Semantic Web (2012) 1–6
IOS Press

datos.bne.es: a Library Linked Data Dataset

Editor(s): Pascal Hitzler, Kno.e.sis Center, Wright State University, Dayton, OH, USA; Krzysztof Janowicz, University of California, Santa Barbara, USA
Solicited review(s): Sören Auer, University of Leipzig, Germany; Tomi Kauppinen, University of Münster, Germany; Danh Le Phuoc, DERI Galway, Ireland

Daniel Vila-Suero [a], Boris Villazón-Terrazas [b] and Asunción Gómez-Pérez [a]
[a] *Ontology Engineering Group, Facultad de Informática, Universidad Politécnica de Madrid, Spain*
E-mail: {dvila,asun}@fi.upm.es
[b] *iSOCO, Avda. Partenón. 16-18, 28042, Madrid, Spain*
E-mail: bvillazon@isoco.com

Abstract. We describe the datos.bne.es library dataset. The dataset makes available the authority and bibliography catalogue from the Biblioteca Nacional de España (BNE, National Library of Spain) as Linked Data. The catalogue contains around 7 million authority and bibliographic records. The records in MARC 21 format were transformed to RDF and modelled using IFLA (International Federation of Library Associations) ontologies and other well-established vocabularies such as RDA (Resource Description and Access) or the Dublin Core Metadata Element Set. A tool named MARiMbA automatized the RDF generation process and the data linkage to DBpedia and other library linked data resources such as VIAF (Virtual International Authority File) or GND (Gemeinsame Normdatei, the authority dataset from the German National Library).

Keywords: Library Linked Data, Cultural Heritage Linked Data, Linked Open Data, Bibliographic Data, Authority Data, FRBR, RDF

1. Introduction

In recent years, the amount of semantically structured knowledge available on the Web as part of the so-called *Linked Open Data (LOD) cloud* has seen a substantial growth in the domain of cultural heritage and, particularly, in digital libraries. Indeed, libraries, museums and archives are showing great interest in publishing their data as Linked Data (LD) [9]. Several national libraries for example started to publish metadata as Linked Data, including the Swedish National Library [8], the German National Library[1], the National Library of France[2] and the British Library[3]. Initiatives like Europeana LOD [7] or VIAF[4] (the Virtual International Authority File) also stress the opportunities offered by Linked Data within the cultural heritage domain. Moreover, the Linked Data value proposition is already producing changes at organizational levels within library organizations as exemplified by the *Stanford Linked Data Manifesto*[5], the development of a new Linked Data-based bibliographic framework by the Library of Congress, or the official support to open data from the Conference of European National Libraries (CENL).

The benefits of publishing library data as Linked Data have been recently summarized [2] by the W3C Incubator Group on Library Linked Data. In particular, the following key benefits of *Library Linked Data (LLD)* have been identified: i) provides enhanced and more sophisticated navigation through information, ii) increases the visibility of cultural data, iii) supports integration of cultural information and digital objects into research documents and bibliographies, iv) offers

[1] http://bit.ly/Kdisdu
[2] http://data.bnf.fr
[3] http://bnb.data.bl.uk/
[4] http://viaf.org
[5] http://bit.ly/vcwhVe

a more durable and robust semantic model than metadata formats that rely on specific data structures, v) facilitates re-use across cultural heritage datasets, thus enriching the description of materials with information coming from outside the organization's local domain of expertise, and vi) allows developers and vendors to avoid being tied to library-specific data formats such as MARC[6] (MAchine Readable Cataloguing) or Z39.50[7].

In this context, at the end of 2011, the National Library of Spain (Biblioteca Nacional de España, BNE) has transformed into RDF and linked the *Authority and Bibliography* catalogues that contain around 7 million authority and bibliographic records. BNE authorities and publications are directly linked to resources in other datasets: VIAF, DBpedia, SUDOC, LIBRIS, and the GND (the authority file from the German National Library). datos.bne.es contains approximately 58 million triples generated from MARC 21 records and 587,000 links to other resources in different languages (e.g., German, French, English, Swedish). The following sections provide a description of key features of the datos.bne.es dataset, and a guide for potential data consumers.

