

Publishing Linked Data - There is no One-Size-Fits-All Formula

Boris Villazón-Terrazas, Daniel Vila-Suero, Daniel Garijo,
Luis M. Vilches-Blázquez, María Poveda-Villalón,
José Mora, Oscar Corcho, and Asunción Gómez-Pérez

OEG-DIA, FI, Universidad Politécnica de Madrid, Spain
{bvillazon,dvila,dgarijo,lmvilches,mpoveda,jmora,ocorcho,asun}@fi.upm.es

Abstract. Publishing Linked Data is a process that involves several design decisions and technologies. Although some initial guidelines have been already provided by Linked Data publishers, these are still far from covering all the steps that are necessary (from data source selection to publication) or giving enough details about all these steps, technologies, intermediate products, etc. Furthermore, given the variety of data sources from which Linked Data can be generated, we believe that it is possible to have a single and unified method for publishing Linked Data, but we should rely on different techniques, technologies and tools for particular datasets of a given domain. In this paper we present a general method for publishing Linked Data and the application of the method to cover different sources from different domains.

Key words: Linked Data, Publishing Linked Data

1 Introduction and Motivation

So far, Linked Data principles and practices are being adopted by an increasing number of data providers, getting as result a global data space on the Web containing hundreds of LOD datasets [3]. Moreover, Linked Data generation and publication does not follow a set of common and clear guidelines to scale out the generation and publication of Linked Data.

Furthermore, given the variety of data sources from which Linked Data can be generated, we believe that it is possible to have a single and unified method for publishing Linked Data, but we should rely on different techniques, technologies and tools for particular datasets of a given domain. The rest of the paper is organized as follows: Section 2 introduces our method for publishing linked data, then, Section 3 describes the application of the method to cover different sources from different domains, and finally, Section 4 presents the conclusions.

2 A Method for Publishing Linked Data

In previous work [6] we have already presented the Linked Data Generation Process as one that follows an iterative incremental life cycle model. In that

work we proposed a method that covers the following activities (1) specification, for analyzing and selecting the data sources, (2) modelling, for developing the model that represents the information domain of the data sources, (3) generation, for transforming the data sources into RDF, (4) linking, for creating links between the RDF resources, of our dataset, with other RDF resources, of external datasets, (5) publication, for publishing the model, RDF resources and links generated, and (6) exploitation, for developing applications that consume the dataset. Each activity is decomposed into one or more tasks, and some techniques, technologies and tools are provided for carrying out them. It is worth mentioning that the order of the activities and tasks might be changed base on particular needs of the data owners and publishers. Moreover, we are continuously getting feedback about this method, and therefore, we are improving it constantly. Figure 1 depicts the main activities.

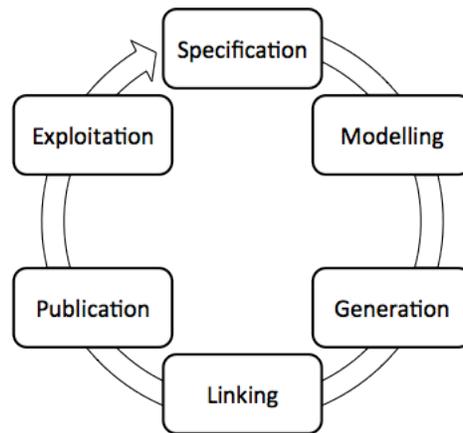


Fig. 1. Main Activities for Publishing Linked Data (extended from [6])

3 Application of the Method to Different Domains

In this section we present the application of the method to cover different sources from different domains.

3.1 GeoLinkedData

GeoLinkedData¹ aims at enriching the Web of Data with Spanish geospatial data into the context of INSPIRE themes². This initiative has started off by

¹ <http://geo.linkeddata.es/>

² The INSPIRE Directive addresses 34 spatial data themes needed for environmental applications. <http://inspire.jrc.ec.europa.eu/index.cfm/pageid/2/list/7>

publishing diverse information sources belonging to the National Geographic Institute of Spain, IGN, and the National Statistic Institute in Spain, INE. Such sources are made available as RDF knowledge bases according to the Linked Data principles [3].

