Semantic Web 0 (0) 1–12
IOS Press

Migration of A Library Catalogue into RDA Linked Open Data

Editor(s): Christoph Schlieder, Universität Bamberg, Germany
Solicited review(s): Carlo Meghini, Istituto di Scienza e Tecnologie dell'Informazione (ISTI-CNR), Italy; Trond Aalberg, Norges Teknisk-Naturvitenskapelige Universitet, Norway

Gustavo Candela [a,*,**], Pilar Escobar [b], Rafael C. Carrasco [b] and Manuel Marco-Such [b]
[a] *Biblioteca Virtual Miguel de Cervantes, Alicante, Spain*
[b] *Departamento de Lenguajes y Sistemas Informáticos, Universidad de Alicante, Alicante, Spain*

Abstract. The catalogue of the Biblioteca Virtual Miguel de Cervantes contains about 200,000 records which were originally created in compliance with the MARC21 standard. The entries in the catalogue have been recently migrated to a new relational database whose data model adheres to the conceptual models promoted by the International Federation of Library Associations and Institutions (IFLA), in particular, to the FRBR and FRAD specifications. The database content has been later mapped, by means of an automated procedure, to RDF triples which employ basically the RDA vocabulary (Resource Description and Access) to describe the entities, as well as their properties and relationships. This RDF-based semantic description of the catalogue is now accessible online through an interface which supports browsing and searching the information. Due to their open nature, these public data can be easily linked and used for new applications created by external developers and institutions. The methods applied for the automation of the conversion, which build upon open-source software components, are described here.

Keywords:, Linked Open Data, Bibliographic and Authority Data, Cultural Heritage, Semantic Web

1. Introduction

Applying the *linked open data* concepts to the cultural heritage domain has become an active and challenging field [22]: many libraries, museums, and archives are currently exploring ways to convert their data into RDF[1] and to develop new interfaces providing a richer experience to the users of cultural heritage websites.

In parallel, modern standards for cataloguing are emerging as an alternative to the traditional ones (such as ACCR2 [3]). For example, RDA (Resource, Description and Access) is a cataloguing standard [17] for descriptive metadata supporting resource discovery. RDA follows the concepts and terminology of the Functional Requirements for Bibliographic Records (FRBR, [15]) and the Functional Requirements for Authority Data (FRAD, [24]) —and it is working to adopt the Functional Requirements for Subject Authority Data (FRSAD, [20])—, a family of models promoted by the IFLA which define entities, relationships, and attributes that should be used to describe resources. Recently, a linked data and semantic web representation of the elements and relationships of RDA was published.[2]

This paper describes the steps applied for the automation and control of the migration process from a MARC21 collection of records to a set of RDF triples containing bibliographic metadata in RDA, schematically represented in figure 4. The process relies on the creation of a relational database according to the

*Corresponding author.
**e-mail: gustavo.candela@cervantesvirtual.com

[1] The Resource Description Framework (`http://www.w3.org/RDF`) is a graph based data model which is widely used for semantic web and linked data applications.

[2] http://www.rdaregistry.info

FRBR family of conceptual models, and provides controlled generation of linked data in RDA. The implementation is strongly based on the currently available open-source technology.

2. Related work

Many libraries and organizations are in the process of transforming their legacy metadata into various RDF-based semantic descriptions, mainly FRBR-based. An early survey on FRBRization techniques was prepared by the Online Computer Library Center (OCLC) [9]. A more recent survey [7] provides a taxonomy of semi-automated techniques based on three criteria: type of FRBRization (methods), model expressiveness and specific enhancements to improve quality or performance.

Usually, the FRBRization builds an FRBR catalogue by applying mapping rules between the source bibliographic metadata and the FRBR attributes. For example, the TELPlus prototype developed an FRBR repository for the European Library [12,21] by applying rule-based interpretation of fields enhanced with cluster deduplication and evaluation metrics.

The LC Display Tool provided by the Library of Congress [27] was a simple XSLT template which transforms MARC data into XML and HTML formats. This approach can lead to very large files (due to the rich variety of relationships available in FRBR) which are difficult to visualize.

A different approach based on musical content was implemented at the Indiana University Library with the Variations project [26] where several XML schemas where used to publish FRBR records[3]. An interface[4] was created by the project in order to retrieve and explore the catalogue.

LibFRBR is a toolkit which can be used to convert bibliographic records into FRBR structures based on the Koha open-source integrated library system and also provides an interface for library cataloguers [6].

FRBR-ML [30] is based on an XML intermediate model which was designed to ease exporting data in various semantic formats. This tool takes MARC-XML records as input and produces a set of FRBR records and their relationships. The output is semantically enriched by linking external information sources.

The GLIMIR project [13,32] has developed software to create clusters of creations within WorldCat[5] which are manifestations of a single expression or expressions of a single work.

Some initiatives such as the RDA Steering Committee (RSC) and the International Working Group on FRBR and CIDOC CRM Harmonisation, are defining metadata according to international models for user-focused linked data applications. In January 2014, the RDA Steering Committee published stable forms of RDA elements and controlled vocabularies. These vocabularies provide elements, guidelines, and instructions based on FRBR principles. RDA elements applies to each of the FRBR entities as RDF properties and sub-properties, and a set of RDA values vocabularies to populate specific RDA elements such as carrier type or media type.

FRBROO[6] is an elaborated version of FRBR implemented as an extension of CIDOC CRM. The FRBROO ontology facilitates the interchange of bibliographic and museum information.

An increasing number of cultural institutions are applying semantic web technologies and creating *linked open data* projects. For example, the Library of Congress Linked Data Service (`id.loc.gov`) provides access to authority data such as the LC subject headings and the MARC geographic areas.

