Library Data Integration: The CoBiS Linked Open Data Project and Portal

Luisa Schiavone[1(✉)], Federico Morando[2], Davide Allavena[2], and Giorgio Bevilacqua[3]

[1] INAF Turin Astrophysical Observatory, Pino Torinese, Italy
luisa.schiavone@inaf.it
[2] Synapta Srl, Turin, Italy
info@synapta.it
[3] Department of Humanities, University of Turin, Turin, Italy
338853@edu.unito.it

Abstract. The CoBiS is a network formed by 65 libraries. The project is a pilot for Piedmont aiming to provide the libraries with an infrastructure for LOD publishing, creating a triplification pipeline designed to be easy to automate and replicate. This was realized with open source technologies, such as the TARQL and JARQL tools that use SPARQL queries to describe the conversion of tables (CSV) or trees (JSON) into graphs (RDF data). The first challenge consisted in making possible the dialog of heterogeneous data sources, coming from four different library applications and different types of data. As a second step, the information contained in the catalogs was interlinked with external data sources.

1 Introduction

The **CoBiS** (Coordinamento delle Biblioteche Speciali e Specialistiche di Torino i.e. *Coordination of Special and Specialized Libraries of Turin*) is an informal network of 65 libraries, collaborating to provide continuing professional development and to offer a better service to their users. CoBiS libraries are heterogeneous from many points of view: holdings, cataloguing software and OPACs. The **LOD project** started in 2015 as a training program, in collaboration with Prof. Vivarelli from the University of Turin. The program was divided in various topics: copyright, collaboration between libraries and Wikipedia, (Linked) Open Data.

2 Purpose

Six libraries from the CoBiS decided to participate in a **pilot project** with the purpose of **providing a unique access point to the collections of CoBiS libraries**. CoBiS bibliographic data were divergent from different perspectives: file formats, semantic frameworks and authority files.

G. Serra and C. Tasso (Eds.): IRCDL 2018, CCIS 806, pp. 15–22, 2018.
https://doi.org/10.1007/978-3-319-73165-0_2

Linked Open Data technologies provide the means to engage such interoperability problems, both from a technical and a semantic point of view. Many national libraries (e.g. BNE [1] and BNF [2]) have converted their catalogues from MARC to RDF and have published their data also through a SPARQL endpoint.

In Italy, the SHARE Catalogue project[1] has converted MARC records from multiple university libraries in RDF; however, no SPARQL endpoint is available to query the database.

By developing and implementing effective tools and procedures, we were able to convert CoBiS datasets, composed of different data formats, into RDF, to interlink them with external authoritative sources (Wikidata, VIAF and InternetArchive) and to publish them as Linked Data, giving also access to the enriched dataset via a SPARQL endpoint.

This work led CoBiS datasets to become a connecting piece in the Linked Open Data Cloud; as a result, data are not only becoming more interoperable among them, but they are also open in order to facilitate the collaboration with online communities.

3 The Project

With respect to the **Linked Open Data stack**, Fig. 1 shows an overall picture of the **project**.

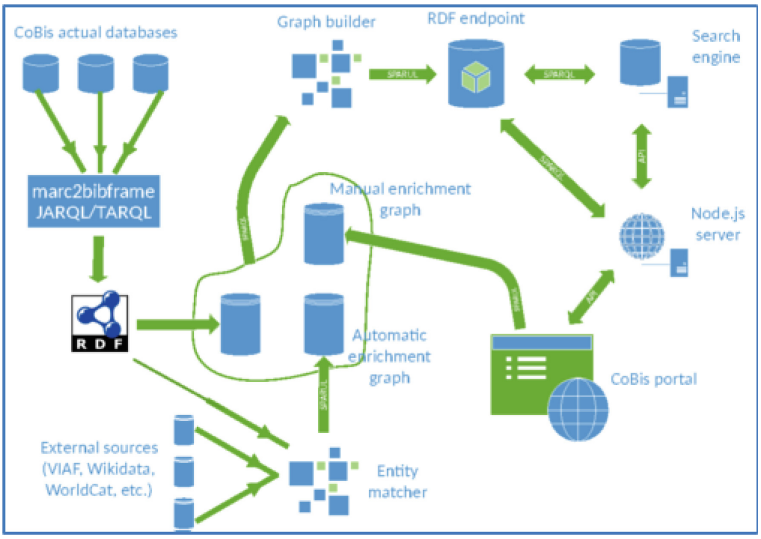

Fig. 1. The project architecture

[1] http://catalogo.share-cat.unina.it/sharecat/clusters.

The first activity of the project consisted in transforming bibliographic data into RDF triples.

Created using various cataloguing softwares, CoBiS data are encoded in different file formats (CSV, XML and JSON) and are structured according to different data models (MARC 21, UNIMARC, Dublin Core OAI-PMH).

