JLIS.it 9, 1 (January 2018)
ISSN: 2038-1026 online

DOI: 10.4403/jlis.it-12447

JLIS.it

How RDA is essential in the reconciliation and conversion processes for quality Linked Data

Tiziana Possemato [a]

a)AtCult. http://orcid.org/0000-0002-7184-4070

Contact: Tiziana Possemato tiziana.possemato@atcult.it
Received: 10 October 2017; **Accepted:**31 October 2017; **First Published:** 15 January 2018

ABSTRACT

RDA (Resource Description and Access), was initially released in 2010 and, as it is particularly appropriate for use by libraries, archives and museums, it replaces the Anglo-American Cataloguing Rules, Second Edition (AACR2). It provides a new structure for the organization of bibliographic data based on the Functional Requirements for Bibliographic Records (FRBR), with more emphasis on identifiers and relationships than on descriptions. In November 2016 the RDA Steering Committee announced steps toward progressive adoption of the IFLA Library Reference Model (LRM). RDA supports the Linked Data environment also through the representation of RDA entities, elements, relationship designators and vocabulary encoding schemas in Resource Description Framework (RDF, the syntax of the semantic web) in the RDA Registry. The paper is concerned with the application of the RDA standard within the field of Linked Data and how it may be used to improve the quality of the data produced to reach the advantages that the semantic web can bring to the cultural heritage sector. More specifically it will look at a series of Share Linked Open Data (SHARE-LOD) projects.

KEYWORDS

RDA; Linked Data; Semantic web; Cultural heritage; Share Linked Open Data (Share-LOD).

CITATION

Possemato, T. "How RDA is essential in the reconciliation and conversion processes for quality Linked Data". *JLIS.it* 9, 1 (January 2018): 48-60. doi: 10.4403/jlis.it-12447.

JLIS.it 9, 1 (January 2018)
ISSN: 2038-1026 online

DOI: 10.4403/jlis.it-12447

Introduction

After opening seminar devoted on *RDA in Europe*, the second day of the EURIG Annual Meeting is dedicated on *RDA towards Linked Data*. The aim is to address the fundamental advantages of RDA as a way of implementing Linked Data models and its technologies.

How RDA is essential in the reconciliation and conversion processes for quality linked data

RDA (Resource Description and Access), was initially released in 2010 and, as it is particularly appropriate for use by libraries, archives and museums, it replaces the *Anglo-American Cataloguing Rules, Second Edition* (AACR2). It provides a new structure for the organization of bibliographic data based on the Functional Requirements for Bibliographic Records (FRBR), with more emphasis on identifiers and relationships than on descriptions. By 2013 many major national and research libraries had implemented the new standard. In November 2016 the RDA Steering Committee (RSC, www.rda-rsc.org) announced steps toward progressive adoption of the IFLA Library Reference Model (LRM, approved by the IFLA committees in August 2017), replacing the Functional Requirements family of models.

It is useful to recall that RDA support the Linked Data environment also through the representation of RDA entities, elements, relationship designators, and vocabulary encoding schemas in Resource Description Framework (RDF, the syntax of the semantic web) in the RDA Registry.

This paper is concerned with the application of the RDA standard within the field of Linked Data and how it may be used to improve the quality of the data produced in order to reach all the advantages that the semantic web can bring to the cultural heritage sector. More specifically it will look at a series of projects that start from analysis and manipulation of authority and bibliographic records and convert them in Linked Open Data following the BIBFRAME model. The Share Linked Open Data (SHARE-LOD) projects try to make visible and tangible a theoretical bibliographic context to experiment in a concrete environment the usability and re-usability of data. The common aim of these projects is also to make possible a revolution in creating, sharing and consuming of info, that starts by a record-oriented approach to arrive to an entity-oriented vision.

The theoretical context of SHARE-LOD projects

New standards, models and technologies offer new ways to approach entity identification and the relationships between entities, recognised as the key elements in the creation of new entity detection and entity identification processes. The context of the SHARE projects, is illustrated in Figure 1. If we consider the contribution of the new international RDA guidelines, as well as of the Linked Open Data philosophy and technology, both of these conceptual and structuring models refer to ways of approaching entity identification and the relationships between entities, and are therefore recognized as key drivers in the construction of new entity detection and entity identification processes.