Regarding the *specification activity*, the data sources taken for this project are composed of (1) a set of spreadsheets published by the Spanish Statistical Institute (INE), (2) a set of geospatial relational databases, Oracle and MySQL, from the Spanish Geographical Institute (IGN). Within this activity we also have analyzed the data sources and defined the URI design for the resources.

After the specification activity, we need to define the ontologies to be used for *modelling* the domain of those data sources. We have created a modular ontology network³ by reusing available knowledge resources. The ontology network consist of the following subontologies (1) SCOVO⁴ for modelling the statistical data, (2) FAO Geopolitical Ontology⁵ for representing the Spanish administrative division, (3) hydrOntology⁶ for describing the hydrographical domain, (4) WGS84⁷ vocabulary for representing information about spatially-located things, and (5) Time Ontology⁸ for describing temporal concepts.

As for the *generation activity* we have used (1) ODEMapster⁹ for transforming relational database content into RDF, (2) NOR2O¹⁰ for converting spreadsheets into RDF, and (3) geometry2rdf¹¹ for transforming geospatial databases into RDF. It is worth mentioning that the geometry stored in the databases were simple (stored in normal database columns) and complex (stored in BLOB columns).

Regarding the *linking activity*, we have used Silk¹²[1] for discovering *owl:sameAs* links between our dataset and external datasets, i.e., GeoNames¹³ and DBpedia¹⁴. We also used sameAs validator¹⁵, an application for validating the links discovered by Silk.

Once we had the legacy data transformed into RDF, we needed to store and publish the data in a triplestore. In our scenario we decided to rely on Virtuoso

³ <http://geo.linkeddata.es/web/guest/modelos>

⁴ <http://vocab.deri.ie/scovo>

⁵ <http://www.fao.org/countryprofiles/geoinfo.asp>

⁶ <http://www.oeg-upm.net/index.php/es/ontologies/107-hydrontology>

⁷ <http://www.w3.org/2003/01/geo/>

⁸ <http://www.w3.org/TR/owl-time/>

⁹ <http://www.oeg-upm.net/index.php/en/technologies/9-r2o-odempaster>

¹⁰ <http://www.oeg-upm.net/index.php/en/technologies/57-nor2o>

¹¹ <http://www.oeg-upm.net/index.php/en/technologies/151-geometry2rdf>

¹² <http://www4.wiwiw.fu-berlin.de/bizer/silk/>

¹³ <http://geonames.org/>

¹⁴ <http://dbpedia.org/About>

¹⁵ <http://oegdev.dia.fi.upm.es:8080/sameAs/>

Universal Server¹⁶ and Pubby¹⁷. For the metadata information we have used VoID¹⁸ for expressing the metadata about our dataset.

Finally, for the *exploitation* we have developed a prototype application¹⁹ that browses and visualizes the RDF data.

3.2 AEMETLinkedData

In AEMETLinkedData²⁰ we have generated Linked Data from Spanish meteorological data. Within this initiative we are publishing information resources from the *Agencia Estatal de Meteorología* (Spanish Meteorological Office, AEMET), as Linked Data.

Within the *specification activity*, among all of the data made available in the FTP server from AEMET, we have focused on surface meteorological observing stations, and more precisely in measurements taken in ten minute interval times. AEMET has around 250 automatic weather stations of this network, registering pressure, temperature, humidity, precipitation and wind data every 10 minutes. Data from the different stations are publicly available online in comma separated values (CSV) files, compressed with gzip, updated every hour and kept for seven days. This means, every hour six new files are added, corresponding with periods of ten minutes, and every day a new folder is created to store the files for that day.

As for the *modelling*, the development of the AEMET ontology network²¹ has been performed following an iterative approach based on the reuse of existing knowledge resources, both ontological (including ontologies and Ontology Design Patterns) and non-ontological resources as proposed by the NeOn methodology [5]. The AEMET ontology network follows a modular structure consisting of a central ontology that links together a set of subontologies that describe different sub domains involved in the modelling of the meteorological measurements. The subontologies are (1) Measurement ontology that models the knowledge related to meteorological observations and reuses the SSN Ontology²², (2) Location ontology that models knowledge about locations and reuses the WGS84²³ vocabulary, such as administrative limits and coordinates, (3) Time ontology that reuses the W3C Time Ontology²⁴ and models knowledge about time such as temporal units, temporal entities, instants, intervals, etc., and (4) Sensor ontology that models the network of sensors and weather stations.