In order to define a unique data model we used two main **ontologies**: Schema.org and Bibframe, the latter being developed by the Library of Congress with the specific aim to facilitate the necessary transition from the traditional catalographic system (based on MARC) to a bibliographic *environment* integrated in the *web of data* [3]. We also used selected properties from RDFS, OWL, DCTerms, FOAF, and Culturalis[2], with the purpose of providing more semantic interoperability.

During the first phase of the project, we tried to convert data in RDF triples using the Library of Congress tool marc2bibframe[3] to process MARC 21 and using the RML mapping language for all the other formats.

RML is a *generic mapping language defined to express customized mapping rules from heterogeneous data structures and serializations to the rdf data model* [4]. The RML conceptual model, based on R2RML W3C standard[4], perfectly fitted with the needs of the project, but had some limitations with respect to technical performances, i.e. processing time, and the verbosity of the mapping language.

When tabular data were available, we experimented the TARQL tool, described by its developers as *SPARQL for Tables: a command-line tool for converting CSV files to RDF using SPARQL 1.1 syntax*[5].

TARQL proved to be both technically efficient and not very time-consuming in terms of writing new mappings, thanks to the use of the standard SPARQL 1.1 syntax, with all its features.

Since many of our input data had a tree-like structure (XML or JSON format) Synapta, in a joined effort with other open source developers from FactsMission[6], realized and published JARQL, a new open source software tool for converting JSON data to RDF. As TARQL, from which it takes inspiration, JARQL uses the SPARQL 1.1 syntax and constructs queries to describe its mappings[7].

By improving those technical tools, we were able to define a **triplification pipeline** where data are extracted from local sources, mapped (using a SPARQL query) to selected ontologies, and converted into an RDF graph, which is periodically updated. A sample JSON input and the related SPARQL mapping is shown in Fig. 2.

[2] http://culturalis.org/.
[3] https://github.com/lcnetdev/marc2bibframe.
[4] https://www.w3.org/TR/r2rml/.
[5] http://tarql.github.io/.
[6] https://factsmission.com/.
[7] http://jarql.linked.solutions/.

Fig. 2. JSON input and related JARQL mapping using SPARQL 1.1 syntax

According to the W3C recommendations [5], CoBiS entities (books and authors) are unambiguously identified by URIs, so that they can be connected to external sources.

CoBiS libraries were using different authority files (e.g. SBN or local systems), but the graph structure of RDF supported the generation of a unified Authority file. Interlinks to external sources, like VIAF and Wikidata, helped us to identify and de-duplicate authors under a shared procedure.

For **link generation** [6] we used both automatic algorithms and manual approaches. In order to minimize false positives, automatic matches proceeded mostly through SBN identifiers: in this way, the newly created graph was interlinked with VIAF and Wikidata.

Due to the size of the VIAF database and to the absence of a public SPARQL end-point, we had to recreate locally a Linked Data graph from the VIAF RDF dump.

In order to support manual matches, we exploited **OLAF (Open Linked Authority File)**[8], our crowd-sourcing interface for creating an authority file based on SPARQL queries.

By analyzing different RDF-structured databases, OLAF suggests potential relations of identity between similar entities (see Fig. 3).

Candidate matches are submitted to domain-experts and, if validated, they are annotated in the CoBiS RDF graph, using the owl:sameAs property. The interaction between automatic computation and manual validation improves the quality and the **reliability** [7] of the data, by allowing domain-experts to directly contribute to the Linked Open Data Cloud.

[8] https://olaf.synapta.io/.