The Bibliothèque nationale de France published `data.bnf.fr` in 2011 by aggregating information about authors, works, and subjects which was scattered among various catalogues. These data are published in RDF using a vocabulary based on the FRBR model where objects are referenced through ARK identifiers.[7] The information is stored in a database which contains the data in different formats, including RDF, JSON, and HTML. [28]

The British National Bibliography Linked Data Platform (`bnb.data.bl.uk/docs`) provides access to the British National Bibliography (BNB), implements the SPARQL query language [25] and delivers RDF and JSON outputs. The dataset has been modelled using existing RDF vocabularies, such as Dublin Core, the Bibliographic Ontology (BIBO), and Friend of a Friend (FOAF). Exceptionally —for example, due to insufficient granularity of those vocabularies— a new term was coined and documented. FRBR was not

[3] http://www.dlib.indiana.edu/projects/vfrbr/

[4] Scherzo, `http://webapp1.dlib.indiana.edu/scherzo`

[5] `https://www.worldcat.org`

[6] http://www.ifla.org/files/assets/cataloguing/frbr/frbroo_v2.2.pdf

[7] The *Archival Resource Key* identifiers are persistent references to web-accessible objects.

initially used [8], since the identification of the entities in the source MARC records required extensive work. The records were therefore normalized for improved matching and later transformed into RDF using XSLT and Jena Eyeball.

The German National Library supplies its data in the RDF standard via its Linked Data Service (LDS; `http://www.dnb.de/EN/lds`) since 2010. The vocabulary is based on Dublin Core and BIBO and complemented with some elements from other vocabularies, for example, RDA, ISBD (International Standard Bibliographic Description), and GND (Gemeinsame Normdatei). The records can be also retrieved in BIBFRAME format, an RDF-based replacement for MARC21. The National Library of Spain (BNE) has recently migrated its databases to RDF and published [23] them at `datos.bne.es`. The transformation is assisted by specific software [33], which supports RDF generation from MARC21, and the vocabulary is strongly based on FRBR and ISBD.

The Europeana linked data at `data.europeana.eu` ensure a high level of consistency and interoperability by abstracting the original data to a common format (the Europeana Data Model). Unfortunately the richness of the original descriptions is partially lost in the homogenization process.

3. The transformation process

Traditionally, the descriptive metadata of bibliographic content —stored, for example, in MARC records— were created and interpreted by humans. Even if those records followed cataloguing rules such as AACR2 and ISBD [29], the textual descriptions therein could not be easily read and interpreted by computers, see for instance, the rich description under field 534 in figure 1. The FRBR family of conceptual models and the RDA specification provide a modern framework which facilitates the automatic processing of the information. However, the transformation of the old records into the new format has a significant cost and is not an easy task [1], since libraries usually host large catalogues which should be manually revised. Therefore, software tools for the automation of the migration process are called for, and the experience of the Biblioteca Virtual Miguel de Cervantes in their implementation is described below.

```
001 ff97f774-82b1-11df-acc7-002185ce6064
003 BVC
041 $aspa
080 821.134.2-2"16"
100 $aVega, Lope de, $d1562-1635
245 $aEl caballero de Illescas
    $h[Libro electrónico]
    $c/Lope de Vega
260 $aAlicante
    $bBiblioteca Virtual Miguel de Cervantes
    $c2002
534 $aPublicación original:
    Madrid, por Juan de la Cuesta,
    a costa de  Miguel de Syles, 1620.
650 $aTeatro español $ySiglo 17º)
700 $aCuesta, Juan de la, $d 1604-1627
    $eimpresor
700 $aSiles, Miguel de $eeditor
```

Fig. 1. A MARC21 record for the novel *El caballero de Illescas*.

3.1. Preprocessing of sources

A MARC21 record describes one entry in the bibliographic catalogue or authority file,[8] and consists of text fields which are identified by a three-digit number, see figure 1. The text in one field can be split into sub-fields which are distinguished with a dollar sign followed by a single-character identifier. Since some fields are required (for example, field 245 containing the title) while some others are optional or user-defined, the homogeneity of the data across libraries cannot be guaranteed. Furthermore, the content of a field can be expressed with different conventions, in different languages, or it may contain typos: these features represent a challenge when MARC21 records must be shared between libraries.

The transformation of MARC records into FRBR is not a simple task [1]. Some issues are common, see [2,21], while some other are particular to a library. For example, the 200,000 records in the Biblioteca Virtual Miguel de Cervantes are provided by a large number of institutions in Spain and Latin America where variable cataloguing practices are applied. Some of the challenges and the measures taken are listed below.

a) *Missing or inconsistent uniform title.* Uniform titles identify expressions or manifestations of a single work. However, often the uniform title is missing (only 2% of records contain a uniform title) or it has been inappropriately selected (for example, the source language has been sometimes appended to it, as in *Don Quijote de la*

[8] An authority files compiles the unique terms and possible variations used to describe names, titles, and subjects.

Mancha, inglés). Since records with identical uniform title are not guaranteed to describe the same work, the preferred title has been used to cluster works instead. Further work will be required to obtain a wider granularity.

b) *Variable encodings.* Some information is encoded using different fields at different institutions. For example, the MARC control number (needed to link back the original record) has been found under fields 001, 856 and 909. Works using multiple languages have been sometimes encoded using multiple language subfields (one per language) while in other cases a single subfield lists all the language codes with custom separators. Specific rules have been created to parse the records with a common provenance.

c) *Markup errors.* MARC tags are introduced manually and therefore, a number of mistakes can be expected. For example, some titles (MARC field 245) include a responsibility statement after the ISBD separator (a slash) instead of the required MARC prefix $c. The parser compiles a list of rules in order to handle such mistakes or inconsistencies.

d) *Textual errors.* Many titles were found to contain spurious characters or unbalanced parenthesis. The migration also allowed to identify such typos and improve the normalization of titles.

e) *Multiple publication statements.* Statements about publishers and distributors (MARC code 260) are not distinguished when the exportation employs a DublinCore-based gateway. For example, publication date of the source work and its digital version are both tagged dc:date. Again, specific rules for each provenance have been implemented.

f) *Unspecified roles.* Secondary personal entries sometimes contain no information about the role played in the creation. By default, the contributor is associated to a particular manifestation (as in the case, for example, of a publisher). A set of rules has been defined for those cases where the context helps in the interpretation of the content. For example, the keyword *trad.* indicates a translator which must be associated to an expression.

g) *No unique identifiers for creators.* Since no authority record number, such as VIAF or ISNI, has been associated to the creators, ambiguity arises when authors have identical names. Also similar names may correspond to a single author due to name variations and typos. An open-source software [5] has been applied in order to identify names which indeed correspond to a single person.

h) *Analytics cataloguing.* Analytics cataloguing creates separate record for each item found in a larger resource, such as article within a journal, newspaper or serial. The information in MARC field 773 (host item) has been parsed in order to detect if the host resource is in the library catalogue. In such case, a relation isPartOf is added to the database.