2. datos.bne.es

datos.bne.es is the main result from the project *Linked Data at the BNE*, supported by Biblioteca Nacional de España in cooperation with the Ontology Engineering Group (OEG) at the Universidad Politécnica de Madrid (UPM). There is a strong commitment from the BNE to maintain and to periodically update and improve the RDF dataset. This commitment is in line with the shift towards open web standards and licenses. datos.bne.es is licensed under the CC0 1.0 Universal Public Domain Dedication license. The following sections present the main characteristics of the RDF dataset.

2.1. Source data

The last version of the RDF dataset is the result of transforming:

– Authority records[8]: These records contain metadata describing people, organizations, work titles, and subject headings.

– Bibliographic records[9]: These records form a representative subset of the BNE catalogue, and provide metadata belonging to modern and ancient monographs, electronic records, manuscripts, periodical publications, maps, engravings, photographs, printed music, sound and audiovisual recordings.

More information about the characteristics of the source data, as well as links to the BNE catalogue, can be found in the `datos.bne.es` portal.

2.2. Modelling

IFLA (International Federation of Library Associations and Institutions) ontologies, widely agreed upon by the library community, have been used to represent the resources in RDF. datos.bne.es is one of the first international initiatives to thoroughly embrace the ontologies developed by IFLA [10], namely FRBR (Functional Requirements for Bibliographic Records) [4], FRAD (Functional Requirements for Authority Data) [3], FRSAD (Functional Requirements for Subject Authority Data) [5], and ISBD (International Standard for Bibliographic Description) [6]. The first version of the datos.bne.es dataset focused on standard and widely used vocabularies.

Throughout this paper we will use *compact URIs* for identifying vocabulary elements (e.g., `frbr:C1001`), prefixes can be resolved to namespaces using the `prefix.cc` service[10].

Table 1
Classes and number of instances within datos.bne.es

Label	URI	Count
Manifestation	`frbr:C1003`	2,390,103
Work	`frbr:C1001`	1,969,526
Person	`frbr:C1005`	1,163,764
Expression	`frbr:C1002`	1,114,719
Thema	`frsad:C1001`	497,644
Corporate body	`frbr:C1006`	282,879

The following classes from *FRBR* form the core of the vocabulary: (1) Person, (2) Corporate body, (3) Work, (4) Expression, and (5) Manifestation. Also, the class Thema from the FRSAD ontology has been used to model the subject authority data. Table 1 present the number of instances. Additionally, other elements have been used from a number of vocabularies, namely

[6] `http://www.loc.gov/marc/`
[7] `http://www.loc.gov/z3950/agency/`
[8] `http://www.loc.gov/marc/authority/`
[9] `http://www.loc.gov/marc/bibliographic/`
[10] `http://prefix.cc`