JLIS.it 9, 1 (January 2018)
ISSN: 2038-1026 online

DOI: 10.4403/jlis.it-12447

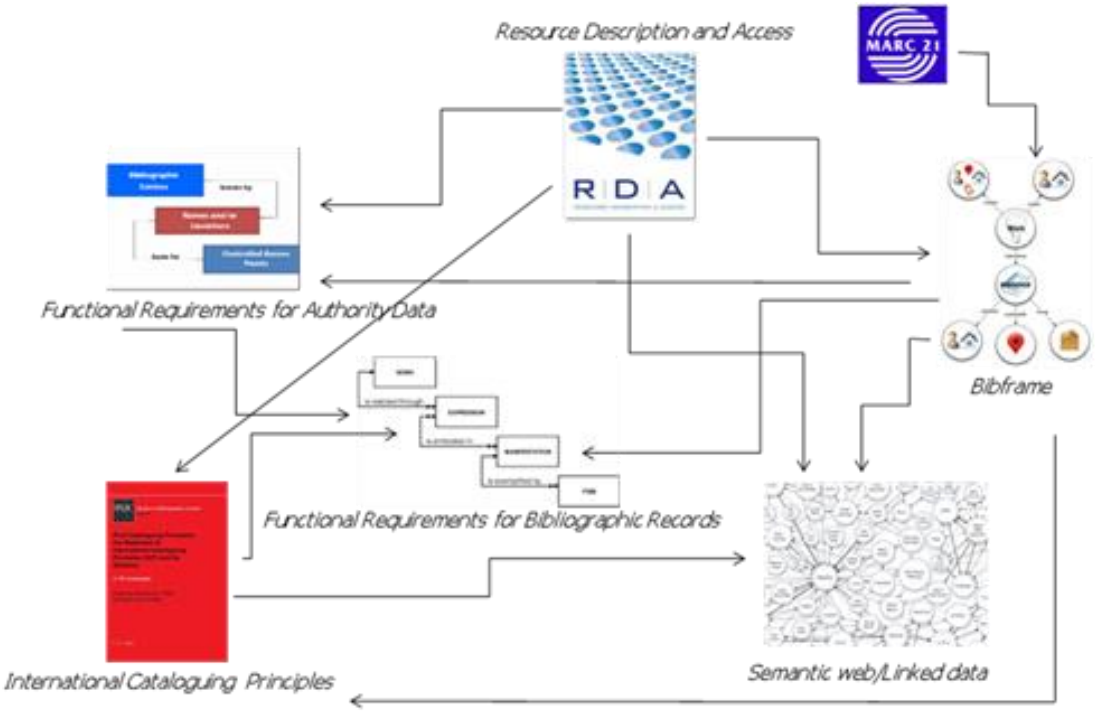


Figure 1: A summary of the theoretical context of the SHARE-LOD projects.

One of the main functions of RDA is to identify and link the entities deriving from the FRBR and FRAD models:

Identifying: FRBR Group 1 and Group 2 entities, and in future also Group 3, are identified by selecting attributes. Identifying involves recording the attributes of an entity by way of a procedure very similar to that of creating authority files for that entity. This allows for systematic guidelines for the identification of all entity types covered by the FRBR model: persons, families, corporations, works, expressions, manifestations and even items.

Linking: The entities of FRBR Groups 1 and 2, and in future also Group 3, are linked through the creation of relationships.

RDA is considered a content standard in line with the FR family, it deals with defining the essential elements for describing and providing access to a resource. The essential RDA elements for describing resources have been selected in line with the FRBR/FRAD evaluation of each attribute and relationship to facilitate the following user functions:

- identifying and selecting a manifestation;
- identifying works and expressions embodied in a manifestation;
- identifying the creator or the creators of a work.

The terms *identify* and *link* summarise the two fundamental objectives of RDA:

JLIS.it 9, 1 (January 2018)
ISSN: 2038-1026 online

DOI: 10.4403/jlis.it-12447

- *identify* a resource by selecting a group of attributes that distinguish it from another resource;
- *link* the resource to other, related resources by creating relevant relationships.

The structure of the RDA Toolkit clearly expresses the importance that the concepts of identification and relationship contribute to the standard:

Section 1: Recording Attributes of Manifestations & Items

Section 2: Recording Attributes of Works & Expressions

Section 3: Recording Attributes of Agents

Section 4: Recording Attributes of Concepts, Objects, Events & Places

Section 5: Recording Primary Relationships between Works, Expressions, Manifestations & Items

Section 6: Recording Relationships to Agents

Section 7: Recording Relationships with Concepts, Objects, Events & Places

Section 8: Recording Relationships between Works, Expressions, Manifestations & Items

Section 9: Recording Relationships between Agents

Section 10: Recording Relationships between Concepts, Objects, Events & Places

These concepts of *identification* and *relationship* also form the basis of the Linked Data model, and can be summarised simply by Sir Tim Berners-Lee's four rules:

1. Use URIs as names for things: give *unique names* to things;

2. Use HTTP URIs so that people can look up those names: the names assigned to things must also be machine readable;

3. When someone looks up a URI, provide useful information, using the standards (RDF, SPARQL): things must be self-explanatory (dereferencing);

4. Include links to other URIs, so that they can discover more things: create links with other objects (any object can become the subject of a new statement).

Following on from the first two main components of SHARE-LOD, the third and final is BIBFRAME. It was initially designed in 2012, is a data model that uses the principles of Linked Data and aims to provide an alternative to MARC. The MARC (MAchine-Readable Cataloging) format was developed in the 1960s and since then has become the international standard format for the encoding and exchange of bibliographic data. BIBFRAME[1] proposes three core classes: Work, Instance, Item; Persons, Families or Corporate bodies are within an Agent relationship with the Work in the data model. While libraries hold a wealth of well-organized information, the MARC format is not suited to the Semantic Web as the linear and static nature of the information it contains

[1] http://www.loc.gov/bibframe.

JLIS.it 9, 1 (January 2018)
ISSN: 2038-1026 online

DOI: 10.4403/jlis.it-12447

cannot easily be harnessed and linked to other, related resources. Version 2.0 of BIBFRAME, a schematic of which is shown in Figure 2, was released by the Library of Congress in November 2016 and updates, inclusive of community input, are ongoing.

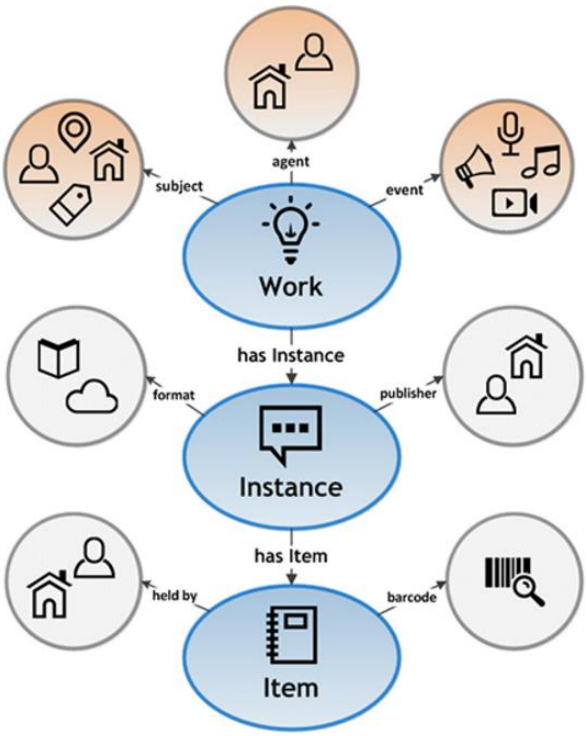


Figure 2: A schematic of the BIBFRAME data model version 2.0.

