# datos.bne.es and MARiMbA:

# An insight into Library Linked Data

Daniel Vila-Suero and Asunción Gómez-Pérez

## 1. Introduction

In recent years, the amount of semantically structured data available on the Web as part of the so-called "Linked Open Data (LOD) cloud" (Heath and Bizer, 2011) has witnessed a substantial growth. Libraries, museums and archives are showing great interest in publishing their data as Library Linked Data (LLD). Several national libraries have published their data as LD, including the Swedish National Library (Malmstem, 2008); the Library of Congress (LoC) (Summers et al., 2008); the German National Library (DNB) [i]; the National Library of France (BnF) [ii], the British Library (BL) [iii], and Biblioteca Nacional de España (BNE, National Library of Spain) (Vila-Suero et al., 2013). Europeana (Isaac and Haslhofer, 2013) and VIAF (Virtual International Authority File) [iv] are examples of larger scale LLD publication from multinational organizations. Other relevant initiatives are (i) the Stanford Manifesto [v], produced during the Stanford Linked Data Workshop; (ii) the new bibliographic framework from the Library of Congress [vi] and the BIBFRAME vocabulary [vii]; and, (iii) the support provided by the Conference of European National Libraries (9) (CENL) to open data and reuse following LD best practices and technologies.

The benefits of publishing Library Linked Data were summarized by the W3C Incubator Group on Library Linked Data (Baker et al., 2011). These benefits are the following: (i) LLD provides enhanced navigation through and discovery of cultural information; (ii) it increases the visibility of cultural data on the Web; (iii) it offers integration of cultural information and digital objects into research documents and bibliographies by means of open web standards; (iv) it provides a more durable and robust semantic model than metadata formats that rely on specific data structures; (v) it facilitates re-use across cultural heritage datasets, thus enriching the description of materials with information from outside the organization's local domain of expertise; and (vii) it allows developers and vendors to avoid being tied to library-specific data formats, such as MARC (MAchine Readable Cataloging) and Z39.50 [viii].

As highlighted above, current library data are usually stored and handled through specialized formats, especially the MARC format. Therefore, some efforts within the library field have focused on transforming MARC 21 records into RDF (Harper and Tillet 2007) (Malmstem, 2008) (Vila-Suero, 2011). In this paper, we aim at exposing our experience in publishing LLD from MARC records of BNE, the *datos.bne.es* dataset, following a method powered by our tool MARiMbA (Vila-Suero, 2011) [ix]. We also present our experience gained in applying the FRBR (Functional Requirements for Bibliographic Records) (IFLA, 1998) and ISBD (International Standard for Bibliographic Records) (IFLA, 2011) vocabularies to MARC records, leveraging LD best practices. Since standardized practices for publishing and integrating LLD across libraries are not yet widely discussed, we expect that this work can contribute to reflecting on the evolution of such practices.

The rest of the paper is organized as follows. Section 2 presents an overview of the *datos.bne.es* case study and the process followed along its development. Sections 3 to 9

describe the activities of the process. Finally, Section 10 provides some conclusions.

## 2. datos.bne.es project: An overview

Since 2006, the Spanish Ministry of Culture has been pursuing a way to improve the interoperability of the authority control and between authority files of Spanish libraries. In this line, they have proposed the creation of a national authority file, managed by the BNE that could serve as tool of reference for both Spanish and Latin-American libraries. The rationale for building and maintaining such authority system is to avoid duplication of records, to increase cataloguing quality and extensibility, and to save operational costs.

In this context, motivated by the growing interest in LOD and semantic technologies, in 2011 BNE and the Ontology Engineering Group from "Universidad Politécnica de Madrid" started a project with the purpose of transforming the authority and bibliographic catalogues into RDF following LD best practices.

### *2.1. Initial considerations*

This section describes three main factors about the *datos.bne.es* case study that have influenced some of the design decisions presented in this paper and that apply to other LLD initiatives.

The first factor relates to the **nature of the data sources** transformed into RDF: *MARC 21 records*. MARC 21 is a standard digital format developed by the LoC in the '60s for the representation and communication of bibliographic and related information in machine-readable form. Since then it has been one of the most widely used standards for the storage and communication of bibliographic information. However, as a highly specialized and relatively old format, MARC presents several drawbacks that need to be taken into account when transforming MARC into a more "semantic" and "open" format such as RDF, for example.

1. MARC records present a "flat" internal structure, as opposed to richer structures such as relational databases, making it more difficult to map their structure to richer models like FRBR.
2. During decades MARC has evolved together with cataloguing rules and practices. This evolution has produced an impact on the use of the different metadata elements within library catalogues, making it challenging to clearly define the semantics of MARC's metadata elements.

The second factor is the importance of **encouraging the participation of library domain experts** (e.g., cataloguers) in the LLD process, especially in the analysis of data sources and in the mapping from MARC 21 records to the RDF vocabulary since library catalogues are built on a set of highly specialized evolving practices, rules, data models, and methods.

Finally, another important factor is the **quality of data sources**. It is worth noting that although library catalogues contain high quality data curated by trained professionals, there are still issues to be solved. These issues will be analyzed in greater detail in Section 7 and range from problems at the data level (e.g., MARC codes errors) to higher-level errors (e.g., lack of authority records for certain works). Most of the issues reflect the evolution of the catalogue and are produced by changes in the cataloguing rules and by migration from one

system to another over the years, among others. The LLD generation process has allowed us to semi-automatically detect deficiencies in the data sources. Therefore, data curation emerges as an important added value offered by the LLD process.

## *2.2. Method and process overview*

In order to carry out the transformation, linkage and publication of the BNE linked dataset, we have followed a method, based on a modification and extension of Villazón-Terrazas et al., (2011), which consists of the following activities: *Specification*, *Modeling*, *Generation*, *Publication*, *Linking*, *Data curation*, and *Exploitation*. Each of these activities is then decomposed into several tasks. We have followed an iterative-incremental lifecycle along the case study development. In particular, we have carried out two iterations, as shown in Figure 1. The set of activities and tasks will be described in the following sections and are summarized in Table 1

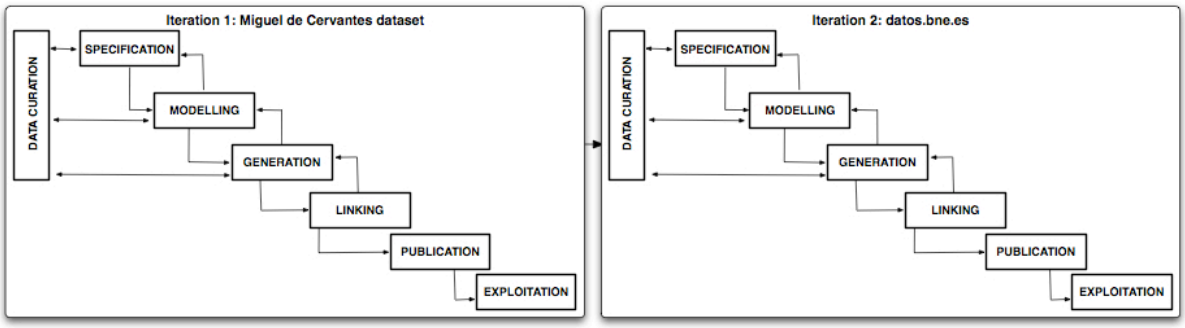


Fig. 1. An Iterative-incremental lifecycle model

**First iteration:** *Miguel de Cervantes* dataset**.** This iteration, discussed in Vila-Suero and Escolano (2011), aimed at transforming a subset of records related to "Miguel de Cervantes". To explain this in an intuitive manner, the subset included all works by "Miguel de Cervantes", all related publications, all authorities (persons, organizations and subjects) related to these publications and, finally, all works related to these authorities. In total, the data source is composed of 8,552 bibliographic records and 41,972 authority records in the MARC 21 format with the ISO 2709 encoding standard. The RDF dataset was transformed into RDF using IFLA (International Federation of Library Associations and Institutions) vocabularies, namely FRBR, FRAD (Functional Requirements for Authority Data) and ISBD, and it was linked with VIAF.

