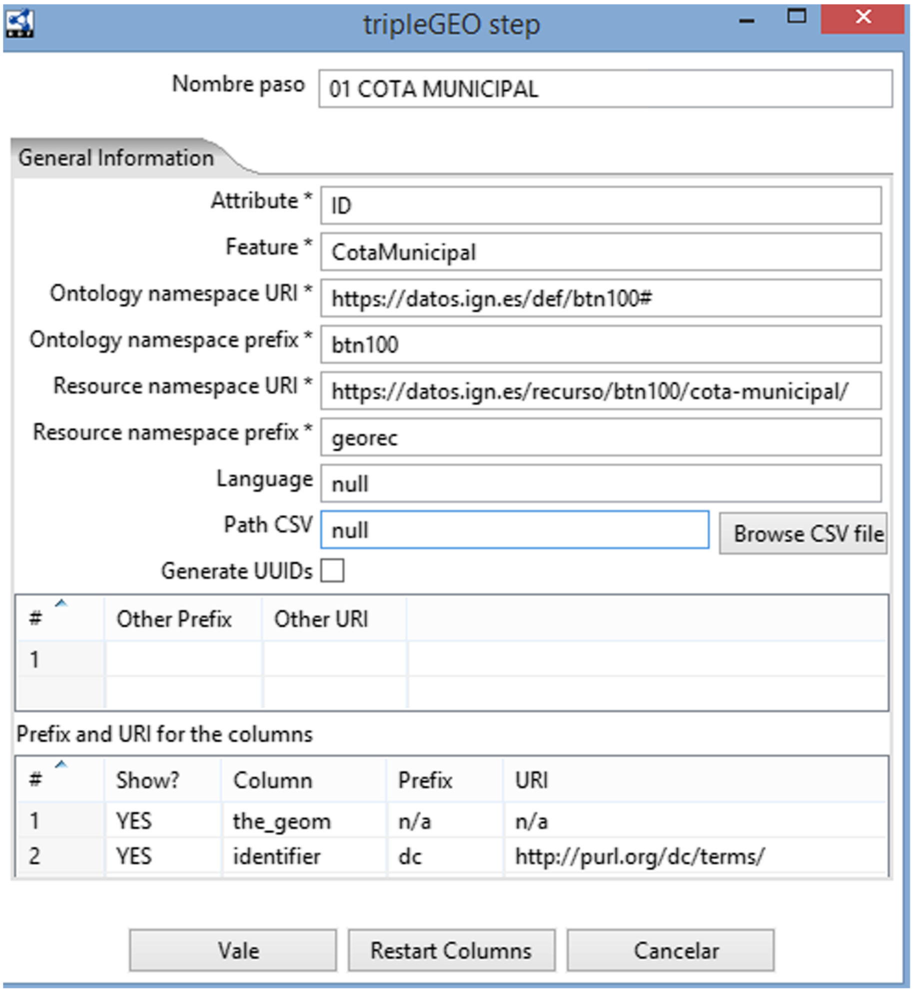
The metadata for our dataset is based on the BTN100 metadata file[30] from the Spanish government open data portal. This file was generated according to the DCAT vocabulary[31].

Exploitation

We are displaying the dataset over a map using Map4RDF[32]. This tool allows end users to visualize and interact with our Linked Dataset. Map4RDF connects to our endpoint in order to provide the faceted browser interface for each BTN100 theme and all their concepts. When a user selects a facet, Map4RDF queries our triplestore and provides the visualization for the instances of the selected facet including their respective GeoSPARQL geometries. The instance for visualization is available at http://certidatos.ign.es/map/. In Fig. 6 we show a visualization example. We exemplify the Spanish provinces painted at the map.