The question at hand: how to identify an entity?

The concept of entity identification is highly relevant in the construction of pathways for researching and locating resources. This is why it has traditionally been considered a highly important aspect of cataloguing. However, although the importance of identification is deeply rooted in the cataloguing tradition, in practice it has often fallen short of expectations: for example, the use of attributes such as date of birth (and death where necessary) to identify a person has not been widely used. Sometimes this occurred through negligence and sometimes in an attempt to save money but, over time, it is an approach that has revealed itself to be highly questionable. One example is the entity - person *Francesco Guicciardini*, an Italian politician who lived between the 19th and 20th centuries. The lack of biographical data associated with his name in many bibliographic catalogues caused him to be confused with the more famous Italian writer, historian and politician of the same name who lived between the 15th and 16th centuries, causing much uproar. The two figures, clearly identified in encyclopaedias, were less clearly distinguished in bibliographic catalogues, causing the two entities to be falsely merged in the processes of reconciliation, as shown in Figures 3 and 4.

JLIS.it 9, 1 (January 2018)
ISSN: 2038-1026 online

DOI: 10.4403/jlis.it-12447

Visualizzazione completa del record

Scegli formato: Standard Scheda catalografica Tag nomi Tag MARC Abstract

Autore	Guicciardini, Francesco
Titolo	La Cassa Nazionale di previdenza per la invalidita e la vecchiaia degli operai / Francesco Guicciardini.
Pubblicazione	Firenze : ..., 1901.
Descrizione	32 p. ; 23 cm
Soggetto	Previdenza sociale
CDD	360

Figure 3: The bibliographic record for the manifestation La Cassa nazionale di previdenza per la invalidità e la vecchiaia degli operai, whose author, Francesco Guicciardini, is not uniquely identified with a date of birth and death.

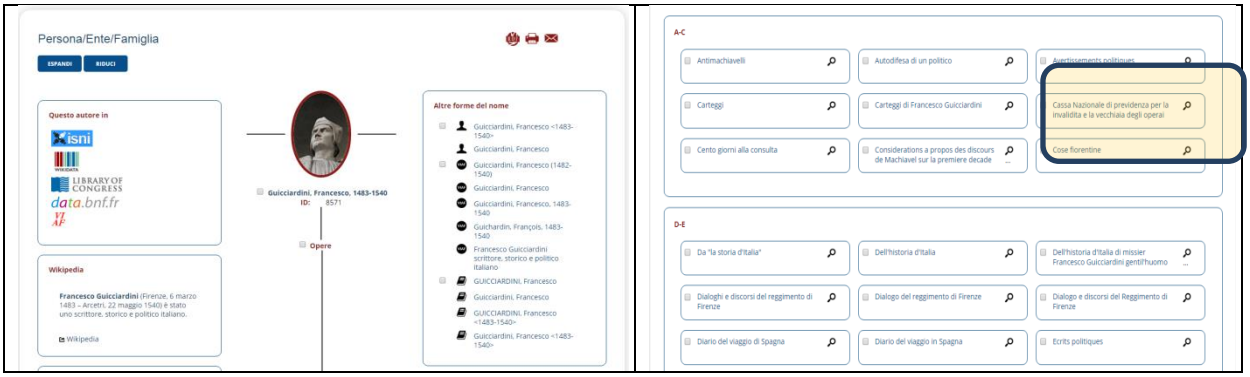


Figure 4: The outcome of a data reconciliation process that wrongly associates the title Cassa nazionale di previdenza per la invalidità e la vecchiaia degli operai with Francesco Guicciardini, 1483-1540.

These developments allow for new cooperative scenarios between institutions and corporations, further removed from a complex *reductio ad unum* approach and physical merging. With the new generation of Authority control and discovery tools enforcing cross-institutional processes of cooperation, integration and virtualization. This creates data enrichment opportunities that were previously inconceivable, putting the focus on identifying entities and discovering their relationships with other entities.

Data reconciliation, enrichment and conversion

With the on-line presence of different catalogues and authority files available in various formats, where possible in an open access model, the concept of authority control has evolved into the grouping of an entity's identifying attributes from different sources. The process is best known as *reconciliation* and consists of creating a cluster of data that all refer to the same entity.

The term *reconcile*, from the Latin *reconciliare*, made up of *re-* and *conciliare* i.e. 'to bring together, conciliate', immediately clarifies the concept behind the process in question: bringing together the different name variants referring to the same entity. The most common understanding of the term is even more significant: 'to bring to agreement, restore to peace and harmony'. Indeed, this *harmony* is the ultimate objective of data reconciliation: reconciling data begins with the assumption that an

JLIS.it 9, 1 (January 2018)
ISSN: 2038-1026 online

DOI: 10.4403/jlis.it-12447

entity may be known by different names deriving from differences in culture, cataloguing rules and linguistics as well as simple typographical errors; accepting this variety and making it into a virtue.