**Second iteration:** *datos.bne.es* dataset**.** The goal of this iteration was to transform both the complete set of authority records and a subset of the bibliographic catalogue into RDF. The subset selected included records describing modern and ancient monographs, electronic records, manuscripts, periodical publications, printed music, sound and audiovisual recordings, maps, engravings, and photographs. This selection was intended to maximize the representativeness of the records while keeping a reasonable quality of the produced Linked Data with regards to the application of FRBR. More specifically, in line with our iterative and incremental approach, we performed several experiments to asses how the different sets of bibliographic records responded to the process of applying FRBR and based on this analysis we selected those that produced better results. This paper describes the second iteration,

which produced the current version of the *datos.bne.es* dataset. The remaining sections of the paper will focus on describing each of the activities and tasks (shown in Table 1) performed along this iteration. As will be discussed in Section 10, for the next iteration we plan to include the remaining bibliographic records after a careful analysis of their suitability to the current data model.

| | 1.**Specification** (Section 3) | 2.**Data Curation** (Section 7) | 3.**Modelling** (Section 4) | 4.**Generation** (Section 5) | 5.**Linking** (Section 6) | 6.**Publication** (Section 8) | 7.**Exploitation** (Section 9) |
|---|---|---|---|---|---|---|---|
| **Goal** | Analyzing and describing data (data sources and RDF data) characteristics | Fixing and improving both the data sources and the RDF | Creating a vocabulary to describe the RDF resources | Producing RDF resources from the data sources | Connecting the RDF dataset to other relevant datasets | Making the dataset available on the Web | Defining and developing applications that make use of the RDF dataset |
| **Tasks** | 1.Identify and analyze the data sources<br><br>2. Design the URIs<br><br>3. Definition of license and provenance information | 1. Data sources curation<br><br>2. RDF data curation | 1.Analyze and select domain vocabularies<br><br>2.Develop the vocabulary<br><br>3. Vocabulary for representing provenance information | 1.Select, extend or develop the technologies for producing RDF<br><br>2. Create mappings between the vocabulary and the data sources<br><br>3. Transform the data sources into RDF | 1.Select target datasets to link the entities in the dataset<br><br>2. Discover the links with the target datasets<br><br>3. Validate the links | 1. Publish the dataset<br><br>2. Publish metadata describing the dataset<br><br>3.Enable effective discovery of the dataset | 1.Develop or configure applications on top of the dataset |

Table 1. LLD main activities and tasks

## 3. Specification

The goal of the specification activity is to analyze and describe the data sources that will be transformed into LD and the dataset that will be produced. This activity can be further decomposed into three tasks: *Identifying and analyzing the data sources* (Section 3.1); *Designing the URIs* (Section 3.3); and *Defining the license and provenance information* (Section 3.4). Section 3.2 introduces MARC 21, the data sources format.

### *3.1. Identifying and analyzing the data sources*

Within this task we identify and select the BNE data sources to be used for publishing LLD. In addition, we need to search and compile all the available data and documentation about those resources, including purpose, data model and implementation details, and to identify the main entities described within the data sources and the relationships among them.

More than five million authority records and over eight million bibliographic records comprise the BNE catalogue. These records use the *authority* and *bibliographic* MARC 21

formats (introduced in Section 3.2). The records share some common characteristics but also present some differences as summarized in Table 2. The current version of the *datos.bne.es* dataset is both the result of transforming the complete set of authority records and a representative subset of the bibliographic records.

| | **Authority data source** | **Bibliographic data source** |
|---|---|---|
| **Purpose** | To carry information (metadata) concerning the authorized form of names and subjects to be used in access points to MARC 21 records. | To carry information (metadata) about bibliographic resources. These bibliographic resources conform the holdings of the library and include resources like printed and manuscript textual materials, maps, music, video, etc. |
| **Data model** | MARC 21 Format for Authority records | MARC 21 Format for Bibliographic records |
| **Main concepts** | Persons, Organizations, Conferences, Congresses, Subjects or topics, Works, Versions of Works (e.g. translation of a Work) | Publications including maps, manuscripts, electronic records, software, musical scores, sound and audiovisual recordings, among others |
| **Components** | Records composed by fields, subfields and indicators | |
| **Implementation details** | The records can be implemented in two different encodings: *ISO 2709* and *MARCXML* | |
| **Unique Identifiers** | Field 001 | |

Table 2. Data sources specification

### *3.2. MARC 21 in a nutshell*

The MARC 21 specification defines the logical structure of a machine-readable library record. Each record is divided into *fields* identified by three-digit *tags*. For example, the field with tag *100* contains the established form of a personal name in a record that conforms to the *Format for Authority data*.

The specification distinguishes two types of fields: *control*, and *data fields*. *Control fields* contain control numbers or other coded information used for processing records but do not contain *indicators* or *subfield* codes. *Data fields* contain information about the resource or resources described within the record and are typically subdivided into one or more *subfields* identified by a *subfield code* preceded by a delimiter (e.g. $). Additionally, some *data fields* are further defined by two character positions called *indicators* in order to further specify additional attributes. The meaning of the *subfield codes* and the two character *indicators* varies according to the *field tag* they precede.

*Field tags*, *subfield codes*, and *indicators* are known as *content designators*. The main purpose of the specification is to define the meaning of the possible values for these *content designators*. *Access points* are the fields of the record that enable users and librarians to find bibliographic records.

Finally, the specification [x] defines different communication formats: *authority*, *bibliographic*, *classification*, *community information*, and *holdings*. For example, the *Format for*

*Authority Data* defines the content designators for creating records by encoding the authorized forms of names used for constructing access points in other records.

For instance, Figure 2 depicts an extract from an authority record produced by BNE [xi], corresponding to the author Miguel de Cervantes. In the record we can find several *control fields*; for instance, the field *001* contains the BNE identifier. Moreover, the record contains a number of *data fields*. For example, the field *100* is the main access point to the record and contains information about the main entity being described by the record, whereas the subfield *$d* contains information about the dates associated with the described entity.

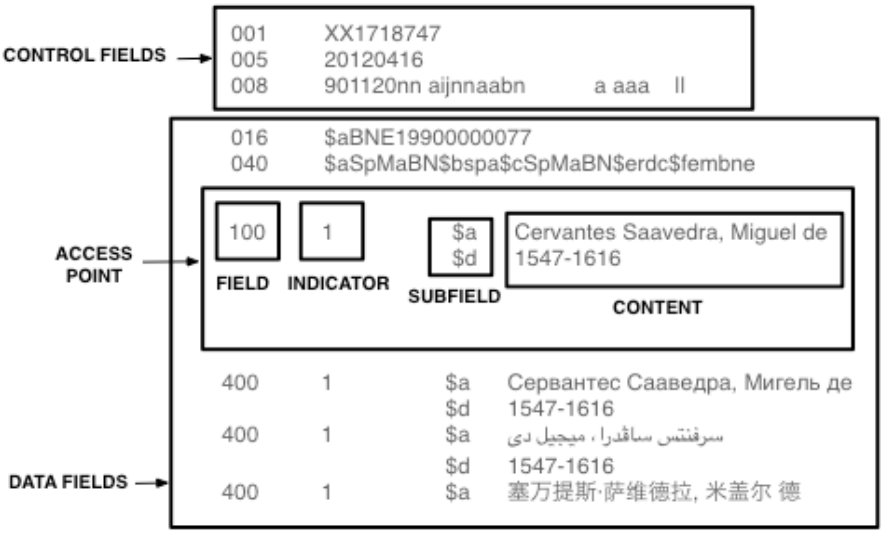


Fig. 2. Extract from Miguel de Cervantes' MARC 21 authority record

### *3.3. Designing the URIs*

This task defines the URIs that will be used as identifiers for the RDF dataset resources. We distinguish two type of URIs: (i) *Vocabulary URIs*, which identify the terminological components (RDF classes and RDF properties) for describing the entities and their relationships and attributes in the RDF dataset [xii]; and (ii) *Data URIs*, which identify the resources (also referred to as instances or individuals) that we are publishing.

Regarding *Vocabulary URIs*, we have reused a number of vocabularies [xiii]. For instance, IFLA namespaces are http://iflastandards.info/ns/, http://iflastandards.info/ns/fr/frbr/frbrer/ for the *FRBR* vocabulary; and http://iflastandards.info/ns/isbd/elements/ for the *ISBD elements*. Therefore, we have reused the URIs provided by the vocabulary publishers and have not minted any URI for the vocabulary components used within the dataset, which implies that the publishers (IFLA here) control and maintain these resources.

On the other hand, we have designed the *Data URIs* that identify the *datos.bne.es* resources. We have exclusively used HTTP URIs. The BNE is responsible for providing access to these resources when some application sends an HTTP request to such URIs. For creating the URIs we have used the *natural keys* pattern. The *natural keys* pattern is described in Dodds and Davis (2012) as the pattern of minting URIs algorithmically derived from existing unique identifiers. This pattern is a good match for identifying resources created out of MARC 21

records from a single catalogue since the field *001* or *control number* is used for uniquely identifying these records within the catalogue. In addition, the *control number* has been traditionally used for exchanging records between organizations; this control number can be useful for linking the RDF resources with other external datasets (as is the case with VIAF). In the *datos.bne.es* dataset we append the control number to the *base URI* http://datos.bne.es/resource/. It is worth noting that we have included the word *resource* in the namespace for our data items, so that in the future we will be able to separate these data elements from possible vocabulary elements created by the BNE with another namespace such as http://datos.bne.es/vocabulary/. For example, given that the control number of *Miguel de Cervantes*' record is *XX1718747,* by appending it to the base URI we identify *Cervantes* by http://datos.bne.es/resource/XX1718747.