Maintenance

Despite a maintenance activity is not included in the followed guidelines, we considered it is important to ensure the dataset will always have the most current version of the data source. In order to automatize the generation and



Nombre paso: 01 COTA MUNICIPAL

Attribute *: ID

Feature *: CotaMunicipal

Ontology namespace URI *: https://datos.ign.es/def/btn100#

Ontology namespace prefix *: btn100

Resource namespace URI *: https://datos.ign.es/recurso/btn100/cota-municipal/

Resource namespace prefix *: georec

Language: null

Path CSV: null

Generate UUIDs ☐

#	Other Prefix	Other URI
1		

Prefix and URI for the columns

#	Show?	Column	Prefix	URI
1	YES	the_geom	n/a	n/a
2	YES	identifier	dc	http://purl.org/dc/terms/

Fig. 5 TripleGeo configuration for RDF generation. An example of the configuration structure for a specific feature

updating of the Linked Dataset we developed a Pyhton script[33].

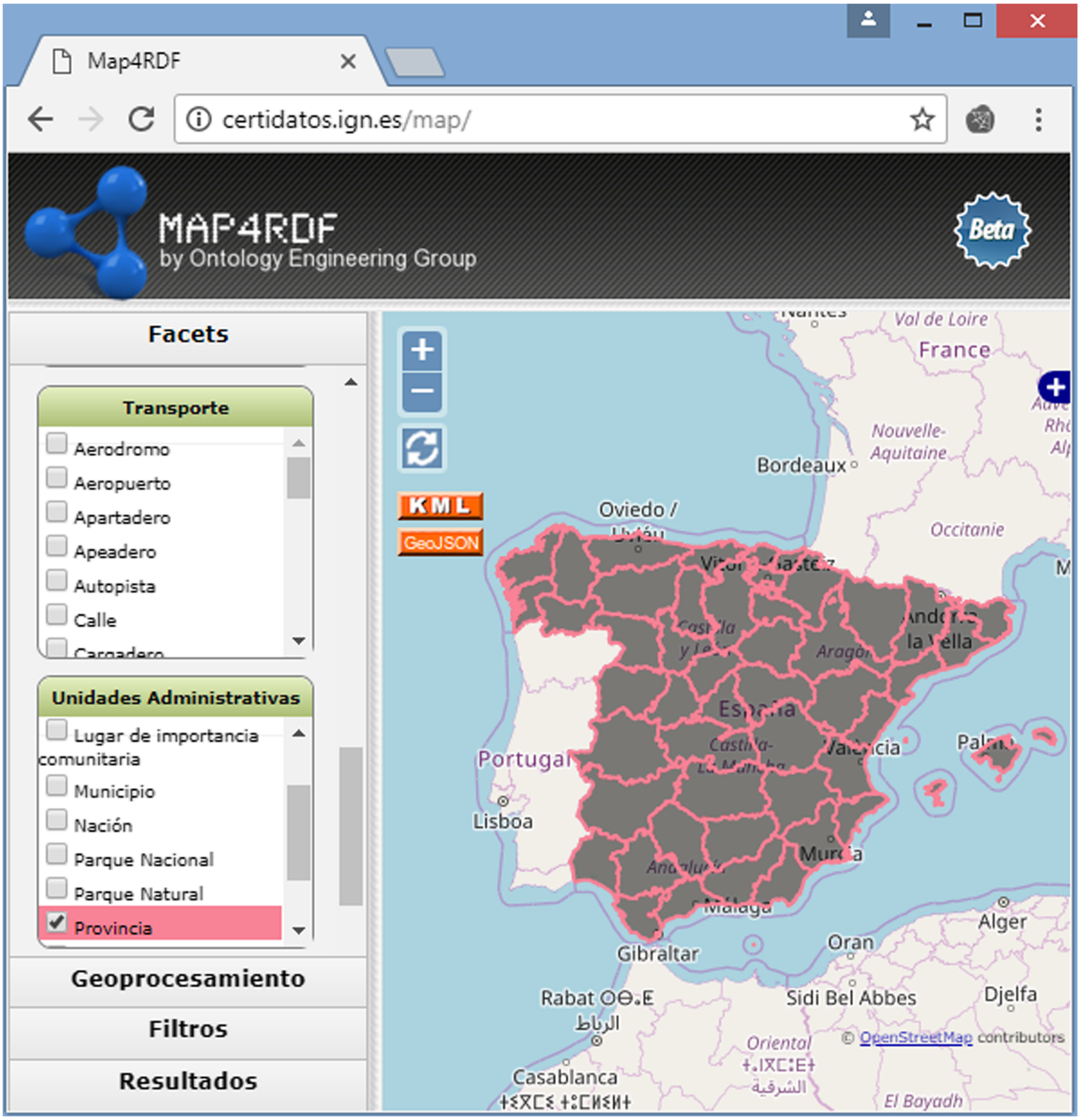
The script, which will be periodically executed, starts by downloading the BTN100 data source, then it makes a testing process between the downloaded source and the previous one. If a change is detected, the script identifies the elements that need to be updated. Then it generates, via GeoKettle, the new RDF files and publishes the updates in the SPARQL endpoint. Finally the script also updates the thesauri and sameAs files in the triplestore.