The pre-processing applied a set of parsers (implemented in Java upon the the MARC4j library) to normalize the information contained in data fields such as titles, roles or languages.

3.2. Definition of an FRBR-FRAD relational model

The FRBR family of conceptual models [15] are intended to be independent of any cataloguing code or implementation and they identify the principal entities, their attributes and the relationships between them. The FRBR model defines the products of intellectual or artistic endeavour (work, expression, manifestation, and item) and is complemented with the FRAD model, which defines the entities responsible for the content (person, family, and corporate body), and with the FRSAD model, which defines the entities that serve as the subjects of creations (concept, object, event, and place), see figure 2.

Traditional data storage systems, in particular relational databases, are much more mature than semantic ones, and they offer reliable, extensively tested implementations. Inspired by the IFLA conceptual models, an Entity-Relationship (ER) model, schematically represented in figure 3, has been defined to store the Biblioteca Virtual Miguel de Cervantes descriptive metadata. Some additional elements were incorporated to the model in order to address the catalogue specificities. For example, *Collection* entities were needed to host arbitrary groupings of objects, such as works in a bibliography, items with a common provenance (e.g., a partner library holdings or items in a personal archive), which are not properly creations and usually have no associated descriptive metadata. Since authors are often the subject of a book in a library with a focus on literature, the subject element from Dublin Core have been used to describe creations having a particular agent as *subject*; conversely, agents play different *roles* when contributing to a document such as printer,

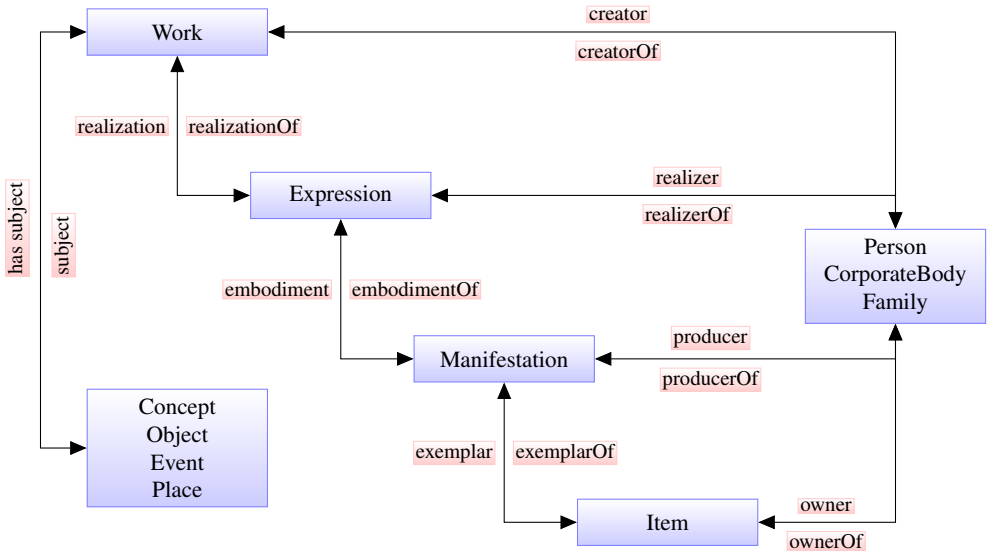

Fig. 2. Entities defined in FRBR (Work, Expression, Manifestation, Item), FRAD (Person, CorporateBody, Family), and FRSAD (Concept, Object, Event, Place) with their primary relationships.

editor or illustrator. A generic relationship between entities (*partOf*) was defined in order to describe nested inclusions, for example, journals publishing volumes, made of issues containing articles. The online manifestations are connected to their URL with the *homepage* attribute. Entities for the UDC and Unesco classifiers and for VIAF persons were also added to the model. Since the RDA technical guidelines were created while several aspects of FRBR were still in flux, they include some additional entities (such as *Agent*) and rename some relations: for example, the FRBR *embodiment* becomes *manifestationOfExpression* in RDA, see figure 5.

As can be seen in figure 3, the abstract class *creation* generalizes the basic FRBR entities (*work*, *expression* and *manifestation*). This class has been added in order to avoid redundant descriptions and duplicate coding, since many properties, such as subject, are common to all types of entities.

3.3. Migration of MARC records into the FRBR database

The application of the FRBR model to an existing MARC collection needs to identify, create and connect FRBR entities [2]. Once the MARC records were normalized and enhanced through the applications of the actions listed in Section 3.1, the transformation was implemented in three consecutive steps:

a) Identification of FRBR entities.
b) Extraction of relationships between entities.
c) Semi-automatic clustering of entities.

The sequential nature of the migration process allows for simple incremental construction and update.

The identification of FRBR entities required the implementation of a detailed mapping between the original metadata and the FRBR attributes, in particular for those records containing multiple references to persons, subjects or related works. Duplications were minimized by searching for creators with similar names and compatible dates [5]. In parallel, complex subject headings were decomposed into their elementary components to reduce the number of different subject entities.