Please note that throughout this paper we will use compact URIs, also known as *CURIES* (http://www.w3.org/TR/curie/), for identifying vocabulary elements (e.g., *frbr:C1001*) and that the prefixes can be resolved to namespaces with the *prefix.cc* service [xiv].

### *3.4. Defining license and provenance information*

Licensing datasets is a topic of discussion within the LLD domain. However, since the recent announcements made by several important organizations, such as Europeana [xv], CENL, or the Harvard Library [xvi], there seems to be a shift toward open licenses. More specifically, the CENL agreement to support *Creative Commons' Public Domain* license [xvii], also known as CC0, has already produced positive effects, exemplified by the releases of LLD datasets under the CC0 license from the DNB, the British Library, and *datos.bne.es*, among others.

Defining the provenance information is also an important task when publishing LLD. In *datos.bne.es,* we have to identify the following *aspects of provenance*: (i) the creator and publisher of the data; and (ii) temporal information (e.g., data creation and retrieval date). A more detailed discussion about the specific provenance elements that we provide for *datos.bne.es* is presented in sections 4.4 and 5.3.

## 4. Modeling

The goal of the modeling activity is the design and implementation of the vocabulary that will be used to describe the RDF resources to be be published following the LD principles. In this section we present the tasks identified in Table 1 for this activity. Such an activity can be further decomposed into three tasks: analyzing and selecting the domain vocabularies (Section 4.1); developing the vocabulary (Section 4.3); and choosing the vocabulary for representing the provenance information (Section 4.4).

### *4.1. Analyzing and selecting domain vocabularies*

According to Heath and Bizer (2011) the main recommendation for the *modeling* activity is to reuse as much as possible available and widely used vocabularies. As we will discuss in this section, several vocabularies and domain ontologies with varying potential and suitability for modeling library resources can be found.

Some general-purpose vocabularies such as the Friend-of-a-Friend (FOAF) ontology, or the Dublin Core Metadata Initiative vocabularies [xviii] are extensively used in LLD initiatives such as VIAF or DNB.

On the other hand, a number of domain-specific vocabularies created within the library community to describe bibliographic and authority data such as ISBD, FRBR, FRAD, FRSAD (Functional Requirements for Subject Authority Data), FRBRoo (FRBR-object oriented), MADS/RDF (Metadata Authority Description Schema in RDF), or the more recent RDA (Resource Description and Access) vocabularies are partially based on some of FRBR notions. Where the ISBD vocabulary mimics the bibliographic record on a catalogue card, the FR oriented models (including RDA) rely on a new conceptual model of the bibliographic universe using different levels of abstraction. Currently, within the new bibliographic framework initiative by the LoC, a new vocabulary, named BIBFRAME and built on existing models such as FRBR and RDA, is being developed and publicly discussed, and it represents a future alternative to those mentioned above. Additionally, the Europeana Data Model (EDM) is of significant relevance to libraries due to the role of the Europeana Project as a leading player in the dissemination of cultural materials.

Finally, besides those vocabularies developed within the library community, some more loosely modeled bibliographic vocabularies such as BIBO, the SPAR vocabularies, or SKOS (Simple Knowledge Organization System), as well as a suitable vocabulary for representing subject authority data can also be used.

In our case study, the IFLA vocabularies, widely agreed upon by the library community, have been used to represent BNE entities in RDF. *datos.bne.es* is one of the first international initiatives to thoroughly apply the vocabularies developed by IFLA (Vila-Suero and Escolano, 2011); These vocabularies are FRBR, FRAD, FRSAD, and ISBD. The main reasons for selecting IFLA vocabularies are the following:

- The BNE has traditionally put significant effort in building an authority catalogue that describes not only subject headings, persons and organizations, but also titles (i.e. MARC 21 subfield $t); translations (i.e. MARC 21 subfield $l); parts of works (i.e. MARC 21 subfields $n and $p); or arranged statements for music (i.e. MARC 21 subfield $o). These authority records and the relationships between them can naturally be mapped to FRBR classes and relationships as Persons, Corporate Bodies, Works, and Expressions such as *is creator of* or *is embodied in*, among others. Table 3 shows the distribution of these entities within the catalogue data studied for datos.bne.es

- Bibliographic records can be naturally mapped to FRBR Manifestations and linked to the related authority records by FRBR relationships (IFLA, 1998). Additionally, the ISBD elements vocabulary provides a good coverage of the fields and subfields of MARC 21 bibliographic records and is intimately related to the cataloguing rules used by BNE.

- It is important to note that more general vocabularies, such as FOAF, BIBO, or even the EDM, do not offer straightforward mechanisms for representing the relationships between the aforementioned entities. For example, between a work and its translations, or between a person and a work.

| Label | URI | Type | Nº of times |
|---|---|---|---|
| Manifestation | frbr:C1003 | Class | 2,390,103 |
| Work | frbr:C1001 | Class | 1,969,526 |
| Person | frbr:C1005 | Class | 1,163,764 |
| Expression | frbr:C1002 | Class | 1,114,719 |
| Thema | frsad:C1001 | Class | 497,644 |
| Corporate body | frbr:C1006 | Class | 282,879 |
| language | dcterms:language | Relationship | 3,112,900 |
| is creator of | frbr:P2010 | Relationship | 2,129,222 |
| is created by (person) (person) | frbr:P2009 | Relationship | 2,129,222 |
| is embodiment of | frbr:P2004 | Relationship | 1,246,773 |
| is embodied in | frbr:P2003 | Relationship | 1,246,773 |
| is realized through | frbr:P2001 | Relationship | 1,054,736 |
| is realization of | frbr:P2002 | Relationship | 1,054,736 |
| same as | owl:sameAs | Relationship | 587,520 |
| subject | dcterms:subject | Relationship | 249,560 |
| has title of individual work by same author | isbd:P1117 | Property | 2,474,351 |
| has place of publication, production, distribution | isbd:P1016 | Property | 2,435,661 |
| has title proper | isbd:P1004 | Property | 2,390,161 |
| has specific material designation and extent | isbd:1022 | Property | 2,386,325 |
| has name of person | frbr:P3039 | Property | 1,163,764 |
| has title of work | frbr:P3001 | Property | 1,969,526 |

Table 3. Classes, relationships and number of times they appear within *datos.bne.es*

### *4.2. FRBR in a nutshell*

The study of FRBR was initiated by IFLA in the '90s and it follows entity-relationship techniques to identify the "things" that the bibliographic data describes, their attributes, and their relationships to other "things". As a result of this approach, the FRBR study proposes an entity-relationship model and a set of associated user tasks (find, identify, select and obtain). In this paper, we are mainly interested in *entities*, *relationships* and *attributes*.

The **entities** represent the objects of interest to users of library data and they are organized into three groups. However, in our case study we focus only on the following two groups:

- **Group 1 entities (Work, Expression, Manifestation, and Item)** represent different aspects of intellectual or artistic products. A *Work* is an abstract entity that defines a distinct intellectual creation, which is recognized through its individual realizations or expressions. An *Expression* is also an abstract entity and can take several forms, such as alphanumeric or musical. A *Manifestation* is the physical embodiment of a certain expression (e.g. a certain edition of the written form of a work). Finally, an *Item* is a concrete entity and represents a single exemplar of a manifestation (e.g., one of the copies of a certain edition of the written edition of a work).

- **Group 2 entities (Person and Corporate body)** represent the agents involved in the creation, distribution, and dissemination of intellectual products.

The model also defines the relationships among the entities. "Primary" relationships are those that link entities within the primary group (Group 1) and that are essential for the

organization of the bibliographic data proposed by the model. "Responsibility" relationships define core connections from primary entities to Group 2 entities.

Each of the entities defined in the model has associated a set of attributes. For instance, the entity Person has associated the following attributes: name of person, dates associated with the person, title of person, and other designation associated with the person.

Although the FRBR report has been around for more than a decade, its implementation on library systems is relatively limited (Hickey et al., 2002) (Hegna and Murtomaa, 2002) (Aalberg, 2008). The main problem behind its application lies in the difficulty of adapting existing catalogue data to FRBR due to the fact that very often higher-level entities like FRBR *Expression* or *Work* are not explicitly present within MARC-based catalogues, which are record-oriented and where one record can describe several distinct entities (e.g., the author, the manifestation, and even the associated expression and work).