¹⁶ <http://virtuoso.openlinksw.com/dataspace/dav/wiki/Main/>

¹⁷ <http://www4.wiwiw.fu-berlin.de/pubby/>

¹⁸ <http://www.w3.org/TR/void/>

¹⁹ <http://geo.linkeddata.es/web/guest/visualizacion-beta>

²⁰ <http://aemet.linkeddata.es/>

²¹ http://aemet.linkeddata.es/models_en.html

²² http://www.w3.org/2005/Incubator/ssn/wiki/Report_Work_on_the_SSN_ontology

²³ <http://www.w3.org/2003/01/geo/>

²⁴ <http://www.w3.org/TR/owl-time/>

The *RDF generation* is produced with ad-hoc Python scripts²⁵. These scripts are executed in two processing steps, integrating with ease the generation of RDF and tasks as crawling the FTP server where the CSV files are located. The first step generates the information about the automatic stations, which is static, and thus needs to be executed only once. The second step generates the information about the observations on a regular basis, keeping the data updated.

Regarding the *publication*, the generated RDF is stored in Virtuoso, which integrates with Pubby for the publication of the results and makes them available for humans as well as computers.

As for the *exploitation*, we have developed a simple application²⁶ that consists on a viewer based on a modified version of map4rdf²⁷, a tool that uses Google Web Toolkit framework to visualize and access Linked Data resources. Each automatic weather station is geolocated, enabling its representation in the map with a marker. Since there are approximately 300 stations, they can be all represented at the same time.

3.3 El Viajero

El Viajero is a dataset that exposes descriptive metadata and provenance of news, blogs and posts belonging to the Prisa Digital Group²⁸ in the domain of travel journalism. El Viajero makes use of several heterogeneous datasets, and it is integrated within a service which exploits it in order to help users navigating and browsing the contents of the dataset.

Regarding the *specification*, the data sources analyzed come from a variety of newspapers and digital platforms belonging to the Prisa Digital Group: “Suplemento El País”, “Guías Aguilar”, and “Canal Viajar”. Users can also create their own blogs to post their traveling experiences (around 600 different blogs exist), apart from commenting the news posted by the reporters of the group. The news are stored using an extension over the International Press Communications Council (IPTC)²⁹ Standard called News Industry Text Format (NITF)³⁰, which is also used by other major publishing agencies in Europe, such as AFP³¹, ANSA³² or AP Digital³³. The data is separated from the metadata in different kinds of files: one for the structure of the news, another for the content of each piece of news, another one for the metadata of the sources of the news like images or videos and another one for storing statistics about the news (such as views, date of the last view, etc). The blogs and posts are stored in two different private databases: one with metadata about the posts (creator, number of comments,

²⁵ http://aemet.linkeddata.es/technology_en.html

²⁶ http://aemet.linkeddata.es/browser_en.html

²⁷ <http://oegdev.dia.fi.upm.es/map4rdf/>

²⁸ <http://www.prisa.com>

²⁹ <http://www.iptc.org/site/Home/>

³⁰ <http://www.iptc.org/cms/site/index.html?channel=CH0107>

³¹ <http://www.afp.com/>

³² <http://www.ansa.it/>

³³ <http://www.apdigitalnews.com/>

body, date of creation, etc.) and the other one with metadata about the blogs (title, header image, description, owner, etc.).

The resources have been *modeled* following a layered approach. In the first level DC³⁴, SIOC³⁵, MPEG7³⁶ and WGS84³⁷ vocabularies are used for describing the metadata of the different resources on the platform (e.g., post metadata, photo and video metadata, location metadata, etc.), but not their evolution. In the second level, our ontology³⁸ connects the previous vocabularies to the Open Provenance Model (OPM) [4] by extending the OPM core ontology (OPM-O³⁹) in the third layer.