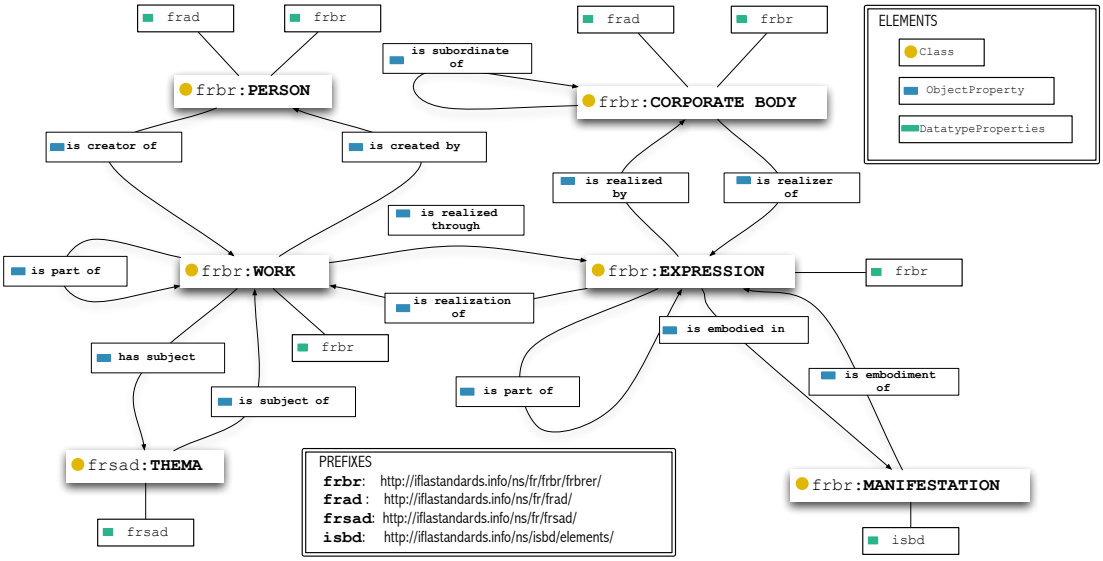

Fig. 1. Overview of the ontology network

ISBD, RDA (Resource Description and Access) Group Elements 2[11] and Relationships for WEMI[12], Dublin Core terms[13], SKOS[14], MADS/RDF[15]. Figure 1 provides an overview of the vocabulary described above.

Table 2
Object properties and their usage across the datos.bne.es

Label	URI	Count
language	dcterms:language	3,112,900
is creator (person) of	frbr:P2010	2,129,222
is created by (person)	frbr:P2009	2,129,222
is embodiment of	frbr:P2004	1,246,773
is embodied in	frbr:P2003	1,246,773
is realized through	frbr:P2001	1,054,736
is realization of	frbr:P2002	1,054,736
same as	owl:sameAs	587,520
subject	dcterms:subject	249,560

2.3. *Extraction and mapping process*

The transformation process from the MARC 21 format to RDF has been carried out using MARiMbA[16], a tool developed to support the mapping process, allowing librarians to use a combination of RDFS and OWL vocabularies of their choice to model the data. The tool performs a pre-processing step that generates a set of spreadsheets, based on the source records. This spreadsheets function as mapping templates where the librarians can manually map the different MARC 21 metadata elements to appropriate RDFS and OWL classes and properties.

MARiMbA exploits MARC 21 record structure, namely *field codes* (e.g., `100`, `245`), *subfield codes* (e.g., `$a`, `$t`), and *heading information fields*[17]. Although the meaning of these codes is defined by the MARC 21 standard, not all codes are always used and their usage may vary across different libraries. Taking this into account, one of the design principles behind the approach followed by MARiMbA is to derive the mappings from the actual data (i.e. by pre-processing the source records). After pre-processing the source data, the librarians are presented with the current metadata elements used across the source records so they can manually assign the correspondences to RDFS and OWL classes and properties.

The pre-processing step generates three spreadsheets: (1) *Classification mapping*: where the user is presented with all possible combinations of heading field and subfields (e.g., `100 $a$t`), used for assign-

[11] http://rdvocab.info/ElementsGr2/
[12] http://rdvocab.info/RDARelationshipsWEMI/
[13] http://purl.org/dc/terms/
[14] http://www.w3.org/2004/02/skos/core
[15] http://www.loc.gov/standards/mads/rdf
[16] http://marimba4lib.com

[17] http://www.loc.gov/marc/authority/ad1xx3xx.html

ing an RDF type or OWL class (e.g., `frbr:Work`) to each MARC record in the transformation process; (2) *Annotation mapping*: where the user is presented with all possible combinations of field and subfield (e.g., `100 $t`) for each type of entity (e.g., Work), used for mapping each MARC element to an specific property (e.g., `frbr:titleOfWork`); and (3) *Relation mapping*: where the user is presented with all possible variations of subfields (e.g., `100 $a + $t`) in the heading field for each pair of entities (e.g., Person-Work, Work-Work). This mapping is used for creating a relationship between a pair of records of a certain type given that the string of the heading field of one record contains the string of the heading field of the other and that they present a certain variation of subfields.

Figure 2 shows an example of the mapping process. Given two records with the following heading fields: (1) `100 $a` *Cervantes Saavedra, Miguel de*, and (2) `100 $a` *Cervantes Saavedra, Miguel de* `$t` *Don Quijote de la Mancha*. First, the records are mapped to `frbr:Person` and `frbr:Work` respectively, based on the *classification mapping*. Second, field `$a` is mapped to `frbr:nameOfPerson`, and the field `$t` to `frbr:titleOfWork`, based on the *annotation mapping*. Finally, both resources are related through `frbr:isCreatorOf` after a string comparison and the analysis of their variation of subfields (`100 $a + $t`), based on the *relation mapping*. In addition, the website found at `http://bne.linkeddata.es/mapping-marc21/` has been set up to provide more details about the mapping and transformation processes, as well as the complete set of mappings used in the transformation of the RDF dataset.