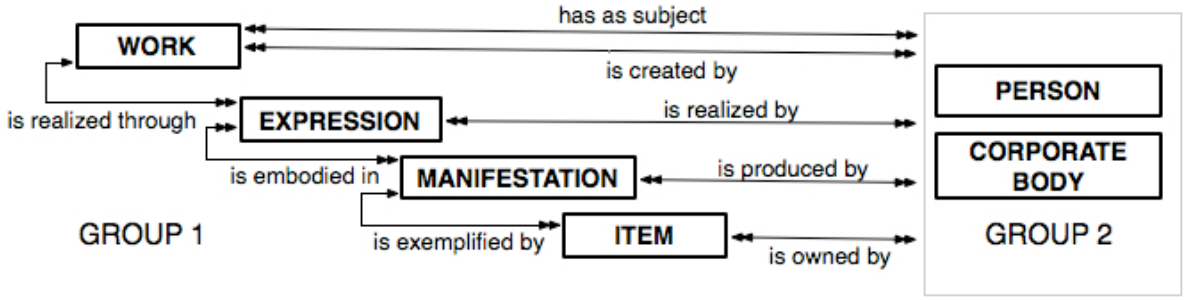


Fig. 3. Primary" and "Responsibility" relationships of FRBR Entities in Groups 1 and 2. (Adapted from (IFLA 1998))

### *4.3. Developing the vocabulary for transforming the data sources*

The goal of this task is to develop a vocabulary for modeling the data represented in MARC 21 records. This task can be decomposed into three steps: first, selecting the classes for modeling the entities (*person*, *work*, and *organization*) that appear in the records; second, selecting the properties for modeling the attributes of the entities (for example, *name of the person*, *title of the work*, and *location of the organization*); and third, selecting the relationships among the entities (a person *is creator of* a work, a work *is published by* an organization).

One important aspect is that the development of the vocabulary for *datos.bne.es* has been driven by the data sources and more specifically by the analysis of the usage of fields, subfields, and indicators across the BNE catalogue. Table 3 presents an overview of the most-used classes, relationships, and properties in the dataset; a high-level overview of the vocabulary developed for *datos.bne.es* is shown in Figure 4.

The classes *Manifestation, Work, Person, Expression,* and *Corporate body from FRBR* form the core of the vocabulary, whereas the class *Thema* from the FRSAD ontology and the class *Concept* from SKOS have been used to model the subject authority data.

The properties for describing bibliographic data have been reused from a number of

vocabularies, namely ISBD, RDA Group Elements 2, RDA Relationships for WEMI, Dublin Core terms, SKOS, and MADS/RDF; whereas the properties for describing authority data have been reused from FRBR, FRAD, FRSAD, and RDA Group Elements 2.

Regarding relationships, both the FRBR "Primary" (*is embodiment of, is embodied in, is realized through,* and *is realization of*) and the "Responsibility" (*is creator of* and *is created by*) relationships have been reused for relating authority and bibliographic data.

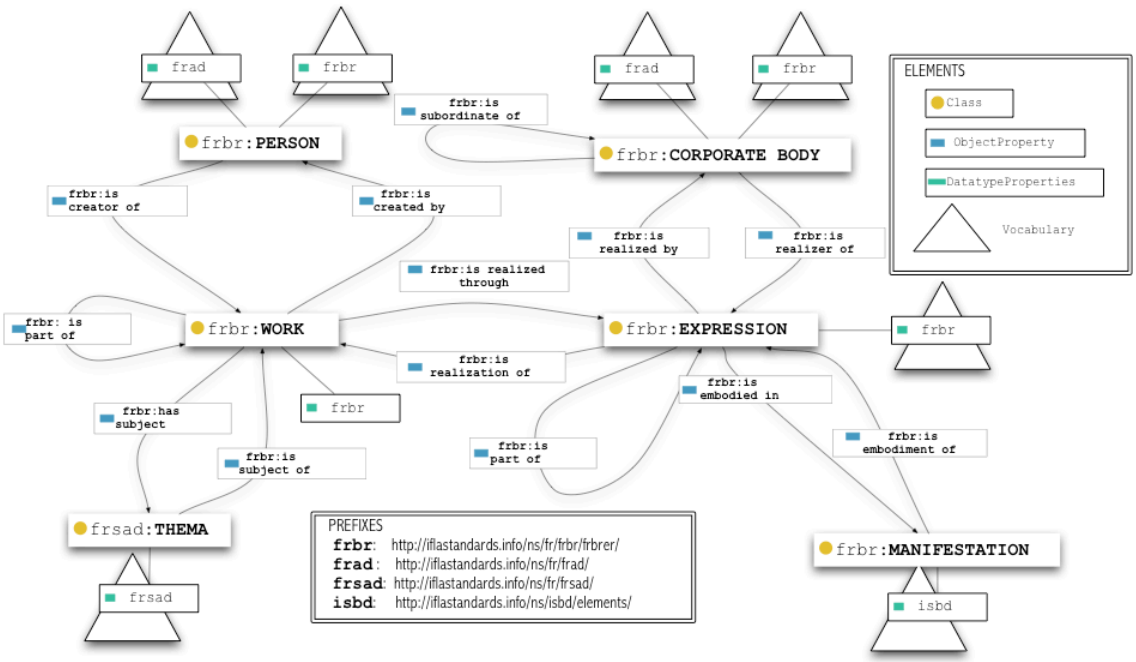


Fig. 4. An overview of the BNE vocabulary based on the FR family of models

### *4.4. Choosing the vocabulary for representing the provenance metadata*

Regarding the description of the provenance of resources, several vocabularies are available, for instance, OPMV (Open Provenance Model Vocabulary [xix]), PROV, and PROV-O, which are being standardized by the W3C [xx]. In *datos.bne.es* we describe the *metadata information* using the vocabularies OPMV and Dublin Core Metadata Terms [xxi].

## 5. Generation

The goal of this task is to transform the data sources into RDF following the decisions taken in the *specification* activity and according to the vocabulary designed and implemented in the *modeling* activity. This activity can be further decomposed into three tasks: first, selecting, extending or developing the technologies for transforming data sources into RDF (Section 5.1); second, mapping the data sources to the vocabulary concepts (Section 5.2); and third, transforming the data sources into RDF (Section 5.3).

### *5.1. Selecting, extending or developing the technologies for producing RDF*

The goal of this task is to identify the appropriate technological support for transforming the data sources into RDF.

In our case study, a review of the state of the art reveals that there are a number of tools for transforming MARC 21 records into RDF following the LD principles; these tools are the *COMET* tool and the *marc2rdf* script [xxii] for transforming bibliographic records with configurable mappings [xxiii]; and the XSLT sheet from the LoC [xxiv] that uses Dublin Core vocabularies. As for programming libraries, *Metamorph* [xxv] is a format-agnostic solution for transforming bibliographic metadata into RDF, which provides a scripting language to specify data transformations (with any vocabulary). It currently includes readers for common library formats (e.g. PICA+, MAB, MARC).

Given the specification (Section 3), the decision of modeling the data using FRBR (Section 4.1) and the review of the state of the art, we can say that the aforementioned solutions are (i) developer-oriented and thus difficult to use by non-technical users; (ii) not suitable for working with the FRBR data model; and (iii) not designed for working with authority and bibliographic records at the same time. As a result of this analysis, we have developed a tool, named MARiMbA, for *datos.bne.es* that fulfills the requirements.

### *5.2. Creating mappings between data sources and the domain vocabulary*

The goal of this task is to create the explicit mapping between the data sources and the domain vocabulary. In our case study, librarians and cataloguers map MARC 21 records to the RDFS/OWL vocabulary presented in Section 4.3 using MARiMbA. In this section we discuss, on the one hand, the mapping process from MARC 21 records to RDF by means of the example presented in Figure 5 and, on the other hand, the specific process followed for *datos.bne.es* using MARiMbA.

In order to facilitate the mapping to domain experts, MARiMbA (i) pre-processes the data sources and provides a summarization in a set of *mapping templates* for defining the mappings; and (ii) provides these *mapping templates* in the form of simple spread-sheets, so the domain experts do not have to learn complex mapping languages and can work with a relatively familiar and general-purpose tool. We decompose the process into (i) RDF Classification; (ii) RDF Description; and, (iii) RDF Interrelation. MARiMbA produces three types of *mapping templates*, one for each step. The steps and associated *mapping templates* are described in detail below.