As for the *generation activity*, we have developed and reused parsers to produce RDF from the Prisa Digital datasets automatically. The system has been developed using the JENA Framework, combined with the TDB and Virtuoso repositories for the proper publication of the dataset. The data is exposed using Pubby, and additional metadata about the query used to retrieve the exposed data is captured with the Provenance Vocabulary (along with metadata of the dataset itself).

Since *El Viajero* belongs to a Spanish media company, many of the guides' locations refer to Spanish locations. GeoLinkedData contains accurate metadata about the majority of these locations, so we have linked to this dataset the guides within the Spanish territory. For the guides referring to the rest of the world we linked to DBpedia, the most linked dataset in the Linked Data cloud, and which provides many additional details about the locations. We relied on the Silk framework for finding the links to these two datasets, curating manually the results obtained by the tool. Additional links from travel recommendations have been linked to the Linked User Feedback (LUF) dataset⁴⁰.

Thanks to the publication of *El Viajero* as Linked Data, users have available more than 6600 resources about almost 1000 locations around the world. In order to help them organize, browse and explore the contents of the dataset, we have developed an application [2] to expose these contents in an eye catching way by representing them in a map.

3.4 datos.bne.es

datos.bne.es publishes bibliographic data from the Spanish National Library (BNE) catalogue. This initiative has kicked off with the publication, under Linked Data principles, of information from the bibliographic and authorities catalogues, making them available as RDF knowledge bases. Furthermore, these bases are interrelated with other knowledge bases existing within the Linking

³⁴ <http://dublincore.org/documents/dcmi-terms/>

³⁵ <http://rdfs.org/sioc/spec/>

³⁶ <http://metadata.net/mpeg7/>

³⁷ <http://www.w3.org/2003/01/geo/>

³⁸ <http://webenemasuno.linkeddata.es/Ontology/OWLDOC/index.html>

³⁹ <http://openprovenance.org/model/opmo>

⁴⁰ <http://lab.isoco.net/technologies/luf>

Open Data initiative. With this initiative, Spain has joined the ranks of other institutions such as the British Library and the Deutsche Nationalbibliothek that have recently launched similar projects.

Regarding the *specification activity* the data source taken for this project is composed of around 3.9 million bibliographic and 4.2 million authority records in the MARC21⁴¹ format using the ISO 2709⁴² encoding standard, provided as data dump from the BNE catalogue. With the respect to the data quality, we have assumed the records to present necessary quality features like consistency, accuracy, or precision, as the records have been created and curated by highly qualified professionals. Regarding the URI design we have put special focus on minting URIs in a multilingual scenario, taking into account the design of TBox and ABox URIs.

After the specification activity, we define the ontologies to be used for *modelling* the domain of those data sources. Again, the most important recommendation is to reuse as much as possible available vocabularies and ontologies. To this end, we have decided to reuse a very well established set of models as the IFLA Functional Requirements family⁴³ and ISBD elements⁴⁴ that have been agreed upon and used by the library community.

As for the *generation activity* we have used MARiMbA⁴⁵ that provides a tool focused on enabling and easing the mapping from MARC21 to the RDF(S)/OWL models, and the transformation into RDF. In this activity we have emphasized the importance of the mapping in which we establish correspondences between MAR21 records the chosen vocabularies. Furthermore, MARiMbA's main design principle is to provide an easy-to-use mapping framework that allows domain experts (i.e. librarians and cataloguers) to work independently from the IT team and to establish the mapping rules without the need of high technical skills (e.g. knowledge about XML, XSLT or other ad-hoc mapping languages).

Regarding the *linking activity*, we chose VIAF⁴⁶, the Virtual International Authority File from the OCLC, as a candidate dataset since we had to deal with catalogue authoring data. VIAF provides a cluster of authority records of the same entity across several national authority files. This cluster has proven to be useful to establish links to other authoritative datasets such as the German National Library.

Once we have the legacy data transformed into RDF, we need to store and publish the data in a triplestore. In our scenario we have decided to rely on Virtuoso Universal Server and Pubby and to provide an additional web-developers-friendly API (providing simple methods for retrieving persons, works, etc.), using

⁴¹ <http://www.loc.gov/marc/bibliographic/>

⁴² http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=7675

⁴³ <http://www.ifla.org/en/node/2016>

⁴⁴ <http://metadataregistry.org/schema/show/id/25.html>

⁴⁵ www.oeg-upm.net/index.php/en/downloads/228-marimba

⁴⁶ <http://viaf.org/>

Puelia⁴⁷. For the metadata information we have used VoID for expressing the metadata about our dataset.