As we mentioned, Github is used as our environment to deal with file versioning and storing. However, Github does not allow to push files larger than 100 MB. For this reason, the script makes another test in order to check if some of the updated data sources has more than 90 MB. If there is a file with this condition, the script breaks it down into files up to 90 MB and then uploads the resulting files into Github.

Conclusions and future directions

In this paper we have described our updated approach for the previous Spanish GeoLinked Data work, specifically for the representation of the BTN100 catalog. We have presented the process to generate and publish the BTN100 as Linked Data. The dataset has been generated by using the *btn100* ontology, which reuses GeoSPARQL vocabulary. This ontology provides complex geospatial representations and makes data more interoperable with other similar datasets. Our dataset has been tested against competency questions posed by domain experts; it is modular and therefore easily extensible.

In this work we have entirely supported the process by Github, in order to provide a collaborative, distributed and version development tool. Our work also provided an automatic script in order to perform the whole process, from data extraction to publication. This script allows updating the dataset whenever a change is detected in the data source.



Transporte

- □ Aerodromo
- □ Aeropuerto
- □ Apartadero
- □ Apeadero
- □ Autopista
- □ Calle
- □ Cargadero

Unidades Administrativas

- □ Lugar de importancia comunitaria
- □ Municipio
- □ Nación
- □ Parque Nacional
- □ Parque Natural
- ☑ Provincia

Fig. 6 Visualization example. The Spanish provinces are depicted in the map according to the filter selection from the Administrative Units facet

Our model represents all the BTN100 themes; however, we are only generating linked data for territorial units. We are using Virtuoso as the technology behind our endpoint; this is due to our previous experience with this technology. However, Virtuoso does not fully support GeoSPARQL functions; it would be important to complement our work with another triplestore in this domain. Our work addresses a specific need from the IGN, it is available in Spanish. However, the approach is applicable to other scenarios in this domain and not possibly in other languages.

Endnotes

[1] http://www.ign.es/web/ign/portal

[2] http://centrodedescargas.cnig.es/CentroDescargas

[3] https://creativecommons.org/licenses/by/4.0

[4] http://linkedgeodata.org

[5] https://www.openstreetmap.org

[6] http://data.ordnancesurvey.co.uk

[7] https://www.geo.admin.ch/en/geo-services/geo-services/linkeddata.html

[8] http://mayor2.dia.fi.upm.es/oeg-upm/index.php/en/technologies/151-geometry2rdf/index.html

[9] https://www.w3.org/2003/01/geo/wgs84_pos

[10] http://schemas.opengis.net/gml/3.2.1/gml_32_geometries.rdf

[11] https://github.com

[12] http://centrodedescargas.cnig.es/CentroDescargas

[13] http://inspire.ec.europa.eu

[14] http://www.ign.es/resources/licencia/Condiciones_licenciaUso_IGN.pdf

[15] http://lot.linkeddata.es

[16] http://www.opengis.net/ont/geosparql

[17] https://datos.ign.es/def/btn100

[18] http://vocab.linkeddata.es/datosabiertos/def/sector-publico/territorio

[19] http://vocab.linkeddata.es/datosabiertos/def/urbanismo-infraestructuras/callejero