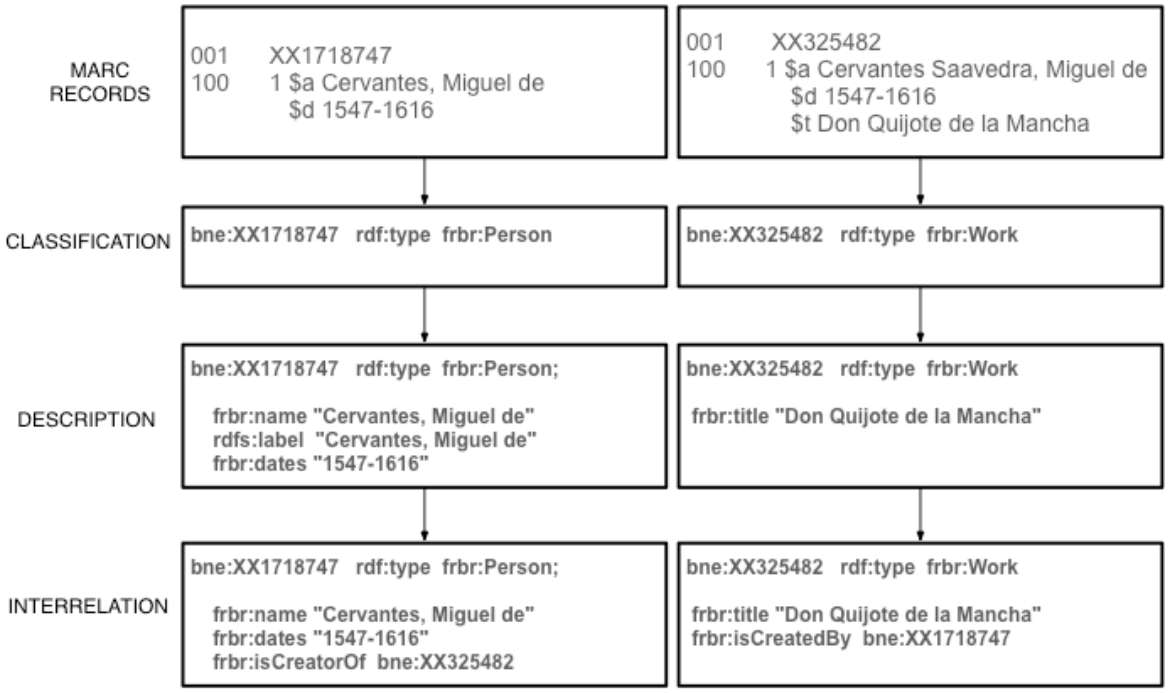


Fig. 5. An overview of a mapping process for authority records (*RDF statements in Turtle serialization*)

**RDF Classification.** Based on the combination of subfields in the main access point of the record (e.g., 100$a$d or 100$a$d$t) and given a record, this step will decide what type of RDF resource is generated (e.g., an *frbr:Person* and a *frbr:Work*). Specifically, the goal is to map a record, a portion of a record, or a combination of several records to one or more RDFS/OWL classes based on the characteristics of the record.

In *datos.bne.es* each MARC 21 record is mapped to one and only one RDFS/OWL class. We differentiate between bibliographic and authority records in the following way:

- The type of authority records is assigned based on the combination of subfields of the record main *access point* (i.e., the fields 1XX). An *RDF classification mapping template* for mapping authority records to classes is provided to the domain experts (see Figure 6). There is one sheet per access point (fields 100, 110, 111, 130, 150, and 151). The first column provides the combinations of the subfields found for the specific field (field 100 in the figure). The second column provides the records that contain that combination. The third column shows an example of the content found for that combination. Finally, the fourth column is where the domain experts assign the mappings from authority records to RDFS/OWL classes.

- No *mapping template* was used for bibliographic records because they are directly mapped to *frbr:Manifestation*. As discussed in Section 4, these records are naturally mapped to manifestations in the BNE catalogue**.**

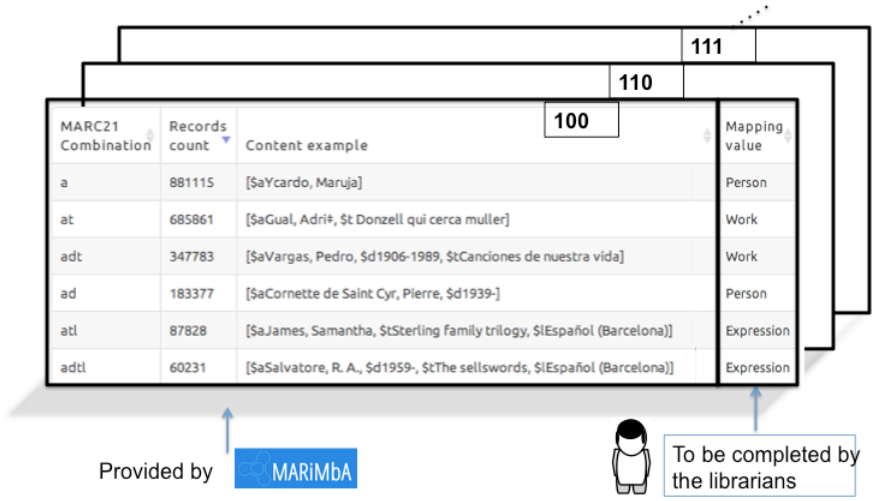


Fig. 6. Example of an *RDF classification mapping template for the field 100*

**RDF Description.** Given that the previous authority record (100$a$d) was mapped to the type *frbr:Person*, this step will decide if the subfield *100$a* is mapped to *frbr:nameOfPerson,* to *rdfs:label*, or to both. Specifically, the goal is to map the fields and subfields of records to one or more RDFS/OWL properties based on the characteristics of the record.

In *datos.bne.es* we use the following four approaches: first, one field can be mapped to one property; second, one field can be mapped to several properties; third, one subfield can be mapped to one property; and fourth, one subfield can be mapped to several properties. However, other approaches (e.g., a combination of subfields mapped to a property) are currently being explored for future versions. As in the previous step, domain experts are provided with a *RDF description mapping template* used with authority and bibliographic records.

Figure 7 presents the structure of the *RDF description mapping template* for the third approach. There is one sheet per entity (*Person*, *Work*, *Manifestation*, *Expression*, *Corporate Body* and *Thema*). The first column provides the combinations of field/subfield for the specific type of entity (Person in the figure). The second column provides the number of records that contain that combination. The third column shows an example of the content found for that combination. Finally, the fourth column is where the domain experts assign the mappings from MARC 21 fields and subfields to RDFS/OWL properties.

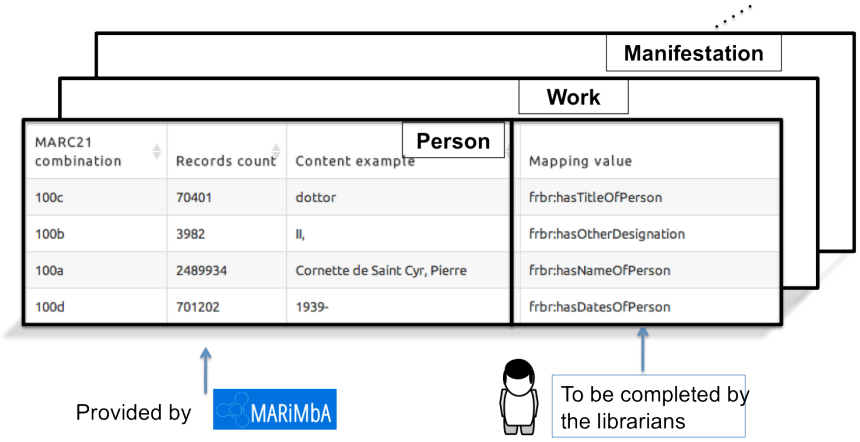

Fig. 7. Example of a *RDF description mapping template* for Person

**RDF Interrelation.** Given that one authority record (100$a$d) has been mapped to *frbr*:*Person,* and another record (100$a$d$t) has been mapped to *frbr:Work*, this step will decide whether a person is the creator of the work. In other words, if a relation between *frbr:Person* and *frbr:Work* should be established. The goal here is to create the mapping rules for establishing relationships between the RDF resources generated in the previous steps.

In *datos.bne.es* we use the FRBR data model for interrelating the RDF resources and focus on "Primary" and "Responsibility" relationships. In particular, we differentiate two cases:

1. To establish **relationships between bibliographic and authority records** we use *frbr:isEmbodiedIn* (between *frbr:Expression* and *frbr:Manifestation*). The relationship is established using a pointer found in the subfield *245$=* of the bibliographic record. This pointer contains a reference to an authority record (the value of the field 001). This record can be

- A *frbr:Work*: There is no *frbr:Expression* in the catalogue, so MARiMbA generates a new *frbr:Expression* with the language code found in the field 008 of the bibliographic record. Finally, the *frbr:Work* and the *frbr:Manifestation* are linked to the new *frbr:Expression.*

- A *frbr:Expression:* The *frbr:Manifestation* is directly linked to the *frbr:Expression.*

2. To establish **relationships between two authority records** (A1 and A2 in Figure 8), the domain experts use the *RDF interrelation mapping template.* The mapping is based on the main *access points* (e.g. "$a Cervantes, Miguel de" and "$a Cervantes, Miguel de $t Don Quijote"). First, the tool checks wether the access point of A1 (e.g., "$a Cervantes, Miguel de") is contained in the access point of A2 (e.g. *"$a Cervantes, Miguel de* $t Don Quijote"). If it is contained, the tool compares their combination of subfields (e.g., *a* and *at*) and extract what we call the *variation of subfields* (e.g., *t*). This *variation* is presented in the first column of Figure 8. There is one sheet per pair of entities (*Person-Person, Person-Work, Work-Expression, Work-Work,* and *Corporate body-Corporate Body*). The first column provides the *variations of subfields* in the main access points for the pair of entities (Person-Work in the figure). Finally, the second column is where the domain experts assign the OWL/RDFS properties that will be used for establishing a relationship between the pair of entities that presents that *variation of subfields.*