The extraction of relationships identifies connections, mainly involving works, as in *creator* or *subject* elements, but also expressions —for example, *translators* and *editors*— and, in a small number of cases, manifestations —for example, *printers* and *illustrators*. Relationships between complex works (for example, a journal with articles or a monograph with chapters) and the simple components are also extracted in this step. A standard practice has not emerged yet, and collections have been sometimes considered a single work, a manifestation of different works, or a collective work made of smaller works. The last approach has been used here in order to describe serial relationships, mapping MARC 773 entries to *isPartOf* relationships.

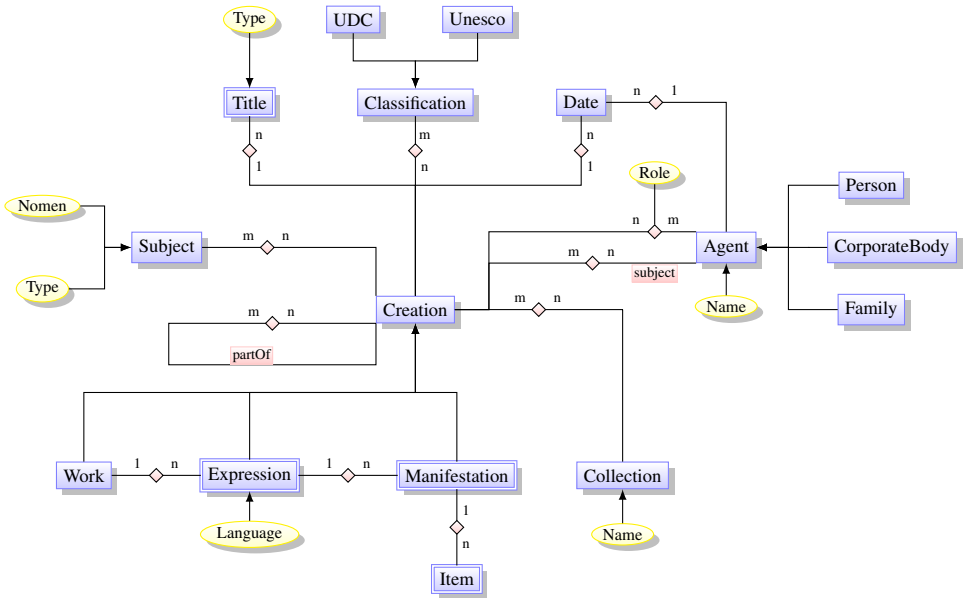

Fig. 3. Diagram of the Entity-Relationship model of the relational database.

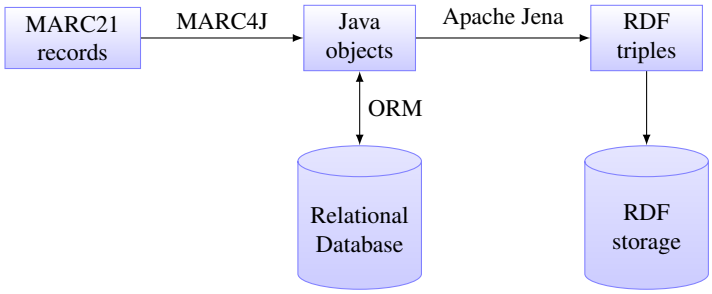

Fig. 4. Schematic representation of the migration and conversion process.

The statement of responsibility (MARC field 245 $c) contains useful information about persons or bodies contributing to the creation of the content, linking usually persons and expressions. Furthermore, reproduction notes (field 533) often relate a document to the source employed to create the digital version, which can be considered expressions of a single work. In order to extract such valuable relationships, these fields were parsed to find keywords such as *edición ilustrada* (an illustrated edition) or *traducción* (a translation) and the results were then mapped to FRBR relationships.

Since patterns are not sufficient to interpret the whole variety of relationships between entities, a web cataloguing interface was implemented for the supervision by librarians of the transformation and clustering process. The interface allows one to retrieve, modify and create relationships and supports the hierarchical navigation through the FRBR structure.

A final step reorganizes the catalogue by grouping manifestations and expressions of the same work, and employs data mining techniques to this purpose. Training sets including difficult cases were prepared by the cataloguing department. Preliminary inspection revealed that uniform titles were not suitable to merge expressions or manifestations of a work, since their main purpose is to provide a normalized form of the title and, only secondarily, to disambiguate works with identical name. The result is that many works sharing their uniform title and author were indeed different creations (for example, many documents had uniform title *Laws* and author *Alfonso XIII, king of Spain*).

The clustering process follows instead the principles of the OCLC FRBR Work-Set Algorithm [31] which identifies sets of works based on the information found in bibliographic and authority records: a key is created for every record by combining author and title and,

secondarily, by using the uniform title (MARC 130) or the title with MARC 7XX fields. Sets contain works which share an identical key.

3.4. From FRBR to RDA Linked Open Data

Two main approaches have been generally applied to the publication of *linked open data*. Transient RDF views are published as a top layer providing real-time access to the original data. Alternatively, persistent RDF views are generated and the data are published in asynchronous time intervals. Since bibliographic archives do not update their data very frequently and some delay is acceptable in delivering the metadata, persistent RDF views provide a more efficient approach in libraries [18]. Moreover, the adoption of RDF systems usually requires a gradual transition to allow heterogeneous data to be carefully adapted and tested while in parallel the personnel gains confidence with the new procedures to create descriptive metadata.

A parser has been implemented in Java that applies mapping rules between the FRBR database and the RDA vocabulary (classes, properties and relationships), based on the RDA recommendations[9]. For every entity in one of the RDA classes vocabulary[10] (e.g., rdac:Work or rdac:Person) an RDF document is created which contains its properties and relationships, as depicted in figure 6. For example:

a) rdaw:titleOfTheWork links a work to the string by which the work is known.
b) rdae:languageOfExpression contains the language used in a particular expression.
c) rdam:carrierType assigns a manifestation to the format used for storage and the type of device required to access the content.