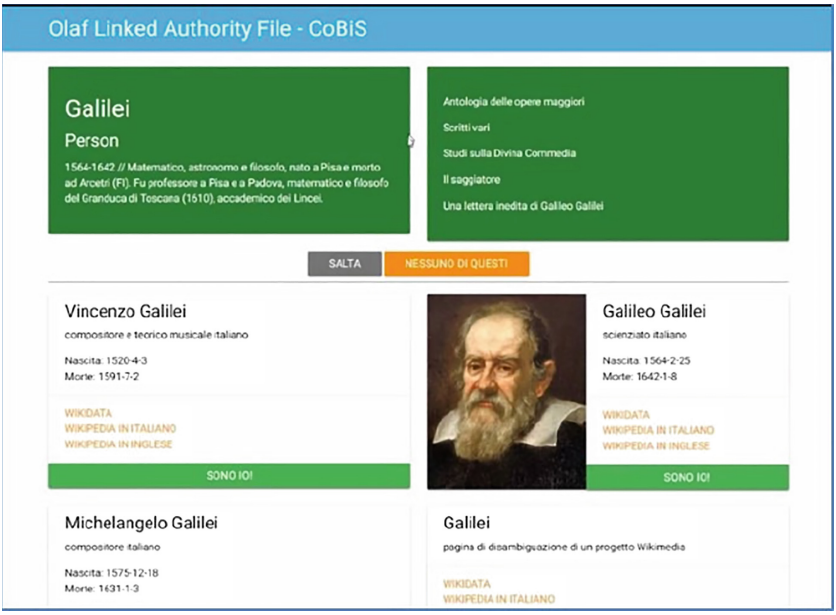

Fig. 3. The *Galilei* search in OLAF

Within the overall interlinking process, **Wikidata** is assuming the role of metadata hub: because of its constant growing and the dynamism of the *wiki* approach [8], its entities are strongly interlinked with VIAF and other databases (e.g. national bibliographic and biographic dictionaries, disciplinary databases, etc.).

On a scenario where the cataloging process is evolving into aggregating shared and interlinked data [9], librarians (especially those of small institutions not contributing to VIAF) are encouraged to use Wikidata as authority file [10], both by connecting existing data and creating new items about authors not yet included in Wikidata (using its publicly editable graphic interface). In those cases, such a practice would minimize redundancy and the overall cost of authority record creation, thus increasing the efficiency of bibliographic record production and maintenance [11].

4 The Portal

The CoBiS LOD Project portal (http://dati.cobis.to.it) is online with its full Linked Data stack, including a public SPARQL end-point configured to support federated queries (http://dati.cobis.to.it/sparql) a full dump of the RDF data, etc.

In the author's page, dynamically generated through SPARQL queries, you have biographical information and a list of interlinked resources coming from Wikidata and other bibliographic repositories (VIAF, Wikidata, LoC, Deutsche National Bibliothek GND, Bibliothèque Nationale de France BNF, Servizio Bibliotecario Nazionale SBN, Dizionario Biografico degli Italiani DBI).

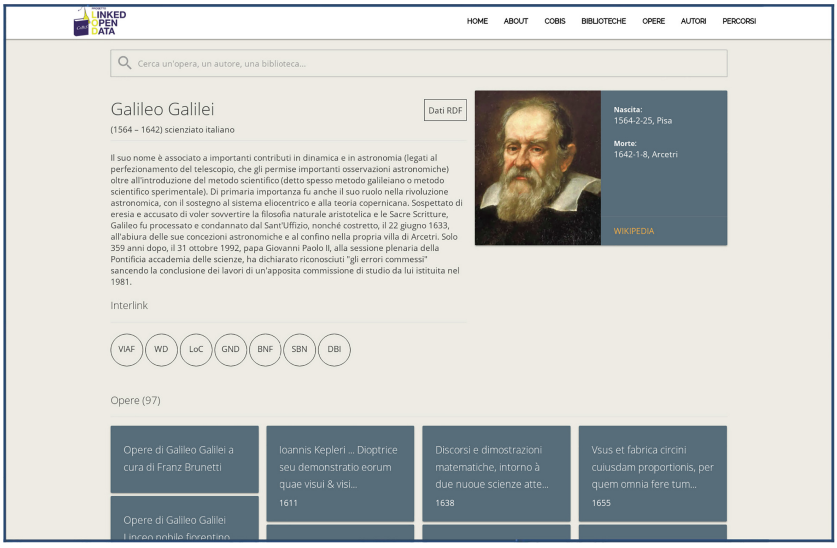

Fig. 4. The *Galileo Galilei* page

On the right there is an infobox with authors data and an image, both fetched live from Wikidata leveraging Linked Data. Clicking on the RDF button, all the triples of the resource can be directly viewed.

At the bottom of the page, all the author's **books inside the CoBiS database** are shown (see Fig. 4). To explore information on such books, you can click one of the boxes or you can use the search bar, looking for a title which is not listed[9].

Figure 5 shows an **example search for the *Dialogo***. On the left side of the page, you see bibliographic details with a collection of interlinked resources. Exploiting the power of Open Data, we are also able to read the **Internet Archive digital copy of the book** (Fig. 6). A physical copy of the book is

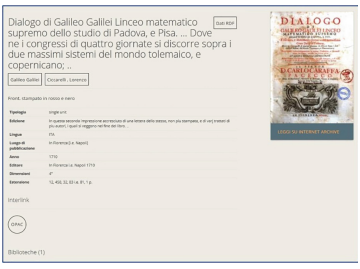

Fig. 5. The Galilei's *Dialogo* file in the CoBiS database

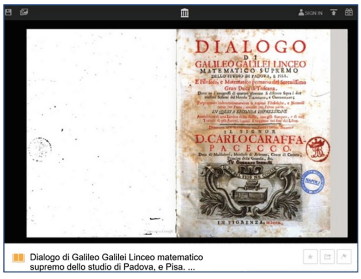

Fig. 6. The Galilei's *Dialogo* in the internet archive

[9] The search bar is powered by Solr - https://lucene.apache.org/solr/.

available in some CoBiS libraries. All details coming from the individual OPACs can be shown by clicking the **OPAC button** (Fig. 5).