Reconciliation, or clustering, has been at the centre of major endeavours such as VIAF and ISNI projects: the idea being to take various classifying data from different projects and sources, and make it available in a way that could be defined as *democratic* (without necessarily privileging one form over another) to better identify the entity in question.

Even wider reconciliation processes form the basis of a number of projects that convert and publish bibliographic catalogues as Linked Open Data, such as the SHARE Catalogue project, introduced at the 2016 Convegno delle Stelline conference and realized within 8 university libraries in the South of Italy, or the more recent SHARE-VDE (SHARE Virtual Discovery Environment) venture being developed by a group of 16 North American institutions, with a united aim, but the ability to maintain individual ILSs, practices and cataloguing traditions.

The enrichment of records, derived through connections to authority files (in a centralized, distributed or local model), other external sources and from clustering data from specific projects, has extraordinary potential to enhance their function. It enables end users to expand their research on the entity increasing their chance of finding new information and resources, while at the same time allowing libraries to consult other authoritative general or specialized sources.

These conditions are also prerequisite for a revolution in the concept of cataloguing and how a catalogue is presented. The change from the record in its entirety having meaning in its rigidity, to the entities as *real things in the world*, recognising how flexibility and diversity can enrich information, as illustrated in Figure 5.

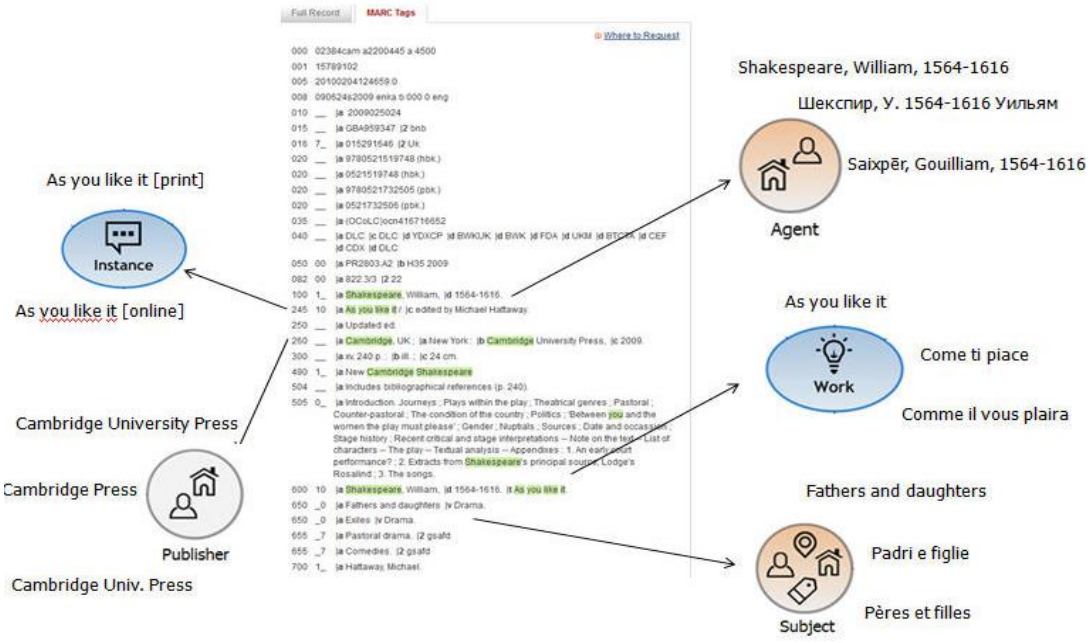
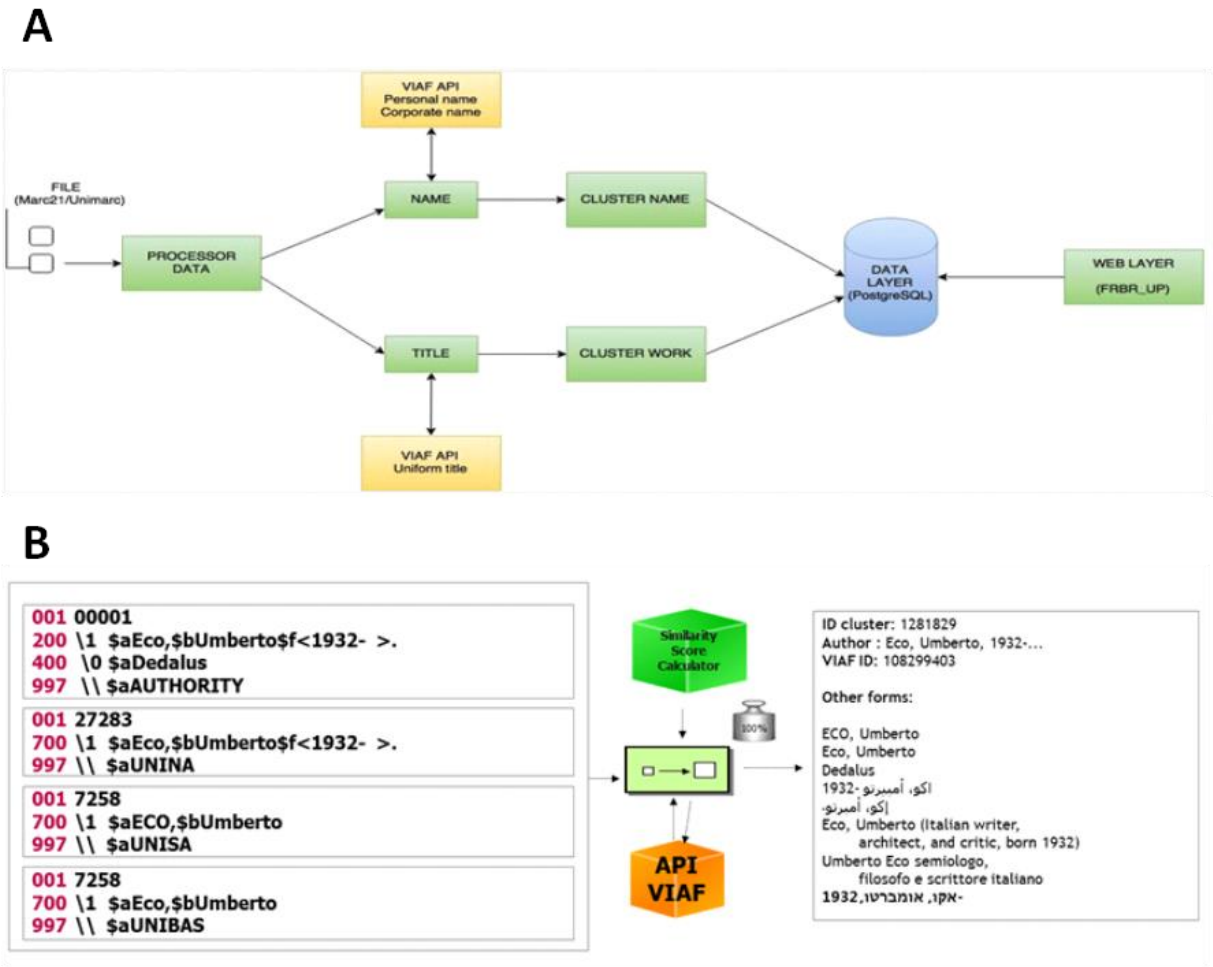


