The first Italian Union Catalogue in Linked Open Data: Share Catalogue as cooperative model in the cultural heritage

Tiziana Possemato (@Cult)
The catalogue presented is part of a broader project of cooperation and service sharing among libraries in Campania, Basilicata and Salento called SHARE (Scholarly Heritage and Access to Research), which includes, with different levels of collaboration:

- the adoption of an organisational model based on cooperation;

- integrated access to bibliographic information and documents relating to their respective analogue and digital collections;

- the creation of common platforms for digital publishing:
  - SHARE Press
  - SHARE Journals
  - SHARE Books
  - SHARE Open Archive
The **SHARE Catalogue** project involved conversion to linked data, publication on different sites and open data portals and the construction of a single portal for the use of the data in the authority and bibliographic catalogues of the participating Universities:

- Università degli Studi di Napoli Federico II (Naples)
- Università degli Studi di Napoli L’Orientale (Naples)
- Università degli Studi Napoli Parthenope (Naples)
- Università degli Studi di Salerno (Salerno)
- Università degli Studi del Sannio (Benevento)
- Università degli Studi della Basilicata (Potenza)
- Università degli Studi del Salento (Lecce)
Project aim: to integrate the considerable *knowledge base* represented by the universities’ different authority and bibliographic catalogues to enrich it with the new and in-flux one generated by the web, creating an integrated information system to provide users with a single access tool for the various Libraries’ OPAC.
**Project objectives** – divided into different but complementary lines:

- the conversion of data according to the reference model indicated by the W3C standards for Linked Open Data, RDF – Resource Description Framework;

- the creation of a data structure based on separate records for Persons and Works, defined by the BIBFRAME model;

- the enrichment of data through a connection to external projects, above all related to Authority files and online encyclopedias (e.g. VIAF, Library of Congress Name Authority file, ISNI, Wikipedia, Wikidata, ...).
SHARE Catalogue

• the publication of datasets on the sites responsible for receiving data in linked open data with Creative Commons license;

• the construction of a consultation portal, equipped with simple and intuitive navigation based on the new BIBFRAME – Bibliographic Framework Initiative – data model.
The Bibliographic Framework as a Web of Data: Linked Data Model and Supporting Services document published by the Library of Congress on November 21, 2012, sets out a new data model designed as an evolution, in linked open data, of the Marc 21 format.

The reflections on the new cataloguing rules focus on some specific points, including:

- a greater level of identification and analysis of the data;
- greater attention to controlled vocabularies;
- more widespread use of terms instead of codes;
- emphasis on relationships;
- greater flexibility in controlled items.
“In translating the MARC 21 format to a Linked Data model it is important to deconstruct and then reconstruct the informational assets that comprise MARC”. The BIBFRAME Model, version 2.0 (published on 2016, 21th of April) consists of the following core classes:

**Work**: The highest level of abstraction, a Work, in the BIBFRAME context, reflects the conceptual essence of the cataloged resource: authors, languages, and what it is about (subjects).

**Instance**: A Work may have one or more individual, material embodiments, for example, a particular published form. These are Instances of the Work. An Instance reflects information such as its publisher, place and date of publication, and format.

**Item**: An item is an actual copy (physical or electronic) of an Instance. It reflects information such as its location (physical or virtual), shelf mark, and barcode.
BIBFRAME 2.0 further defines additional key concepts that have relationships to the core classes:

**Agents**: Agents are people, organizations, jurisdictions, etc., associated with a Work or Instance through roles such as author, editor, artist, photographer, composer, illustrator, etc.

**Subjects**: A Work might be “about” one or more concepts. Such a concept is said to be a “subject” of the Work. Concepts that may be subjects include topics, places, temporal expressions, events, works, instances, items, agents, etc.

**Events**: Occurrences, the recording of which may be the content of a Work
BIBFRAME – Il vocabolario

The **BIBFRAME Vocabulary** is comprised of the RDF **properties**, **classes**, and **relationships** between and among them:

**Classes** include the three core classes (Work, Instance and Item) as well as various additional classes, many of which are subclasses of the core classes.