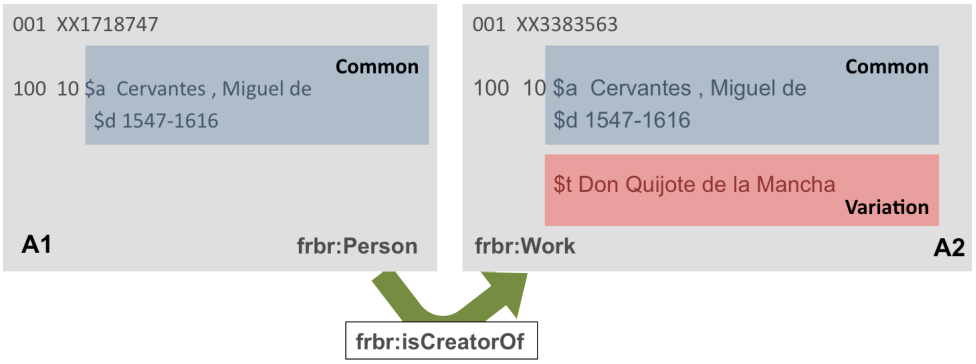

Fig. 8. Example of an RDF interrelation mapping process for authority records

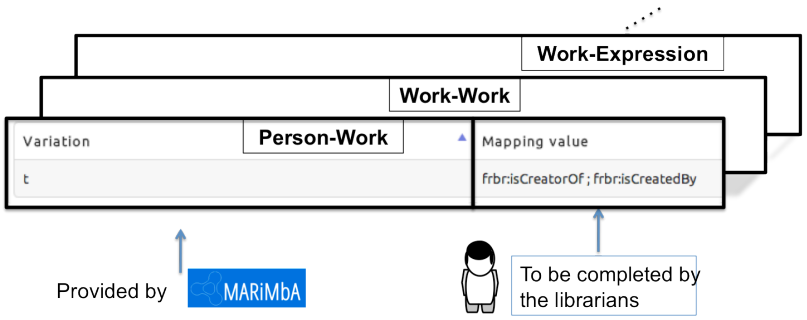


Fig. 9. Example of an *RDF Interrelation mapping template* for persons and works

Finally, it is worth noting that the mappings introduced by the librarians are validated by MARiMbA by checking that the URIs are valid (i.e., if they are present in the vocabulary from which they have been taken, and if they are not misspelled) and that only RDFS/OWL classes are used in the *RDF classification mapping template* and RDFS/OWL classes properties and in the *description* and interrelation *mapping template.*

### *5.3. Transforming the data sources into RDF using MARiMbA*

The final task in the *generation* activity is to automatically produce the *datos.bne.es* RDF dataset. For this MARiMbA takes the following inputs: (i) the MARC 21 data sources and the URI specification described in Section 3; (ii) the domain vocabulary presented in Section 4; and (iii) the *RDF lassification*, *description*, and *interrelation* mappings established by the librarians in the spreadsheets.

Figure 10 depicts the mapping and transformation processes. Given two records with the following *heading fields*: (i) 100 $a Cervantes Saavedra, Miguel de, and (ii) 100 $a Cervantes Saavedra, Miguel de $t Don Quijote de la Mancha, the process followed has three steps: first, the records are mapped to *frbr:Person* and frbr:Work respectively, based on the classification mapping; second, subfield $a is mapped to *frbr:nameOfPerson*, and the field $t to *frbr:titleOfWork*, based on the annotation mapping; and third, both resources are related through *frbr:isCreatorOf* after making a string comparison and analysis of their *variation of subfields* (100$a + $t) based on the relation mapping.

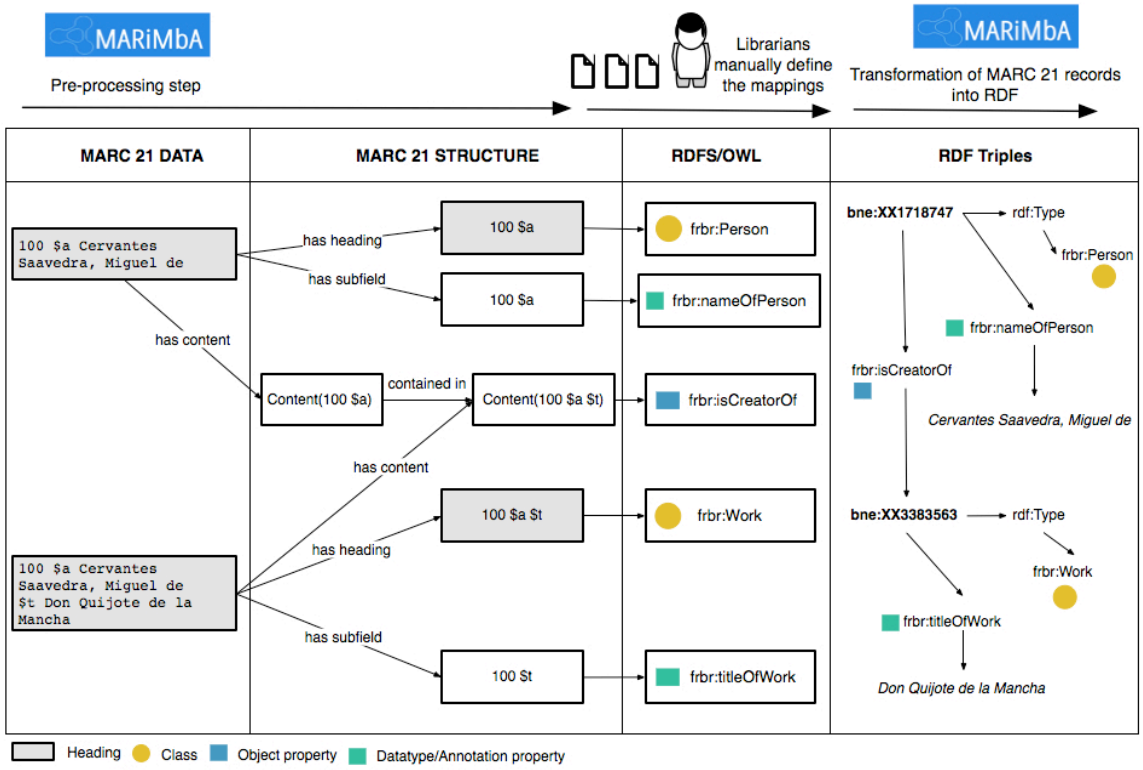


Fig. 10. A mapping and transformation process through a real example. Extended from Vila-Suero et al., (2013)

The website http://bne.linkeddata.es/mapping-marc21/ has been set up to provide more details about (i) the mapping and transformation processes; and (ii) the complete set of mappings used in the transformation of the RDF dataset.

Finally, we describe the license and provenance information in RDF and add this information to each RDF resource description by means of a mechanism called Named Graphs (Carroll et al., 2005) that is currently being studied for inclusion in the W3C RDF 1.1 recommendation [xxvi]. By including such information, we facilitate the automatic data processing by third party applications. Specifically, we provide (see Listing 1) (i) temporal information, including the date in which the data was retrieved (*prv:retrievedBy* and *prv:completedAt*); (ii) the publisher and creator of the data (*dcterms:publisher* and *dcterms:creator*); and (iii) the license (*dcterms:license*).

```
[     rdf:type prv:DataCreation ;
    prv:completedAt
    "2013-03-01T14:15:31.768Z"^^xsd:dateTime ;
    prv:retrievedBy
    [ rdf:type prv:DataAccess ;
      prv:completedAt
    "2013-03-01T14:15:31.768Z"^^xsd:dateTime ;
    ]
void:subset
  [ rdf:type void:Dataset ;
    dcterms:creator <http://oeg-upm.net#this> ;
dcterms:license <http://creativecommons.org/
publicdomain/zero/1.0/> ;
dcterms:publisher <http://datos.bne.es#org>
   ] ;

  ] ;
```

Listing 1. Example of license and provenance information

## 6. Linking

The goal of the linking activity is to include links from the *datos.bne.es* dataset into other relevant RDF datasets in order to allow the consumers to navigate related resources. This activity involves the automatic discovery of relationships between data items in order to increase the external connectivity of the RDF dataset. The activity is decomposed into three tasks: (i) identifying target datasets for linking (Section 6.1); (ii) discovering the outgoing links (Section 6.2); and (iii) validating the outgoing links (Section 6.3).

### *6.1. Identifying target datasets for linking*

The goal of this task is to identify datasets of similar topics or general datasets that can provide extra information to the dataset. The datasets can be looked up through data catalogs such as *datahub.io* [xxvii] or *datacatalogs.org* [xxviii].