RDA provides also additional value vocabularies[11] for some properties. Parsers for some FRBR attributes such as Media Type, Carrier Type and Content Type have been implemented accordingly.

Whenever a relationship could not be described using RDA elements, then popular vocabularies were applied. For example, the OWL-Time ontology[12] has been used to describe temporal events such as publication years; external content, hosted by partner libraries, was described with FOAF elements [4] and subjects triples were created with the Dublin Core[13] property *dc:subject*. Languages and forms of work (genres), which are currently not specified in RDA vocabularies, have been mapped to the codes used by the Library of Congress.[14] Even if all the standard vocabularies listed in Table 2 were used, some issues could not be fully addressed: for example, RDA provides only a generic relationship for containment (*wholePartManifestationRelationship*) which is not rich enough to describe the variety of inclusions in a collection (volumes in a journal, articles in a volume, or books in a series).

Table 1
Design patterns followed by Uniform Resource Identifiers- Dots stand for the common prefix `data.cervantesvirtual.com` and the asterisk for a particular value.

ENTITY	PATTERN
Person	`.../person/*`
Family	`.../family/*`
Corporate Body	`.../corporatebody/*`
Work	`.../work/*`
Expression	`.../expression/*`
Manifestation	`.../manifestation/*`
Item	`.../item/*`
Institution	`.../institution/*`
Language	`.../language/*`
Date	`.../date/*`

The output dataset adheres to established design patterns [10]. For example, the path to the resource provides a readable description of the entity, as shown in Table 1.

Finally, the dataset has been enriched semantically by automatically linking objects to terms in other Linked Open Datasets. For example, links to DBpedia[15] were gathered for persons by using the identifiers provided by the Virtual International Authority File[16]. This enhancement allows queries where the unambiguous VIAF identifier is used to retrieve information about an author.

[9] `http://www.rda-jsc.org/archivedsite/docs/5rda-frbrrdamappingrev.pdf`
[10] http://www.rdaregistry.info/Elements/c
[11] `http://www.rdaregistry.info/termList`
[12] `www.w3.org/TR/owl-time`
[13] `http://dublincore.org/documents/dces`
[14] `http://id.loc.gov/authorities/genreForms`; `http://id.loc.gov/vocabulary/iso639-1/es`
[15] `http://wiki.dbpedia.org/`
[16] `https://viaf.org/`

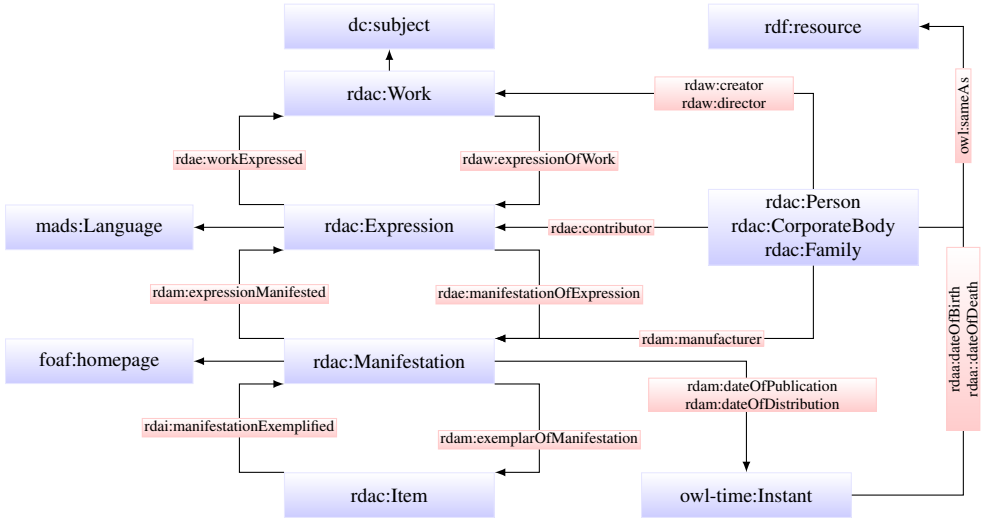

Fig. 5. The ontology (concepts and relations) describing the catalogue entries is based on the RDA, RDF, OWL, FOAF and Dublin Core vocabularies. Tag prefixes denote different name-spaces (the source ontology): RDA Class (rdac), Work (rdaw), Expression (rdae), Manifestation (rdam), Item (rdai), or Agent (rdaa); Resource Description Framework (rdf); Dublin Core (dc); Library of Congress Metadata Authority Description Schema (mads); Friend of a Friend (foaf); OWL time ontology (owl-time).

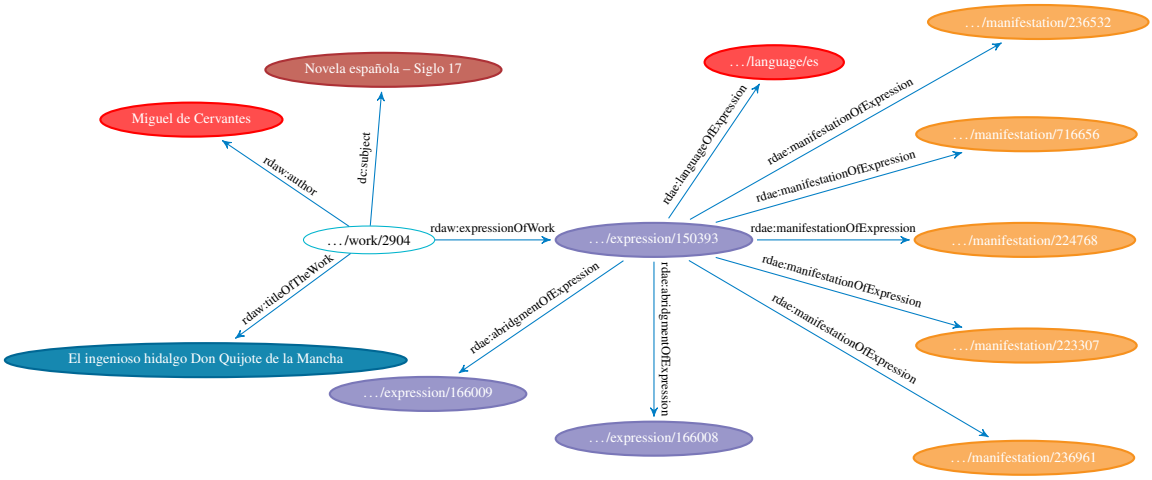

Fig. 6. Overview of the RDF output for the work *El ingenioso hidalgo Don Quijote de la Mancha*.