**Properties** describe characteristics of the resource being described as well as **relationships** among resources. For example: one Work might be a “translation of” another Work; an Instance may be an “instance of” a particular BIBFRAME Work. Other properties describe attributes of Works and Instances. For example: the BIBFRAME property “subject” expresses an important attribute of a Work (what the Work is about), and the property “extent” (e.g. number) expresses an attribute of an Instance.
The SHARE Catalogue project:

The decision to direct a project of data integration from different catalogues to the Semantic Web and linked open data technologies was based on extensive literature that defines its potential and richness.

The goal, therefore, was not only to facilitate interaction, exchange and the reuse of information between dissimilar computer systems (in terms of sw and hw), but, above all, to offer very interesting benefits to both the knowledge base manager (the librarian) and the end user.
General description of the catalogues

The cataloguing data considered in the project is derived from three different LMS (Aleph, Sebina and Millennium) and is structured in Unimarc and Marc 21.

The type of information considered by the project:
• the bibliographic database (bibliographic record);
• the authority database (authority files).

In the analysis of the original data, particular attention was paid to the elements and common cataloguing practices in the different catalogues to identify contact points between the various sources and the elements of diversity and difference.
Survey results:

- little or no widespread use of attributes useful for the identification of the Work and Expression entities;
- few records accompanied by an authorised access point for the title (what REICAT defines as the *uniform title*) or by elements such as the original title to identify an Expression.
General description of the catalogues

This first observation of the data suggested another analysis path:

- a study of the *Person-Family-Body* entities with their relative attributes: an element analysis related to the creator or creators of a work, analysing, mainly, two authority files from the Università Federico II and Università della Basilicata;

- a comparison of the same *Person*, entity recorded in the different catalogues in order to analyse the cataloguing practices of each library and the uniformity (or difference) in the choice of access point.

(* From this point onwards, by *Person*, we refer to an entity of the type Person/Family/Body)
Enrichment of FRBR/BIBFRAME entities

The objective of the subsequent work phase was to provide, through the use of API, web services and SPARQL endpoints available from external sources to the catalogue, data enrichment through the creation as a *cluster* of forms used in different catalogues and coming from the outside, for each entity, and therefore:

- **cluster of names** – the forms, authorised and variants, of names of the People;
- **cluster of titles** – authorised access point and variant forms for the titles of the Works.
Creation of *Person* cluster

**Enrichment of the source data through VIAF**: for each Person entity, all the forms through which it can be represented are recovered from VIAF. The enrichment of the original data takes place in different steps:

- extraction, from the library authority files, the authorised form and variant forms of the name;
- extraction, from catalogue records, the fields in which the item’s creator/creators is registered;
- direct comparison of access points thus obtained with the forms on the VIAF, structured into clusters;
- Creation of the final cluster with all the forms of the related name and associated with a unique local identifier.
Person cluster

Results of a Person cluster for the Richard von Mises entity with the different name forms stemming from:

- Authority files
- VIAF
- Referrals (to authority record)
- Forms used in bibliographic records
Data enrichment

This enrichment from external sources to catalogues also allows, for example, a Russian-speaking user to search for the string

Толстой, Лев Николаевич *

instead of

Tolstoj, Lev Nikolaevic

without any of the participants in the library project having ever created a reference in Cyrillic for this name.
Data enrichment

This person in

[Links to Library of Congress and VVKDATA]

Works

- Tolstoj, Lev Nikolaevic 1828-1910
- Tolstoj, Lev Nikolaevic
- Tolstoj, L. N. 1828-1910 Лев Николаевич
- Tolstoj, Lev
- Tolstoy, Leo (Russian writer, philosopher, 1828-1910)
- Лев Николаевич Толстой русский писатель и мыслитель
- Tolstoj, Lev, 1828-1910
The creation of title clusters for the identification of a Work was rather complex, due to the lack of authority records for titles and the shortage, in the bibliographic records, as mentioned, of authorised access point for this entity type.