Figure 5: The new revolution: from record to entity.

JLIS.it 9, 1 (January 2018)
ISSN: 2038-1026 online

DOI: 10.4403/jlis.it-12447

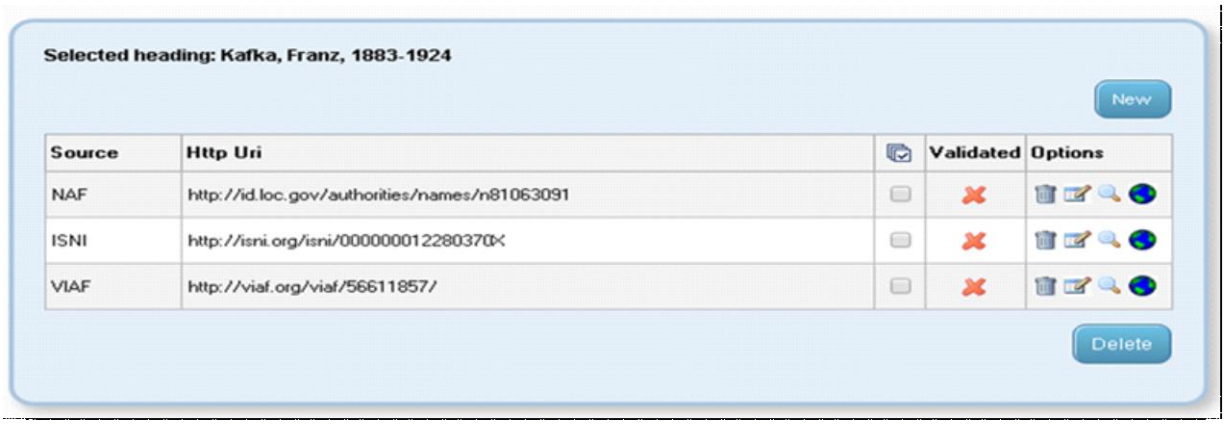
Data reconciliation and enrichment is obtained by means of complex logics and algorithms (data comparison, results filtering, validation etc.), which may be carried out using either automated systems or manual processes, included (where the ILS permits it) in the cataloguing workflow. The relationship between the reconciliation and validation of the results can differ profoundly between the automated and manual processes, as the automated processes assure a high-level of reconciliation and clustering with a low-level of validation of results versus the manual processes with a low-level of reconciliation and clustering and a high-level of validation of results. The best outcome, based on the weighing of parameters during the automated process, can often be a compromise of the two.



Figures 6A and 6B: Authify tool in SHARE-LOD projects to obtain more comprehensive and precise URIs retrieval and automated process of cluster creation for Person- and Work-type entities.

JLIS.it 9, 1 (January 2018)
ISSN: 2038-1026 online

DOI: 10.4403/jlis.it-12447



Figures 7: Using the URI Management System in the WeCat cataloguing module of the OLISuite ILS: this is an example of manual entity enrichment carried out in the cataloguing workflow (the availability of APIs and web services allows the use of external sources - in this case, NAF, ISNI and VIAF).

In SHARE-LOD projects Authify is the tool for automated reconciliation. It is a RESTFul module that offers several full-text search services among names and work clusters, and relator terms detection services. The manual tool is the URI Management System embedded into Casalini Libri's OLISuite WeCat cataloging system (developed by @Cult).

In addition to these there are two more components that are part of the overall system:

1. The *Database of relationships* created from the analysis of bibliographic and authority records with the aim to make evident the relationships that are contained within these records (between author and publishers, author and subjects, publisher and areas of interest, authors and collaborators, titles and ISBN etc.). The final goal of these procedures being to provide a more effective identification of the entities of interest starting from a traditional (record-oriented) environment.
2. The *Knowledge base of clusters* with GET services to retrieve the cluster data and PUT services to create new clusters. A common knowledge base as a web accessible source with reconciled entities identified with RWO URIs can also be made accessible via API/WS or SPARQL endpoint in RDF format.