4. Results and evaluation

The automatic procedure described in section 3 has been applied to transform over 200,000 bibliographic records and 70,000 authority entries, generating about 15 million RDF triples which are published through the gateway *data.cervantesvirtual.com*. The main features of the RDF data set are summarized in table 3 and it provides high quality *linked open data*, what is called *five-star open data* [16]. The repository holds nearly 37,000 links to external repositories, as shown in the table. These links are described through the `owl:sameAs` relationship and they introduce the rich connectivity promoted by the *linked open data* philosophy. The Biblioteca Virtual Miguel de Cervantes data can be downloaded, navigated and queried using a SPARQL endpoint, and they are published under the Creative Commons Public Domain Dedication License[17].

[17] `https://creativecommons.org/publicdomain/zero/1.0`

Table 2
Vocabularies employed in the RDF dataset.

Prefix	Name	URI
rdac	The RDA Classes element set	`http://rdaregistry.info/Elements/c/`
rdaw	The RDA Work properties	`http://rdaregistry.info/Elements/w/`
rdae	The RDA Expression properties	`http://rdaregistry.info/Elements/e/`
rdam	The RDA Manifestation properties	`http://rdaregistry.info/Elements/m/`
rdai	The RDA Item properties	`http://rdaregistry.info/Elements/i/`
rdaa	The RDA Agent properties	`http://rdaregistry.info/Elements/a/`
rdau	The RDA Unconstrained properties	`http://rdaregistry.info/Elements/u/`
foaf	Friend of a Friend vocabulary	`http://xmlns.com/foaf/0.1/`
dc	DCMI Metadata Terms	`http://purl.org/dc/elements/1.1/`
skos	Simple Knowledge Organization System	`http://www.w3.org/2004/02/skos/core#`
dbpedia-owl	DBpedia ontology	`http://dbpedia.org/ontology/`
time	Time Ontology in OWL	`http://www.w3.org/2006/time#`

Table 3
Some features of the RDF dataset.

Main address	`data.cervantesvirtual.com`
Description	`.../void.ttl`
Site-map	`.../sitemap.xml`
Vocabularies	18
No. of classes	20
No. of properties	128
No. of triples	13,131,270
SPARQL access	`.../sparql`
No. of links	7,610 to `viaf.org`
	5,615 to `datos.bne.es`
	45 to `id.loc.gov`
	22,373 to `dbpedia.org`
	1,180 to `youtube.com`

The RDF dataset has been evaluated using several methods:

a) Nearly 40 constraints were defined[18] and the data were validated against them using Clark&Parsia's Stardog ICV[19]. The constraints required, for example, that at least one manifestation must be found for every work and author roles can be only assigned to entities of type work.
b) RDFUnit [19] has been used to test DublinCore triples.
c) Acceptance sampling and manual revision was performed on several hundreds of records.
d) A procedure was implemented testing that the number of manifestations and creators matches the numbers in the original database.

These validation procedures allowed to identify and correct inaccuracies in the dataset. For example, the analysis of a random sample with 112 groups created by the clustering of FRBR entities found 8 false positives where works were grouped incorrectly. The wrong clusters mainly contained works with rather general or vague titles such as *Real Decreto* produced by the same author. The inspection of a random subset of 50 DBpedia links revealed 4 mistakes. In contrast, all roles which were manually revised (50 cases) and all relations to BNE records (50 links) were found to be correct.

Several options to provide SPARQL access to the RDF storage were evaluated, including OpenLink Virtuoso[20], 4Store[21], and Sesame[22]. The last one was selected in order to implement the access to the data, since it is an open-source Java framework which proved to be light-weight and satisfied the requirements by supporting full-text queries, batch indexing, and database transactions.[23] An open-source SPARQL interface[24] was added in order to simplify the creation of queries and the visualization of results.

[18] They are available at `https://github.com/hibernator11/validation`
[19] http://docs.stardog.com/icv
[20] `http://virtuoso.openlinksw.com`
[21] `http://4store.org`
[22] `http://rdf4j.org`
[23] For an extensive comparative study of platforms, see [14].
[24] Yasgui, `http://yasgui.org`

The maintenance of the RDF data generated through the process described above is supported by three automatic procedures for the management of the content:

- Rebuild all RDF triples from the database.
- Incremental addition of new RDF triples.
- Data backup and restore operations.

Fully rebuilding the dataset may require a few hours but the incremental construction runs in real time and can be scheduled to take place periodically so that the published data are synchronized with the database content.

5. Conclusions and future work

The traditional online access of the Biblioteca Virtual Miguel de Cervantes[25] provides only a human readable presentation of the catalogue. The publication of the catalogue as *linked open data* supports instead external usage and exploitation of the data. For example, the free and open knowledge base Wikidata[26] has recently incorporated a new property[27] which identifies authors in the Biblioteca Virtual Miguel de Cervantes. Currently Wikidata contains about 4,500 links to our dataset. The links to DBpedia allow users to perform SPARQL federated queries to retrieve, for example, authors in our library classified by subject or or authors who were influenced by another one.

A tool[28] has been developed to assist browsing *linked open data* by non-expert users. The interface presents search results grouped according to FRBR categories. For example, expressions and manifestations are presented under the work they materialize and contributions related to a particular creator are classified according to the role played in the creation.