Finally, for the *exploitation* we have developed a prototype application that provides easy full-text searching and navigation based on the FRBR model.

4 Conclusions

In this paper we claim that it is possible to have a single and unified method for publishing Linked Data, but we should rely on different techniques, technologies and tools for particular datasets of a given domain. Therefore, we have presented our method for publishing Linked Data and the application of that method to different domains. Table 1 summarizes the vocabularies, technologies, and applications used in the development of our initiatives. The first column represents the six main activities of our methodological guidelines, i.e., modeling, RDF generation, links generation, publication and exploitation. The rest of columns include the technological support used for each phase within the set of our initiatives.

⁴⁷ Puelia is an implementation of the Linked Data API specification, see <http://code.google.com/p/puelia-php/>

Phase	BNE	IGN	AEMET	PRISA	INE
Modeling	  DC	 Wgs84 time W3C	 SSN ontology	 SIOC	Scovo Data cube
RDF generation	 MARIMba	  geometry2rdf NOR2O	 CSV parser	  CSV parser	 NOR2O
Links generation	 DNB VIAF LIBRIS DBPEDIA	Silk DBPEDIA Geonames	Silk Geolinkeddata.es	Silk DBPEDIA Geolinkeddata.es	NOR2O Geolinkeddata.es
Publication	 OPENLINK VIRTUOSO	 Pubby	 ckan	 ckan	 < SITEMAP XML /> sitemap4rdf
Exploitation	 map4rdf		SPARQL		

Table 1. Summary of the vocabularies, technologies, and applications used in our initiatives.

Moreover, we have to take into account that there are (1) multiple features such as geospatial, textual, and temporal among others; (2) multiple formats such as csv, excel, pdf, databases, among others; (3) multiple data models such as MARC21 for bibliographic data, simple or complex geometry for geographic information, etc; (4) multiple features or special characteristics of the data such as dynamicity, provenance, clashes between ids, cleansing and curation of existing data sources, etc; (5) multiple URI design decisions depending on the type of data we are dealing with; (6) multiple types of license of the datasets, depending on the creators and publishers; and (7) multilingualism of datasets and how this affects the creation of mappings and RDF.

Acknowledgments.

This work has been supported by BabelData (TIN2010-17550), myBigData (TIN2010-17060), PlanetData (FP7-257641), and Webn+1 (TSI-020301-2009-24) projects. We would like to thanks to all OEG members involved in the Linked Data initiatives.

References

1. K. Bizer, Volz and Gaedke. Silk - a link discovery framework for the web of data. In *18th International World Wide Web Conference*, pages 559–572, 2009.
2. D. Garijo, B. Villazón-Terrazas, and O. Corcho. A provenance-aware Linked Data application for trip management and organization. In *Proceedings of the 7th International Conference on Semantic Systems, I-Semantics '11*, pages 224–226, New York, NY, USA, 2011. ACM.
3. T. Heath and C. Bizer. *Linked Data: Evolving the Web into a Global Data Space*. Morgan & Claypool, 1st edition, 2011.
4. L. Moreau, B. Clifford, J. Freire, J. Futrelle, Y. Gil, P. Groth, N. Kwasnikowska, S. Miles, P. Missier, J. Myers, B. Plale, Y. Simmhan, E. Stephan, and J. Van den Bussche. The open provenance model core specification (v1.1). *Future Generation Computer Systems*, July 2010.
5. M.-C. Suárez-Figueroa, A. Gómez-Pérez, E. Motta, and A. Gangemi. *Ontology Engineering in a Networked World*. Springer, Berlin, 2012.
6. B. Villazón-Terrazas, L. Vilches-Blázquez, O. Corcho, and A. Gómez-Pérez. Methodological Guidelines for Publishing Government Linked Data Linking Government Data. In *Linking Government Data*, chapter 2, pages 27–49. Springer New York, New York, NY, 2011.