Starting by the end

Current catalogue data predominantly contains descriptions of Manifestations/Instances (following the FRBR or BIBFRAME data model). The objective is now to respond to the need to re-design this data model to include a system that derives data from existing records to produce a new, higher Person / Work layer giving significant advantages for the end user.

In order to achieve this aim the data, after being processed through the steps described above, are presented on a portal equipped with navigational tools based on the BIBFRAME data model characterized by three different layers:

– Person/Works: this level is enriched by data from sources external to the library catalogues for the purpose of extending the research potential;

JLIS.it 9, 1 (January 2018)
ISSN: 2038-1026 online

DOI: 10.4403/jlis.it-12447

- Instances (or Publications): the Instances level is associated with Publications and connected to the overlying layer through relationships with the Works present;
- Item: each Instance (Publication) is linked to information about the data set and the availability of the copy present in the local OPAC of each library.

In order to move progressively toward a record-less approach the platform also addresses the Instances reconciliation aspect. On the Item level, API or web services can be implemented to communicate with the local OPAC. In addition, diversified user interfaces can be applied to meet different user community needs.

The first version of the SHARE platform was developed by @Cult whilst working on a smaller scale project that went into production in spring 2016 and that became a model for other future projects. It involved 8 Italian university libraries that used and continue to use different local systems: some based on MARC21, others on UNIMARC, also applying different cataloguing codes.[2]

A further application is ilibri-up, an enhancement of Casalini Libri's existing ilibri bibliographic database, which will also serve the main link for the ISNI Registration Agency activities of Casalini Libri.

As mentioned previously, the SHARE-VDE project for the creation of a Virtual Discovery Environment involves 16 North American institutions and its main aims are as follows:

Conversion, supply and management of authority and bibliographical data in BIBFRAME taking into account the complexity of the long and heterogeneous transition time;

Development of detection services for entity identification including relator terms, and creation of a common knowledge base of clusters of reconciled results for persons and works;

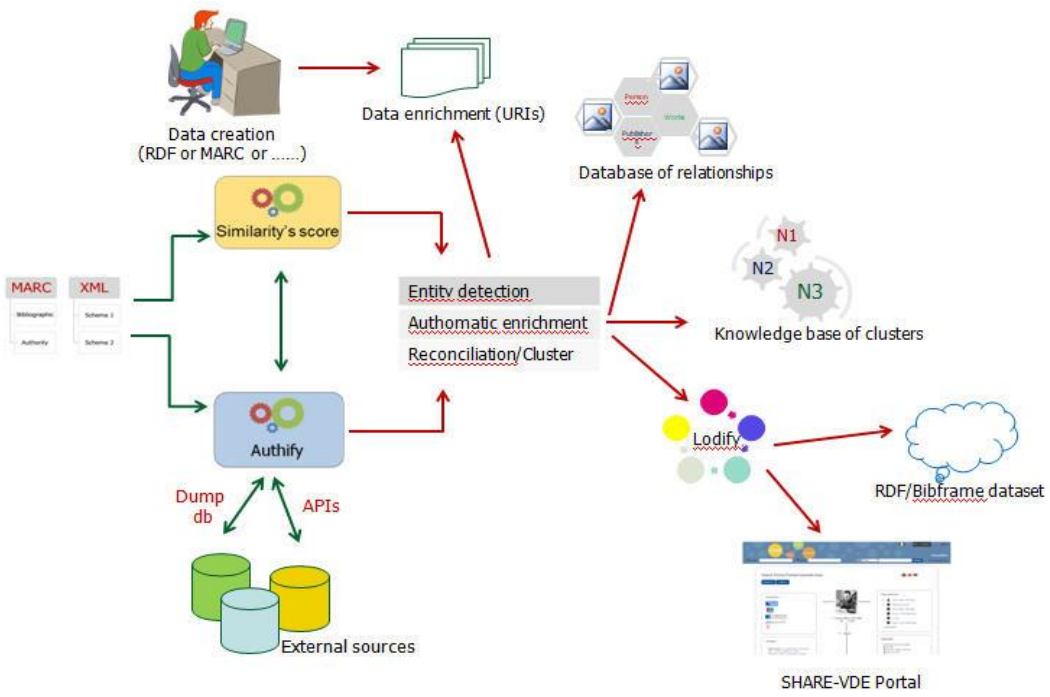
Publication of a FRBR/BIBFRAME three layered platform with build-in instances techniques.

The following series of figures depicts an overview of the SHARE-VDE processes as well as a series of examples. The SHARE-VDE initiative is based on the partnership of Casalini Libri and @Cult.[3]