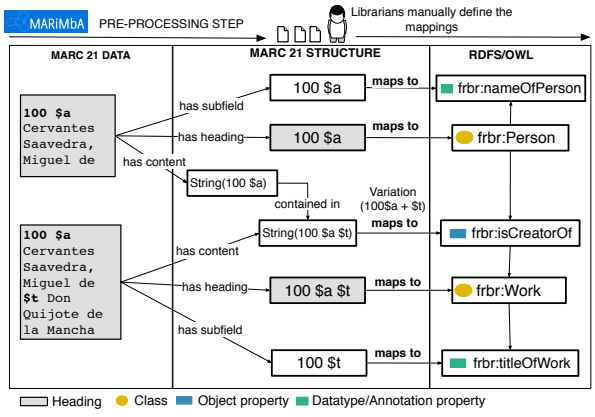

Fig. 2. Mapping process from MARC 21 records

.

2.4. Connectivity

In order to analyse the internal structure and connectivity of the instances of the dataset, Table 2 presents the number of relationships established. From Table 1 it can be observed that there is a high number of manifestations, works, persons, and expressions, which correspond to the core entities of FRBR and in general to the main entities of the bibliographic world. Consequently, among the most established relationships are those that connect these core entities, namely *is creator of* (between a person and a work), *is realized through* (between a work and an expression), and *is embodied in* (between an expression and a manifestation). The total number of relationships established between instances within the dataset is 9,394,711, while the total number of instances is 7,418,635. From these data it can be stated that: (1) although a high number of relationships has been established, this version of the dataset only covers primary relationships from FRBR, (2) the dataset provides a high number of inverse relationships (e.g., is creator of, is created by) which can facilitate traversing the graphs in several directions. In Section 3 a concrete example of the internal structure and connectivity of the dataset will be discussed.

Regarding the external connectivity (i.e. the links to external datasets), datos.bne.es linksets can be divided into two groups: (1) equivalence links using the `owl:sameAs` object property (in descending order): VIAF (454,068), GND[18] (76,413), DBpedia (36,431), Libris (10,884), and SUDOC[19] (9,725), the total number of `owl:sameAs` links is 587,521, and (2) links to Lexvo dataset[20] using the `dcterms:language` property. It is worth mentioning that VIAF already has links to widely-used authority files and a number of links to DBpedia. Therefore, in this version of the dataset we focused on reusing this authoritative and valuable resource. More specifically, given that (1) VIAF links are available online[21] as a plain text file, and (2) most libraries have published their authority files using natural keys[22] to build the URIs of their RDF resources, MARiMbA provides the functionality to generate equivalence links by parsing the aforementioned

[18] `http://d-nb.info/`
[19] `http://www.idref.fr/`
[20] `http://lexvo.org/`
[21] `http://thedatahub.org/viaf`
[22] `http://patterns.dataincubator.org/book/natural-keys.html`

VIAF links and prepending the namespaces to the different keys found in the VIAF links file.

For example, we know that GND URIs follow the pattern `gnd:{GND-ID}` and BNE URIs the pattern `bne:{BNE-ID}`. Using these two URI patterns we can establish links between both datasets by creating `owl:sameAs` statements using GND-ID and BNE-ID pairs found in the VIAF links file. In this way, the GND-ID `11851993X` found in the same VIAF cluster as the BNE-ID `XX1718747` can be used to create the following statement about *Miguel de Cervantes*:

```
@prefix bne: <http://datos.bne.es/resource/> .
@prefix dnb: <http://d-nb.info/gnd/> .

bne:XX1718747 owl:sameAs gnd:11851993X
```

Finally, from the analysis of external links can be concluded that: (1) there is a high number of `owl:sameAs` links to external library datasets, but these links cover only authority entities directly mapped from VIAF links, namely persons, corporate bodies, works and expressions, (2) for a future version of the dataset, links at the level of bibliographic entities (i.e., manifestations) would improve the connectivity of the dataset, (3) a high number of links to the Lexvo language resources has been established providing rich information about such language resources.

2.5. General features

Table 3 provides a summary of the main features and statistics, according to the VoID vocabulary [1].