Some work is still to be done. For example, subject headings can be expressed in different languages, depending on the source library. This question has been addressed by a number of projects in the past [11] and a global solution needs still to be found. Further refinements are also needed for the recognition and extraction of implicit relationships expressed in natural language, such as named entities and temporal expressions. The description of subjects can be also enriched with the creation of a thesaurus based on SKOS, a W3C recommendation for the representation of subject headings. Additionally, limitations to the clustering arise from the fact that records imported from external repositories sometimes lack sufficient metadata or may be expressed in foreign languages. Finally, even if the SPARQL interface provides auto-completion for properties and relationships, further work is also needed to provide easier access to SPARQL for non-expert users.

[25] http://www.cervantesvirtual.com
[26] http://www.wikidata.org
[27] https://www.wikidata.org/wiki/Property:P2799
[28] http://data.cervantesvirtual.com

References

[1] Trond Aalberg and Maja Zumer. Looking for entities in bibliographic records. In George Buchanan, Masood Masoodian, and Sally Jo Cunningham, editors, *Digital Libraries: Universal and Ubiquitous Access to Information, 11th International Conference on Asian Digital Libraries, ICADL 2008, Bali, Indonesia, December 2-5, 2008. Proceedings*, volume 5362 of *Lecture Notes in Computer Science*, pages 327–330. Springer, 2008. DOI https://doi.org/10.1007/978-3-540-89533-6_36.

[2] Trond Aalberg, Tanja Mercun, and Maja Zumer. Coding FRBR-structured bibliographic information in MARC. In Chunxiao Xing, Fabio Crestani, and Andreas Rauber, editors, *Digital Libraries: For Cultural Heritage, Knowledge Dissemination, and Future Creation - 13th International Conference on Asia-Pacific Digital Libraries, ICADL 2011, Beijing, China, October 24-27, 2011. Proceedings*, volume 7008 of *Lecture Notes in Computer Science*, pages 128–137. Springer, 2011. DOI https://doi.org/10.1007/978-3-642-24826-9_18.

[3] American Library Ascociation, CLA, CILIP. *Anglo-American Cataloguing Rules, Second Edition*. ALA Editions, 2005. ISBN 978-0-8389-3555-2.

[4] Dan Brickley and Libby Miller. *FOAF Vocabulary Specification 0.99*. Namespace Document, 14 January 2014 - Paddington Edition. URL http://xmlns.com/foaf/spec/.

[5] Rafael C. Carrasco, Aureo Serrano, and Reydi Castillo-Buergo. A parser for authority control of author names in bibliographic records. *Information Processing and Management*, 52(5):753–764, 2016. DOI https://doi.org/10.1016/j.ipm.2016.02.002.

[6] Naicheng Chang, Yuchin Tsai, Gordon Dunsire, and Alan Hopkinson. Experimenting with implementing FRBR in a Chinese Koha system. *Library Hi Tech News*, 30(10):10–20, 2013. DOI https://doi.org/10.1108/LHTN-09-2013-0054.

[7] Joffrey Decourselle, Fabien Duchateau, and Nicolas Lumineau. A survey of FRBRization techniques. In Sarantos Kapidakis, Cezary Mazurek, and Marcin Werla, editors, *Research and Advanced Technology for Digital Libraries - 19th International Conference on Theory and Practice of Digital Libraries, TPDL 2015,*

Poznań, Poland, September 14-18, 2015. Proceedings, volume 9316 of *Lecture Notes in Computer Science*, pages 185–196. Springer, 2015. DOI https://doi.org/10.1007/978-3-319-24592-8_14.

[8] Corine Deliot. Publishing the British National Bibliography as Linked Open Data. Online: http://www.bl.uk/bibliographic/pdfs/publishing_bnb_as_lod.pdf, 2014.

[9] Timothy J. Dickey. FRBRization of a library catalog: Better collocation of records, leading to enhanced search, retrieval, and display. *Information Technology and Libraries*, 27(1):23–32, 2008. DOI https://doi.org/10.6017/ital.v27i1.3260.

[10] Leigh Dodds and Ian Davis. *Linked Data Patterns: A pattern catalogue for modelling, publishing, and consuming Linked Data*. dataincubator.org, 2012. URL http://patterns.dataincubator.org/book/.

[11] Magda El-Sherbini. Multilingual subject retrieval: Bibliotheca Alexandrina's subject authority file and linked subject data. In Berthold Lausen, Sabine Krolak-Schwerdt, and Matthias Böhmer, editors, *Data Science, Learning by Latent Structures, and Knowledge Discovery [revised versions of selected papers presented during the European Conference on Data Analysis (ECDA 2013), Luxembourg, July 2013].*, Studies in Classification, Data Analysis, and Knowledge Organization, pages 535–546. Springer, 2013. DOI https://doi.org/10.1007/978-3-662-44983-7_47.

[12] Nuno Freire, Rosa Galvão, and Margarida Lopes. FRBR information discovery in traditional catalogues: the TELplus experience. *World Library and Information Congress: 75th IFLA General Conference and Council*, 2009. URL http://conference.ifla.org/past-wlic/2009/135-freire-en.pdf.

[13] Janifer Gatenby, Gail Thornburg, and Jay Weitz. Collected work clustering in WorldCat. *Code4Lib Journal*, 30, 2015. URL http://journal.code4lib.org/articles/10963.

[14] Bernhard Haslhofer, Elaheh Momeni Roochi, Bernhard Schandl, and Stefan Zander. Europeana RDF store report. Technical report, University of Vienna, Vienna, March 2011. URL http://eprints.cs.univie.ac.at/2833/.

[15] IFLA Study Group on the FRBR. *Functional Requirements for Bibliographic Records*, volume 19 of *IFLA Series on Bibliographic Control*. München: K.G. Saur Verlag, 1998.