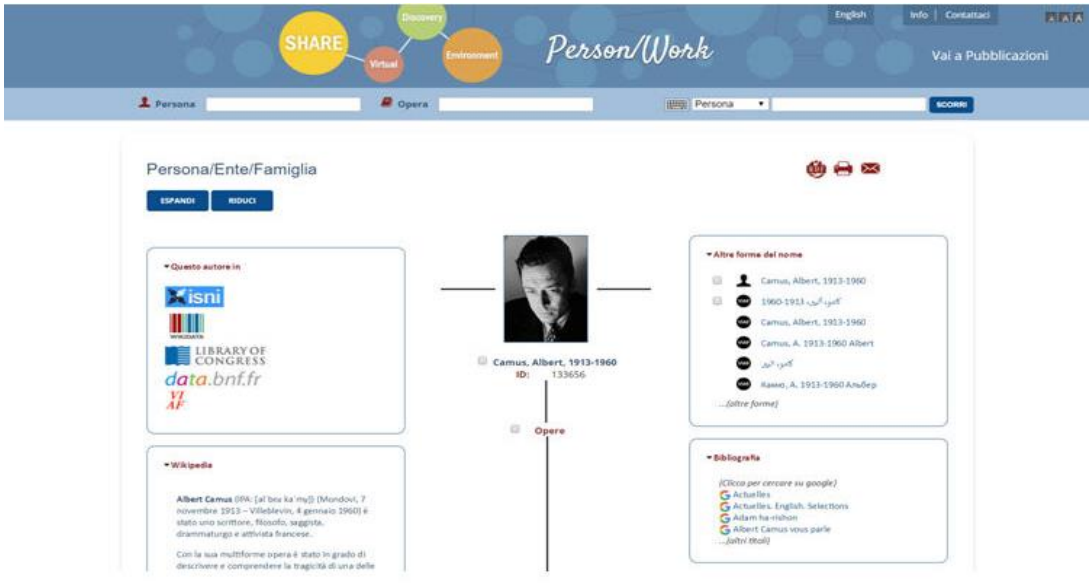
[2] The platform can be viewed at http://catalogo.share-cat.unina.it/sharecat/clusters?l=en.
[3] The platform is accessible at http://www.share-vde.org.

JLIS.it 9, 1 (January 2018)
ISSN: 2038-1026 online

DOI: 10.4403/jlis.it-12447



Figures 8: SHARE-VDE overall processes

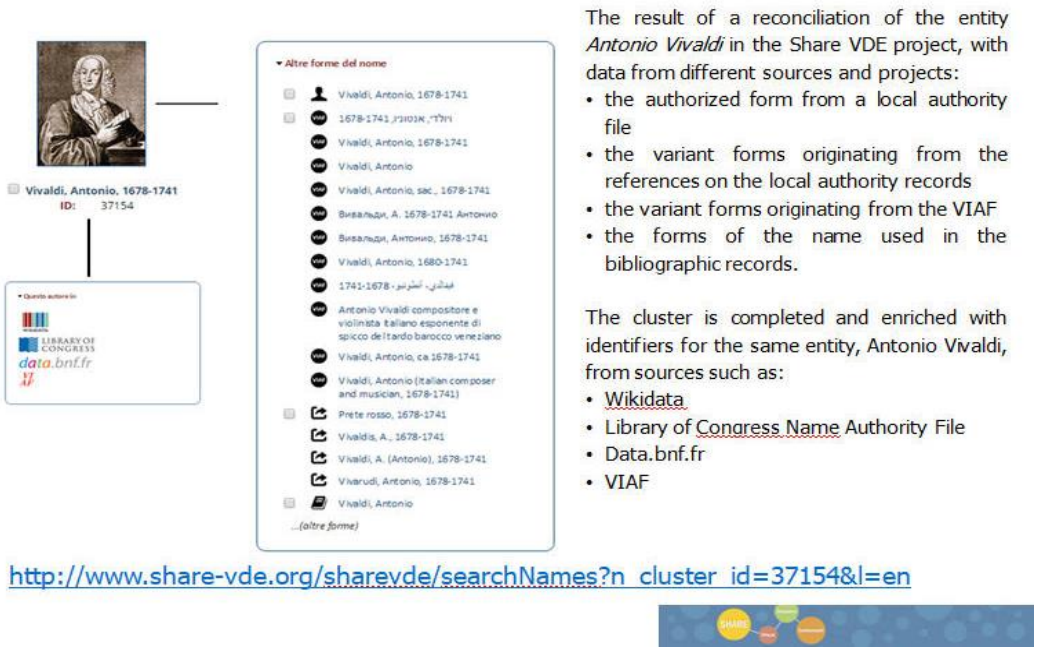


Figures 9: Albert Camus on the SHARE-VDE platform.[4]

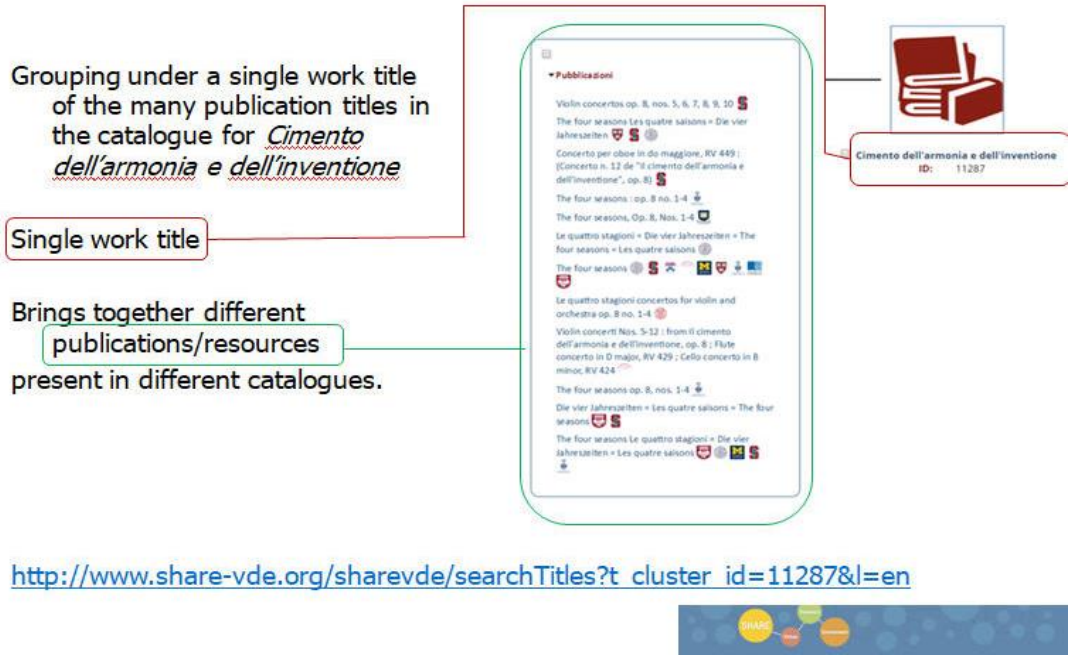
[4] http://www.share-vde.org/sharevde/searchNames?n_cluster_id=133656.