[20] https://datos.ign.es/def/geo_core

[21] http://www.spatialytics.org/projects/geokettle

[22] https://github.com/GeoKnow/TripleGeo

[23] https://www.w3.org/TR/turtle

[24] https://github.com/oeg-upm/btn100/tree/master/tripleGeoplugin

[25] https://github.com/oeg-upm/geo.linkeddata.es-TripleGeoKettle/wiki/How-It-Works#configuration

[26] http://dbpedia.org

[27] http://datos.gob.es

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

[29] http://wifo5-03.informatik.uni-mannheim.de/pubby

[30] http://datos.gob.es/apidata/catalog/dataset/e00125901-base-topografica-nacional-1-100-000.rdf

[31] https://www.w3.org/TR/vocab-dcat

[32] http://oeg-upm.github.io/map4rdf

[33] https://github.com/oeg-upm/btn100/tree/master/proceso-actualizacion

Acknowledgments
We acknowledge the work done by developers who have contributed parts of the software used: Lissete Moscoso, Francisco Siles, Victor Saquicela and Luis Vilches.

Funding
This work has been funded by Centro Nacional de Información Geográfica and DATOS 4.0: RETOS Y SOLUCIONES - UPM Spanish national project (TIN2016-78011-C4-4-R).

Availability of data and materials
All files generated during this work are available in our GitHub repositories. Ontologies and SKOS thesuauri files are available at https://github.com/oeg-upm/ontology-BTN100. Original and transformed files, plugin and scripts are available at https://github.com/oeg-upm/btn100.

Authors' contributions
All authors participated fully in this work from inception to completion. All authors read and approved the final manuscript.

Competing interests
The authors declare that they have no competing interests.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Author details
[1]Ontology Engineering Group, Universidad Politécnica de Madrid, Boadilla del Monte 28660, Spain. [2]Centro Nacional de Información Geográfica, Madrid 28003, Spain.

Received: 17 January 2019 Accepted: 20 February 2019
Published online: 13 March 2019

References

1. de León A, Saquicela V, Vilches LM, Villazón-Terrazas B, Priyatna F, Corcho O. Geographical linked data: a spanish use case. In: Proceedings of the In I-SEMANTICS '10 6th International Conference on Semantic Systems. New York: ACM; 2010.
2. Atemezing G, Corcho O, Garijo D, Mora J, Poveda-Villalón M, Rozas P, Vila-Suero D, Villazón-Terrazas B. Transforming meteorological data into linked data. Semant Web. 2013;4(3):285–90.
3. Corcho O, Pérez IS, Lafuente H, Portolés D, Cano C, Peris A, Subero JM. Publishing linked statistical data: Aragon, a case study. In: Joint Proceedings of the International Workshops on Hybrid Statistical Semantic Understanding and Emerging Semantics, and Semantic Statistics (HybridSemStats). Aachen; 2017.
4. de León A, Wisniewki F, Villazón-Terrazas B, Corcho O. Map4rdf - faceted browser for geospatial datasets. In: Using Open Data: policy modeling, citizen empowerment, data journalism. 2012.
5. Villazón-Terrazas B, Vilches-Blázquez LM, Corcho O, Gómez-Pérez A. Methodological guidelines for publishing government linked data. In: Linking Government Data. New York: Springer; 2011. p. 27–49.
6. Spatial Data on the Web Best Practices. 2017. https://www.w3.org/TR/sdw-bp/.
7. Technical Interoperability Standard for the Reuse of Information Resources. 2013. https://administracionelectronica.gob.es/pae_Home/dam/jcr:a8d2c143-ce9a-4fc7-afe7-ef5d9ba7c4a1/ENGLISH_Interoperability_Agreement_for%20the%20Reuse%20%20of%20Information%20Resources.pdf.
8. Poveda-Villalón M. A reuse-based lightweight method for developing linked data ontologies and vocabularies. In: 9th Extended Semantic Web Conference (ESWC). Berlin: Springer; 2012. p. 833–7.
9. Radulovic F, Poveda-Villalón M, Vila-Suero D, Rodríguez-Doncel V, Garcí-Castro R, Gómez-Pérez A. Guidelines for linked data generation and publication: An example in building energy consumption. Autom Constr. 2015;57:178–87.