In the datos.bne.es, therefore, we have focused on linking authority data (Persons, Corporate Bodies, Works, and Expressions). We have also decided to be linked with the libraries that are part of the VIAF dataset and that have published their authority data as LLD. Thus, we have selected the following datasets: (i) VIAF; (ii) DNB (GND, the authority RDF dataset); and (iii) Libris, and SUDOC. Additionally, as VIAF contains links to DBpedia, which falls in the general-purpose category, we have also selected it as a target dataset. Figure 11 depicts the target datasets using the resource of Miguel de Cervantes as the source of the links.

### *6.2. Discovering the links*

The goal of this task is to discover similar entities in the target datasets. There are several tools for creating links between data items of different datasets, such as the SILK framework (Volz et al., 2009). However, as VIAF mappings are available online [xxix] as a plain text file, MARiMbA generates the links with this mapping file. The rationale for not using tools like SILK is three-fold: first, VIAF mappings are authoritative and validated; second, reusing VIAF mappings speeds up the linking process; and third, the link generation is included in the same tool and users are not asked to learn how to use new software.

For generating the links, MARiMbA benefits from the fact that libraries have published their authority files by means of *natural keys* in order to build the URIs of their RDF resources. Therefore, MARiMbA generates the links by parsing the VIAF mapping file and prepending the namespaces to the different keys found in the file. Listing 2 presents the structure of the URIs created with this technique for the different target datasets.