5 Towards New Challenges

The pilot is only an initial step.

CoBiS plan to include more libraries in a second phase of the project, converting their data into Linked Open Data and linking them to the Linked Open Data Cloud and to other online resources, such as Internet Archive and Wikipedia.

Thanks to linked data, it is possible to query for common items in the CoBiS catalogue and in other major National Libraries catalogues exposing a SPARQL endpoint.

Performing the query in Fig. 7 on the CoBiS SPARQL endpoint[10] towards the French National Library SPARQL endpoint[11] returns a subset of shared books.

```
PREFIX bnf-onto: <http://data.bnf.fr/ontology/bnf-onto/>
PREFIX schemaorg: <http://schema.org/>

select ?cobisInstance ?BNFInstance
where {
  ?cobisInstace <http://schema.org/isbn> ?isbnCobis .
  service <http://data.bnf.fr/sparql> {
    {
    select ?isbn ?BNFinstance
    where {
      ?BNFInstance <http://data.bnf.fr/ontology/bnf-onto/isbn> ?isbn .
    } LIMIT 10000
    }
  }
  FILTER(REPLACE(STR(?isbn), "-", "" ) = ?isbnCobis)
} LIMIT 100
```

Fig. 7. The performed query

We also aim to improve the interlinking, so as to link CoBiS data to other online open data, with specific regard to Wikidata and VIAF and to the most important international projects, such as the linked open data portals of the French and Spanish National Libraries and of the Library of Congress in Washington DC.

[10] http://dati.cobis.to.it/sparql.
[11] http://data.bnf.fr/sparql.

6 Acknowledgements

This pilot project would not have taken place without the aid of the following organizations, that we want to thank:

- *Regione Piemonte* for having believed in the project and for the contribution provided;
- *Politecnico di Torino, Management Committee of the Fund for Development of Research and Education in Information and Communication Technologies* for the financial support to the initial phase of the project;
- *Politecnico di Torino, Nexa Center for Internet and Society (DAUIN)* for the collaboration;
- *Synapta* for the technical realization;
- all the participating Institutes: National Institute for Astrophysics (INAF), Turin Academy of Sciences, Olivetti Historical Archives Association, Alpine Club National Library, Deputazione Subalpina di Storia Patria, National Institute for Metrological Research (INRIM).

References

1. Vila-Suero, D., Villazón-Terrazas, B., Gómez-Pérez, A.: datos.bne.es: A library linked dataset. Semant. Web **4**(3), 307–313 (2013)
2. Simon, A., Wenz, R., Michel, V., Di Mascio, A.: Publishing bibliographic records on the web of data: opportunities for the BnF (French National Library). In: Cimiano, P., Corcho, O., Presutti, V., Hollink, L., Rudolph, S. (eds.) ESWC 2013. LNCS, vol. 7882, pp. 563–577. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-38288-8_38
3. Miller, E., et al.: Bibliographic framework as a web of data: linked data model and supporting services. Library of Congress, Washington DC (2012)
4. Dimou, A., Sande, M.V., Colpaert, P., Verborgh, R., Mannens, E., Van de Walle, R.: RML: A generic language for integrated RDF mappings of heterogeneous data. In: Bizer, C., Heath, T., Auer, S., Berners-Lee, T. (eds.) Proceedings of the 7th Workshop on Linked Data on the Web (2014)
5. LinkedData. https://www.w3.org/DesignIssues/LinkedData.html
6. Bizer, C., Heath, T., Berners-Lee, T.: Linked data - the story so far. IJSWIS **5**(3), 1–22 (2009)
7. Guerrini, M., Possemato, T.: Linked data per biblioteche, archivi e musei. Editrice Bibliografica, Milano (2015)
8. Martinelli, L.: Wikidata: la soluzione wikimediana ai linked open data. AIB Studi **56**(1), 75–85 (2016)
9. Bianchini, C.: RDA e la sfida del web semantico. In: De Castro, F. (ed.) Il punto sul Servizio Bibliotecario Nazionale e le sue realizzazioni nel Friuli Venezia Giulia, pp. 197–206. EUT Edizioni Università di Trieste, Trieste (2014)
10. Guerrini, M.: La filosofia open: paradigma del servizio contemporaneo. Biblioteche oggi **35**, 12–21 (2017)
11. Library of Congress Working Group on the Future of Bibliographic Control: On the Record: Report of the Library of Congress Working Group on the Future of Bibliographic Control. Library of Congress, Washington DC (2008)