[16] Krzysztof Janowicz, Pascal Hitzler, Benjamin Adams, Dave Kolas, and Charles Vardeman. Five stars of linked data vocabulary use. *Semantic Web*, 5(3):173–176, 2014. DOI https://doi.org/10.3233/SW-140135.

[17] Joint Steering Committee (JSC) for Development of RDA. RDA Tool Kit: Resource Description and Access. Online: http://www.rdatoolkit.org, 2012.

[18] Nikolaos Konstantinou, Dimitrios-Emmanuel Spanos, and Nikolas Mitrou. Transient and persistent RDF views over relational databases in the context of digital repositories. In Emmanouel Garoufallou and Jane Greenberg, editors, *Metadata and Semantics Research - 7th Research Conference, MTSR 2013, Thessaloniki, Greece, November 19-22, 2013. Proceedings*, volume 390 of *Communications in Computer and Information Science*, pages 342–354. Springer, 2013. DOI https://doi.org/10.1007/978-3-319-03437-9_33.

[19] Dimitris Kontokostas, Patrick Westphal, Sören Auer, Sebastian Hellmann, Jens Lehmann, and Roland Cornelissen. Databugger: a test-driven framework for debugging the web of data. In Chin-Wan Chung, Andrei Z. Broder, Kyuseok Shim, and Torsten Suel, editors, *23rd International World Wide Web Conference, WWW '14, Seoul, Republic of Korea, April 7-11, 2014, Companion Volume*, pages 115–118. ACM, 2014. DOI https://doi.org/10.1145/2567948.2577017.

[20] Marcia Lei Zeng, Maja Žumer, and Athena Salaba, editors. *Functional Requirements for Subject Authority Data (FRSAD) - A Conceptual Model*, volume 43 of *IFLA series on bibliographic control*. De Gruyter Saur, 2011. ISBN 978-3-11-025323-8.

[21] Hugo Miguel Álvaro Manguinhas, Nuno Miguel Antunes Freire, and José Luis Brinquete Borbinha. FRBRization of MARC records in multiple catalogs. In Jane Hunter, Carl Lagoze, C. Lee Giles, and Yuan-Fang Li, editors, *Proceedings of the 2010 Joint International Conference on Digital Libraries, JCDL 2010, Gold Coast, Queensland, Australia, June 21-25, 2010*, pages 225–234. ACM, 2010. DOI https://doi.org/10.1145/1816123.1816157.

[22] Julia Marden, Carolyn Li-Madeo, Noreen Whysel, and Jeffrey Edelstein. Linked open data for cultural heritage: evolution of an information technology. In Michael J. Albers and Katherine Gossett, editors, *Proceedings of the 31st ACM international conference on Design of communication, Greenville, NC, USA, September 30 - October 1, 2013*, pages 107–112. ACM, 2013. DOI https://doi.org/10.1145/2507065.2507103.

[23] Ricardo Santos Muñoz. Launching of beta version of datos.bne.es, a LOD service and a FRBR-based catalogue view. *SCATNews*, 42(42):13–21, December 2014. URL http://www.ifla.org/files/assets/cataloguing/scatn/scat-news-42.pdf.

[24] Glenn E. Patton, editor. *Functional Requirements for Authority Data - A Conceptual Model*, volume 34 of *IFLA series on bibliographic control*. München: K.G. Saur, 2009. ISBN 978-3-598-24282-3.

[25] Eric Prud'hommeaux and Andy Seaborne, editors. *SPARQL Query Language for RDF*. W3C Recommendation, 15 January 2008. URL https://www.w3.org/TR/rdf-sparql-query/.

[26] Jenn Riley. Enhancing interoperability of frbr-based metadata. In Diane Ileana Hillmann and Michael Lauruhn, editors, *Proceedings of the 2010 International Conference on Dublin Core and Metadata Applica-*

tions, DC 2010, Pittsburgh, Pennsylvania, USA, October 20-22, 2010, pages 31–43. Dublin Core Metadata Initiative, 2010. URL http://dcpapers.dublincore.org/pubs/article/view/1037.

[27] Jodi Schneider. FRBRizing MARC records with the frbr display tool. Technical report, 2008. URL http://jodischneider.com/pubs/2008may_frbr.html.

[28] Agnès Simon, Adrien Di Mascio, Vincent Michel, and Sébastien Peyrard. We grew up together: data.bnf.fr from the BnF and Logilab perspectives. In *IFLA 2014 Satellite Meeting Linked Data in Libraries: Let's make it happen!*, 2014. URL http://ifla2014-satdata.bnf.fr/pdf/iflalld2014_submission_Simon_DiMascio_Michel_Peyrard.pdf.

[29] Standing Committee of the IFLA Cataloguing Section, editor. *International Standard Bibliographic Description (ISBD). Consolidated Edition*. IFLA Series on Bibliographic Control 44. De Gruyter Saur, 2011. URL http://www.ifla.org/files/assets/cataloguing/isbd/isbd-cons_20110321.pdf.

[30] Naimdjon Takhirov, Trond Aalberg, Fabien Duchateau, and Maja Zumer. FRBR-ML: A FRBR-based framework for semantic interoperability. *Semantic Web*, 3(1):23–43, 2012. DOI https://doi.org/10.3233/SW-2012-0044.

[31] Jenny Toves Thomas B. Hickey. FRBR WorkSet Algorithm. Version 2.0, 2009. URL http://www.oclc.org/content/dam/research/activities/frbralgorithm/2009-08.pdf.

[32] Gail Thornburg. A Candid Look at Collected Works: Challenges of Clustering Aggregates in GLIMIR and FRBR. *Information Technology and Libraries*, 33(3), September 2014. ISSN 2163-5226. DOI https://doi.org/10.6017/ital.v33i3.5377.

[33] Daniel Vila-Suero and Asunción Gómez-Pérez. datos.bne.es and MARiMbA: an insight into library linked data. *Library Hi Tech*, 31(4):575–601, 2013. DOI https://doi.org/10.1108/LHT-03-2013-0031.