The result is a great help for end users: dozens of different titles, associated with the publications in the various catalogues, are grouped into a single authorised access point, to simplify the user’s navigation experience and the identification of the appropriate resource.
Grouping under a single work title of the many publication titles in the catalogue for *Promessi sposi*.

One work title brings together more than 70 different publications catalogued by the different libraries, which the end user can access with just one search.

Example for Anna Karenina *
From Work to Instances (publications)

Example for Promessi sposi *

Clicking result on the title Promessi sposi
The **SHARE Catalogue** project consultation portal is characterised by architecture made up of three different levels:

- **Person - Works**
- **Instances (publications)**
- **Item**
The collective information on the **Person**, his **Works**, scope of study and research (biographical note).

At this informational level, enriched by data from sources outside the library catalogues for the purpose of extending the research potential, only the titles of works for which there is, in the single catalogue, at least one instance or publication available are shown.

This selection, which, as you know, operates in a catalogue and not in a bibliography, ensures that the user’s search experience does not become frustrated by the identification of titles of works that are not actually present in the library collections.
the **Instances level**, associated with the **publications**, is created using the bibliographic data of the different library catalogues, connected to the overlying layer, through relationships with the Works present.

Instances/Publications level is achieved through indexing in the SOLR search engine, which allows you to create additional groups through the **facets** function (grouping by date of publication, by language, by author, by publishers, ...) and gives end users a rich set of search and navigation functions.
Each Instance/Publication thus identified is linked to information about the data and the availability of the copy present in the local OPAC of each library.
The simplified navigation

Navigation example starting from the identification of an author from the rather large literary output...

Francesco Guicciardini

... and the selection of one of his major historiographical works, *Storia d’Italia* regardless of the many cataloguing forms that the different publications of the work have adopted in the various catalogues
Transition to the single Catalogue that brings together, in this case, the 18 work records for *Storia d’Italia* by Francesco Guicciardini present in the different libraries.

The navigational path to these results was guided and simplified in comparison to the native catalogues. The search experience should therefore be positive.
Esempio della pagina in RDF della risorsa «Mauro Guerrini»:

http://test-share-cat.oseegenius.it/univsud/searchNames?n_cluster_id=5520

<http://rdf.atcult.it/Person/5520> <http://bibframe.org/vocab/label> "Guerrini, Mauro" .
<http://rdf.atcult.it/Person/5520> <http://bibframe.org/vocab/label> "GUERRINI, Mauro" .
Conclusions

The **SHARE Catalogue** project is one of the main results of a project of cooperation and sharing of experiences and resources between Universities in Campania, Basilicata and Salento.

This is an initiative that aims to facilitate the resource use experience in libraries spread over a vast geographical territory: rendering their catalogues, with their wealth of resources and their specificity, navigable and usable in a few steps according to data organisation models (such as FRBR and BIBFRAME) that arise from an observation of the user’s information and search needs.

In a simple and intuitive mode, it allows researches to find their way in a vast world of information to easily meet their needs.
The project also aims to create a *work group*, formed by cataloguers and other experts willing to share their experience and expertise to improve the information on offer to users.

Sharing catalogues, the analyses carried out during data processing and the subsequent stages in which the project has gradually entered production, the comparison between personal working methods and those of the other group members has increased the awareness of the specificities of their training and the contribution they can make to meet the needs of users and fulfill the mission libraries must have.
Conclusions

Through the **SHARE Catalogue**, the linked open data catalogue, research systems and different publishing platforms created within the project of cooperation and sharing of **SHARE** services, the affiliated universities are promoting a *culture of openness towards knowledge and science*, offering users and, primarily, scholars a consistently coherent set of tools to compile, publish and make freely accessible on the web, according to international standards, their entire research cycle from collection to data classification, from relational structuring to textual presentation, and finally the reworking in the discursive order of an article or a book.
Demo: http://catalogo.share-cat.unina.it/sharecat/clusters

Esempio: http://catalogo.share-cat.unina.it/sharecat/searchNames?n_cluster_id=140063

Tiziana Possemato – @CULT

Thank you