For instance, we know that GND URIs follow the pattern gnd:{GND-ID} and that BNE URIs, the pattern bne:{BNE-ID}. Using these two URI patterns, we can establish links from *datos.bne.es* to GND by creating *owl:sameAs* statements with 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
```

Listing. 2. Example of an owl:sameAs link in *datos.bne.es*

With this technique, MARiMbA generated 587,52 equivalence outgoing links using the *owl:sameAs* object property. The numbers of links to each dataset are the following (in descending order): VIAF (454,068); DNB (76,413);DBpedia (36,431); Libris (10,884); and SUDOC (9,725).

### *6.3. Validating the links*

The goal of this task is to validate the links that have been created during the previous step. In the *datos.bne.es* case study, the links generated have been validated in VIAF and are reliably generated by MARiMbA in an automatic fashion. Therefore, no human supervised validation is needed.

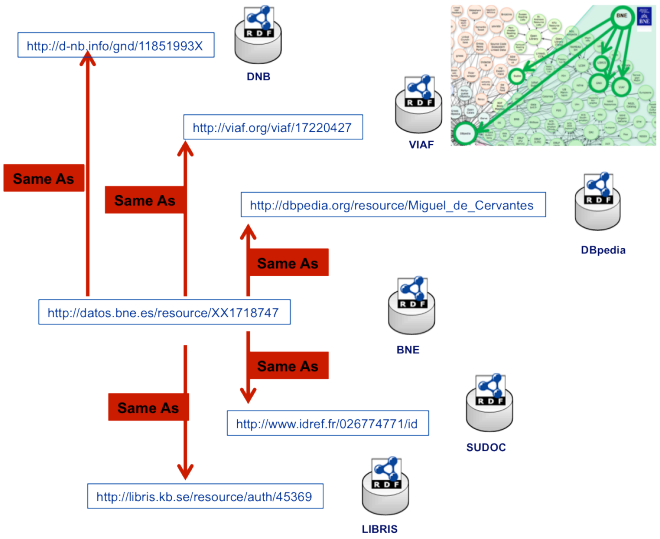


Fig. 11. Target datasets, URIs, and owl:sameAs links for Miguel de Cervantes

## 7. Data curation

The goal of this activity is to assess and ensure the quality of both the data sources and the LD published. Since data quality in the original data sources has a direct impact on the quality of the RDF generated, data curation is a crucial activity in the early stages of the LLD generation process. Therefore, one of the main contributions of our approach is to propose data curation as an activity to be carried out in parallel with the specification, modeling and generation activities, as graphically presented in Figure 1.

Regarding RDF data, there are already several works aiming at providing measures to evaluate the conformance of these data to the LD principles, For instance, Hogan et al., (2012) empirically evaluate a set of concrete guidelines. In *datos.bne.es* we have validated the conformance with these state-of-the-art guidelines. Therefore, in the following sections we will focus exclusively on the data source curation.

The task of data source curation is decomposed into three-subtasks: identifying the data issues; reporting the data issues; and fixing the data issues.

**Identifying data issues.** The LLD generation process, concerning the application of

semantically richer models (e.g. FRBR) and the participation of cataloguing experts, brings a good opportunity to assess, report and fix issues in the MARC 21 data [xxx]. Therefore, for the *datos.bne.es* case study generation process we have identified the following type of issues:

**-** *Coding errors*. The most common issue that emerged from the mapping templates generated by MARiMbA was the incorrect use of certain MARC 21 subfield codes. For instance, in the first iteration the *classification mapping template* showed the combination of subfields 100 $a$f and provided an example (**$a** Chicin, Fred, **$f** (1954-2007)). The librarians were able to identify this incorrect use (note that the correct subfield code is $d and that **f** is the starting character of ***f****echas* - ***d****ates* in Spanish). Other examples of problematic issues found were the following: the *absence of subfield codes* (e.g. 100 $**A**strain, Miguel María), or the *absence of subfield delimiters* (e.g. 100 $aMoix, Llàtzer, $d1955-**t**La Costa Brava, $l Catalán).

- *Format errors*. This type of issue is related to the format and encoding of MARC 21 records. In this regard, two issues were found: first, the content of certain records was not correctly encoded (e.g. 100 $l Espa**Û**ol); and second, the usage of *content designators* did not comply with the MARC 21 format specification (e.g. a high number of records contained an *indicator* in the field 001).

- *Issues derived from the application of FRBR.* Perhaps the most interesting type of issues was related to the application of FR models. In this regard, the most relevant issues found were the following:

  - *Non-compliant records according to FRBR*. For instance, in the *classification mapping* for the field 100, the combination $a$l could not be classified into any of the FRBR entities. The mapping revealed that the language subfield (i.e., $l) was used for including the title information (i.e. $t) and showed the following example: "$a Geary, Rick, **$l** A treasure of Victorian murder".

  - *Authority control issues.* These issues arose especially during the *interrelation* step of the mapping process. Specifically, these issues were related to problems concerning the linking of the *manifestations* to their respective *expressions*. For instance, several thousands of bibliographic records could not be linked to their expression. After evaluation, it was found that there was no authority record created in the authority catalogue for the expression of a work in the original language (e.g., an expression of Don Quijote in Castilian, the original language).

**Reporting and fixing data issues.** In order to report coding errors, format errors, and non-compliant records, MARiMbA automatically generated reports for those content designators that were identified as errors by the librarians in the mapping templates. The report included the list of record identifiers (field 001) classified by the error that was found. In total, MARiMbA reported issues on more than two thousand authority records, and more than twenty thousand bibliographic records. The list of record identifiers and types of issues helped the BNE IT team to automatically fix most of the issues, while other less important issues (e.g., absence of subfields) were assigned to cataloguers to fix them manually.

Regarding authority control issues, MARiMbA automatically reported the issues found in the interrelation step (limited to the problem of linking manifestations to their expressions). The BNE cataloguing experts are currently studying these issues in order to apply changes to the catalogue.

## 8. Publication

The goal of the publication activity is to make available and discoverable on the Web the RDF dataset. This activity is decomposed in three main tasks: *publishing the dataset on the Web*; *publishing metadata describing the dataset*; and *enabling effective discovery of the dataset*.

The first task is to make the dataset available on the Web. In the datos.bne.es case study, we make data available with (I) a SPARQL endpoint that can be accessed under the following URI http://datos.bne.es/sparql; (ii) the LD front-end Pubby, which provides HTTP content-negotiation; and (III) an API (Application Programming Interface) (under http://datos.bne.es/frontend/persons) using Puelia [xxxi], an implementation of the Linked Data API that provides HTTP access to resources and features lsuch as paging or filtering by RDF properties.

The second task is to publish metadata about the dataset. For this purpose, in datos.bne.es we make available the description of the dataset, using VoID [xxxii], a vocabulary for describing RDF datasets. The file can be accessed at http://datos.bne.es/void/bne.ttl.

The final task is to facilitate the reuse by third parties, allowing them to discover the dataset. For this purpose, datos.bne.es is registered in datahub.io under http://datahub.io/dataset/datos-bne-es).

## 9. Exploitation

The goal of the exploitation activity is to develop applications and services that exploit the data and provide rich interfaces to both end users and developers.

Within the context of the *datos.bne.es* use case, we provide the following different domain-specific applications and services:

- **http://datos.bne.es/frontend**: This service provides an API to access and retrieve the data. Its main purpose is to make the usage of data easier for web developers.

- **http://bne.linkeddata.es**: The pilot allows searching for and navigating through authors, their works, and the different translations and editions.

- **http://bne.linkeddata.es/graphvis**: This visualization shows the potential of using graph analysis visual tools to explore the RDF graph data produced according to FRBR. The visualization allows the user to search for and navigate through data related to Miguel de Cervantes.

## 10. Conclusions

In this paper we have discussed the main characteristics of the process followed for the development of the *datos.bne.es* case study. We have also presented MARiMbA, the tool for

transforming MARC 21 records into RDF, linking the dataset with other resources, and reporting issues in the source records. Further, we have defined and discussed a method for generating LLD by means of MARC 21 records and applied the method to real data following an iterative and incremental development lifecycle.

The method here shown is based on previous experiences and guidelines that have been applied to other knowledge domains. Throughout the paper we have demonstrated, on the one hand, how general guidelines can be applied to library data and, on the other hand, we have discussed and extended those activities and steps that are unique to the library domain. In this respect, one of the most interesting aspects of the publication of Linked Data out of current library catalogues is the positive impact that LD principles, such as the use of URIs instead of strings or the concept of formally typed resources, may have on the information architecture of libraries. Examples of this beneficial impact are the inclusion at BNE of resolvable URIs to equivalent resources in external data sets during the cataloguing process (e.g., the inclusion of VIAF or DBpedia URIs when creating a new authority record) or the reorganization that BNE's catalogue is undergoing based on the experiences in the application of FR models in datos.bne.es

Furthermore, in the paper we have shown how to include domain experts in the LLD generation process, thus reducing considerably the cost of mapping the data sources to the RDFS/OWL vocabulary and improving the quality of the mappings. In order to facilitate their participation, MARiMbA drives the mapping process by analyzing the data sources and producing a set of spreadsheets easy to understand and use by library experts. One of the main outcomes of the datos.bne.es project has been the cross-fertilization among the semantic web developers and the library experts, which has resulted in a solid team and in several training courses dedicated to Linked Data have been established within BNE.

In the paper, we have also proposed the *data sources curation* as a crosscutting activity performed in parallel with the specification, modeling, and generation activities. By reporting and fixing issues in the data sources, we increase the quality of the RDF data and the data sources, thus saving costs for the institution. As has already been discussed within the paper, this initiative is still in its infancy but we believe that LLD publication can help to create a "virtuous cycle " that can directly impact on the quality of library data.

Additionally we would like to highlight that datos.bne.es is a living project with many challenges ahead. The project is slowly achieving a number of its initial goals such as improving the interoperability of the catalogue data and positioning the BNE as a high quality data provider [xxxiii]. The next challenges for the project are to promote the use of LLD within internal and external contexts. Regarding internal contexts, the main priority will be the interaction with digital resources from the digital library. As for external contexts, the first step will be the development of a portal for end-users that leverages the potential power of the current BNE graph and improves the interaction with and retrieval of BNE information.

As a closing remark, we believe that the experience and results detailed in this paper can serve as guide and a baseline for future research and development projects and help other institutions on their way to Library Linked Data. More importantly, and in line with the Open Data principles, we have made the results publicly available and accessible on the Web under a public domain license and provided a discussion of the main steps performed to produce them.

## References


Aalberg, T. (2006), "A process and tool for the conversion of MARC records to a normalized FRBR implementation". In Proceedings of the 9th international conference on Asian Digital Libraries: achievements, Challenges and Opportunities. ICADL'06, Berlin, Heidelberg, Springer-Verlag, pp. 283-292.

Alexander, K., Cyganiak, R., Hausenblas M., and Zhao J. (2011), "Describing Linked Datasets with the VoID Vocabulary". W3C Interest Group Note, (http://www.w3.org/TR/void/). Last viewed 12- 01-2013

Baker T., Bermes E., Coyle K., Dunsire G., Isaac A., Murray P., Panzer M., Schneider J., Singer R., Summers E., Waites W., Young J., and Zeng M. (2011), "W3C Library Linked Data Incubator Final Report", W3C, 25 October.

Davis I. and Dodds L. (2010). "Linked Data Patterns". Online resource (http://patterns.dataincubator.org/book/). Last viewed 12-01-2013

Harper, C. and B. Tillett B. (2007), "Library of Congress controlled vocabularies and their application to the Semantic Web." Cataloging and classification quarterly 43.3-4: 47-68.

Heath T. and Bizer C. (2011), "Linked Data: Evolving the Web into a Global Data Space (1st edition)". Synthesis Lectures on the Semantic Web: Theory and Technology, 1:1, 1-136. Morgan & Claypool.

Hegna, K., Murtomaa, E. (2002), "Data Mining MARC to find: FRBR?" In: Proceedings of the 68th

20 IFLA General Conference and Council. IFLA, The Hague, Netherlands.

Hickey, T.B., O'Neill, E.T. and Toves, J. (2002), "Experiments with the IFLA Functional Requirements for Bibliographic Records (FRBR)". D-Lib Magazine 8

IFLA (1998), "Functional Requirements for Bibliographic Records: Final Report". IFLA Study Group on the Functional Requirements for Bibliographic Records. KG Saur. Munich.

IFLA (2011), "ISBD: International Standard Bibliographic Description: Consolidated Edition". De Gruyter Saur. Berlin.

International Standards Office (1996), "ISO 2709:1996 Information and documentation - Format for Information". Geneva.

Isaac, A. and Haslhofer, B. (2013), "Europeana Linked Open Data –data.europeana.eu". Semantic Web Journal, to appear. Available from http://www.semantic-web-journal.net/.

Malmsten, M. (2008), "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.

Summers, E., Isaac, A., Redding, C., and Krech, D. (2008), "LCSH, SKOS and Linked Data". In Proceedings of the 2008 International Conference on Dublin Core and Metadata Applications, DCMI '08, pages 25–33.

Vila-Suero, D. and Escolano, E. (2011), "Linked Data at the Spanish National Library and the Application of IFLA RDFS Models". In IFLA SCATNews Number 35.

Vila-Suero, D. (2001), "W3C Library Linked Data Incubator Group: Use Cases". W3C, 25 October.

Vila-Suero, D., Villazón-Terrazas, B. and Gómez-Pérez, A. (2013), "datos.bne.es: A library linked dataset". Semantic Web Journal, to appear. Available from http://www.semantic-web-journal.net/.

Villazón-Terrazas, B., Vilches, L., Corcho, O., Gómez-Pérez, A. (2011), "Methodological Guidelines for Publishing Government Linked Data". In Wood, D. (Ed.), *Linking Government Data.* Springer, Berlin, pp-27-49.

Volz J., Bizer C., Gaedke M., Kobilarov G. (2009), "Silk – A Link Discovery Framework for the Web of Data". 2nd Workshop about Linked Data on the Web (LDOW2009), Madrid, April.


i http://www.dnb.de/EN/Service/DigitaleDienste/LinkedData/linkeddata_node.html

ii http://data.bnf.fr

iii http://bnb.data.bl.uk/

iv http://viaf.org

v http://www.clir.org/pubs/reports/pub152/LinkedDataWorkshop.pdf

vi http://www.loc.gov/marc/transition/news/framework-103111.html

vii http://bibframe.org

viii http://www.loc.gov/z3950/agency/

ix http://marimba4lib.com

x http://www.loc.gov/marc/specifications/

xi http://catalogo.bne.es/uhtbin/authoritybrowse.cgi?action=display&authority_id=XX1718747

xii In the work presented here there is a parallelism between RDF classes and FRBR entities, and between RDF Properties and FRBR attributes and relationships. In OWL terminology the parallelism would be between OWL classes and FRBR entities, between OWL datatype properties and attributes, and between OWL object properties and FRBR relationships.

xiii The selection and reuse of vocabularies will be discussed in Section 4

xiv http:/prefix.cc

xv http://data.europeana.eu

xvi http://openmetadata.lib.harvard.edu/

xvii http://creativecommons.org/publicdomain/zero/1.0/

xviii http://dublincore.org

xix http://open-biomed.sourceforge.net/opmv/ns.html

xx http://www.w3.org/2011/prov

xxi http://dublincore.org/documents/dcmi-terms/

xxii https://github.com/digibib/marc2rdf

xxiii https://github.com/edchamberlain/COMET

xxiv http://www.loc.gov/standards/marcxml/xslt/MARC21slim2RDFDC.xsl

xxv http://culturegraph.org

xxvi http://www.w3.org/TR/rdf11-concepts/

xxvii http://datahub.io

xxviii http://datacatalogs.org/

xxix http://datahub.io/dataset/viaf

xxx Please note there are commercial quality control tools for MARC 21 data (e.g. *MARC Report* [xxx]) that already provide some of the characteristics defined here (specifically coding errors and format errors). And when the project started these tools where not available in BNE.

xxxi See https://code.google.com/p/puelia-php/

xxxii http://www.w3.org/TR/void

xxxiii The published data is already used by other Spanish institutions like Consorcio Madroño, a consortium of all public university libraries of Madrid, and projects like vestigium.org, a small digital library collecting the works of scientists and humanists from the region of Valencia.