3. Miguel de Cervantes Graph Analysis

This section shows a brief example of the data that can be found within datos.bne.es dataset. More specifically, it discusses some of the characteristics of a representative graph, a portion of the data related to *Miguel de Cervantes Saavedra*, which can help to understand and illustrate the dataset internal structure. Moreover, this experiment can help potential consumers to understand the main entities and the relationships between them. Using datos.bne.es SPARQL endpoint, we build a graph by issuing the following SPARQL query:

```
prefix bne: <http://datos.bne.es/resource/>
CONSTRUCT {
  bne:XX1718747 ?r ?y .
  ?y ?r2 ?y2 . ?y2 ?r3 ?y3
} WHERE {
  bne:XX1718747 ?r ?y .
  ?y ?r2 ?y2 . ?y2 ?r3 ?y3 .
  FILTER(!isLiteral(?y) ). FILTER(!isLiteral(?y2)).
  FILTER regex(?y, "^http://datos.bne.es") .
  FILTER regex(?y2, "^http://datos.bne.es") .
  FILTER (!isLiteral(?y3) ) .
  FILTER regex(?y3, "^http://datos.bne.es")
}
```

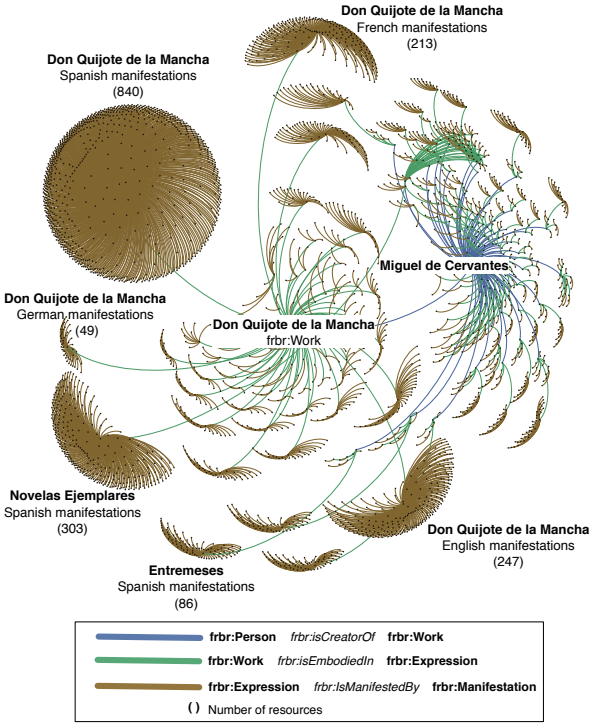

Fig. 3. Visualization of the graph around the RDF resource *Miguel de Cervantes*. Online version at `http://bne.linkeddata.es/graphvis/`

.

This query builds a graph around the RDF resource *Miguel de Cervantes*[23]. Each node of the graph corresponds to one RDF resource, and each edge corresponds to a relationship between two resources. More specifically: (1) the graph describes the works created by *Miguel de Cervantes*, (2) the different expressions of each work (e.g., the English version of *Don Quijote de la Mancha*), (3) the different manifestations of each expression (e.g., a publication from 1898 of the English expression of *Don Quijote de la Mancha*), and (4) the relationships between these resources (e.g., *Cervantes* is the creator of *Don Quijote*).

Figure 3 depicts the graph. In the figure, black dots represent the nodes (3,281 nodes), and edges repre-

[23] `http://datos.bne.es/resource/XX1718747`

Table 3
Overview of the dataset characteristics)

Property	VoID property	Value
VoID file	-	http://datos.bne.es/void/bne.ttl
Homepage	foaf:homepage	http://datos.bne.es
Datahub page	foaf:page	http://thedatahub.org/dataset/datos-bne-es
Publisher	dcterms:publisher	http://dbpedia.org/resource/Biblioteca_Nacional_de_España
License	dcterms:license	http://creativecommons.org/publicdomain/zero/1.0/
Base URI for instances	void:uriSpace	http://datos.bne.es/resource/
SPARQL endpoint	void:sparqlEndpoint	http://datos.bne.es/sparql
Data dumps address	void:dataDump	http://datos.bne.es/datadumps/
Total number of triples	void:triples	58,053,215
Total number of entities	void:entities	7,412,286
Total number of distinct subjects	void:distinctSubjects	7,413,108
Total number of distinct objects	void:distinctObjects	4,716,874

sent the relationships between the nodes (3,974 edges), each type of relationship (i.e. object properties) has a different color (blue edges represent `is creator of`, green edges `is realized through`, and brown edges `is embodied in`). *Cervantes* and *Don Quijote* can be found in the center of the graph, surrounded by high concentrations of brown edges that represent the different manifestations (editions) of the most famous works by *Cervantes*. For example, the most published work of *Cervantes* is the original version of *Don Quijote* with 840 different editions in BNE catalogue.

4. Conclusions

This paper presents a description of the main characteristics of the datos.bne.es dataset and the process followed in its development using the MARiMbA tool. One contribution is that domain experts (librarians and cataloguers) mapped the MARC 21 metadata elements to highly specialized library models provided by IFLA using spreadsheets. In this way, we involved domain experts in the linked data generation, reducing considerably the effort and time spent. MARiMbA approach is flexible enough to allow other cultural institutions to transform their catalogue data into RDF.

Regarding the external connectivity of the dataset, the availability of already validated resources, like VIAF, has facilitated the creation of high quality links between the datos.bne.es dataset and other datasets in the LOD cloud.

Finally, we can say that datos.bne.es represents one of the biggest contributions of high-quality RDF data in Spanish to the LOD cloud. More specifically, it is already being used by initiatives like BibSoup[24] from the Open Knowledge Foundation and JISC (Joint Information Systems Committee). Furthermore, it is planned to position datos.bne.es as a reference data provider for national and regional libraries within Spain.

Acknowledgments. This work is supported in part by the Spanish Project TIN2010-17550 for the BabeL-Data project, and by BNE. We thank the BNE team (M. Jiménez Piano, E. Escolano, M. Hernández, A. Manchado, and R. Sánchez) for their hard work. Finally, we thank the reviewers for their meaningful revisions.

References

[1] K. Alexander, R. Cyganiak, M. Hausenblas, and J. Zhao. Describing Linked Datasets with the VoID Vocabulary. W3C interest group note, W3C, 2011. `http://www.w3.org/TR/void/`.

[2] T. Baker, E. Bermes, K. Coyle, G. Dunsire, A. Isaac, P. Murray, M. Panzer, J. Schneider, R. Singer, E. Summers, W. Waites, J. Young, and M. Zeng. W3C Library Linked Data Incubator Final Report. 2011.

[3] IFLA. *Functional Requirements for Authority Data: a conceptual model*. IFLA Working Group on Functional Requirements and Numbering of Authority Records (FRANAR), München : K.G. Saur. IFLA series on bibliographic control, v. 34. edition, 2009.

[4] IFLA. *Functional Requirements for Bibliographic Records: Final Report*. IFLA Study Group on the Functional Requirements for Bibliographic Records, UBCIM publications ; new series, vol. 19 (Amended and corrected) edition, 2009.

[5] IFLA. *Functional Requirements for Subject Authority Data (FRSAD): a conceptual model*. IFLA Working Group on the Functional Requirements for Subject Authority Records, Berlin ; New York : De Gruyter Saur. IFLA series on bibliographic control, v. 43. edition, 2011.

[24] `http://bibsoup.net/`

[6] IFLA. *ISBD: International Standard Bibliographic Description: Consolidated Edition*. Berlin Boston, Mass. De Gruyter Saur. IFLA series on bibliographic control, v. 34., 2011.

[7] A. Isaac and B. Haslhofer. Europeana Linked Open Data – data.europeana.eu. *Semantic Web Journal, to appear. Available from http://www.semantic-web-journal.net/*.

[8] M. Malmsten. Making a library catalogue part of the semantic web. In *Proceedings of the 2008 International Conference on Dublin Core and Metadata Applications*, DCMI '08, pages 146–152. Dublin Core Metadata Initiative, 2008.

[9] D. Vila-Suero. W3C Library Linked Data Incubator Group Use Cases Report. 2011.

[10] D. Vila-Suero and E. Escolano. Linked Data at the Spanish National Library and the Application of IFLA RDFS Models. In *IFLA SCATNews Number 35*, 2011.