JLIS.it 9, 1 (January 2018)
ISSN: 2038-1026 online

DOI: 10.4403/jlis.it-12447



Figures 10: Entities in cluster: an example of collaboration and sharing.[5]

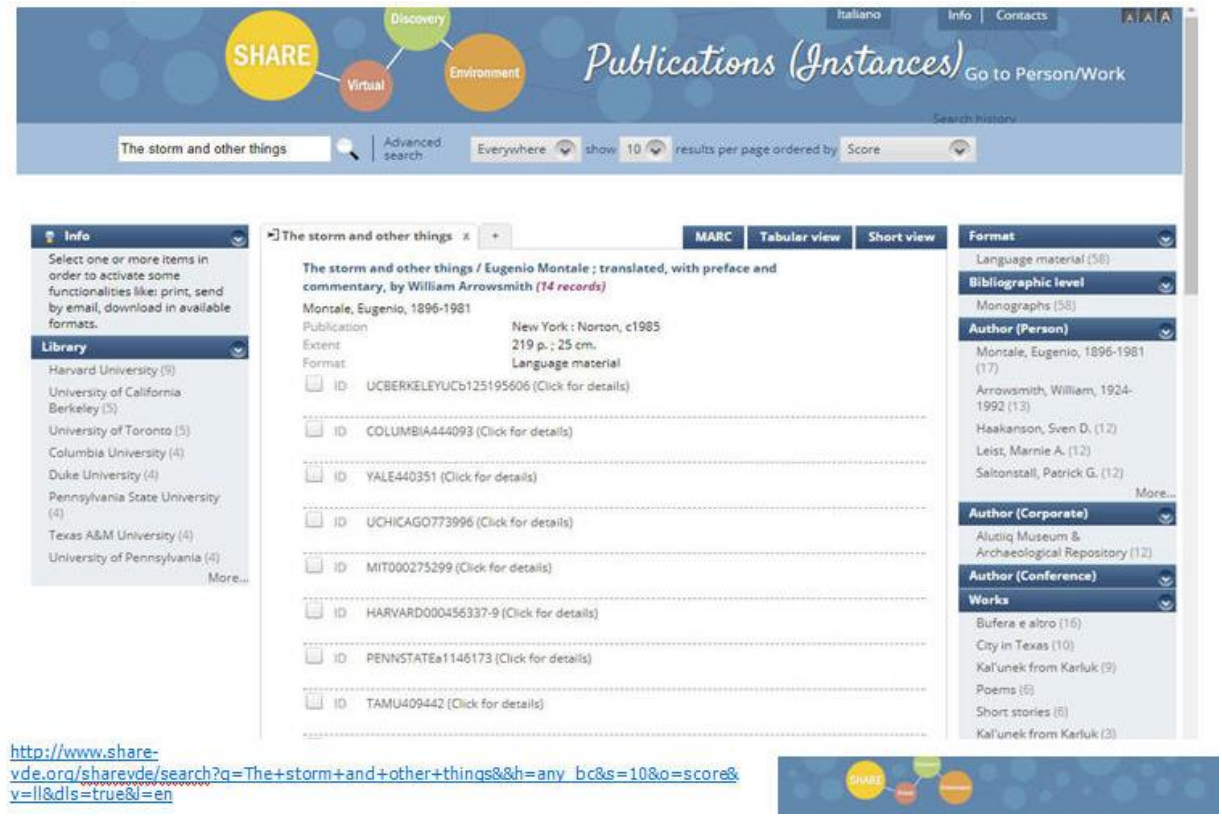


Figures 11: An example of Work/Instances reconciliation.[6]

[5] http://www.share-vde.org/sharevde/searchNames?n_cluster_id=37154&l=en.

JLIS.it 9, 1 (January 2018)
ISSN: 2038-1026 online

DOI: 10.4403/jlis.it-12447



Figures 12: Example of same Instances reconciliation for titles present in different library catalogues.[7]

Conclusions

A great effort is underway to facilitate the sharing and reuse of assets, and tools produced by libraries, museums and other institutions, and to guarantee their availability to a wider public. This endeavour is enriching the World Wide Web with valuable information that would otherwise remain mostly hidden in archives, collections, and catalogues and promotes a culture of open access to knowledge. The result has numerous advantages for all stakeholders. Libraries, archives, and museums all benefit from the possibility of more comprehensive and better structured tools, born out of positive cross-institutional cooperation. These provide end users with a vast wealth of information and create new cooperative tools for professionals within the sector. In line with the philosophy of open data, sharing and reuse, even traditional authority controls are evolving.

The discourse on whether in the future authority control would be centralised or localised has been rendered obsolete by this new way of working. This method focuses on the identification of entities and their relationships, catalysing a landmark transition from authority control to entity identification. Libraries, archives and museums can all benefit from the possibility of more well-structured and sharable data, providing users with a vast wealth of information, and creating new collaborations.

---

[6] http://www.share-vde.org/sharevde/searchTitles?t_cluster_id=11287&l=en.

[7] http://www.share.vde.org/sharevde/searchg=The+storm+and+other+things&&h=any_bc&s=10&o=scores&v=ll&dls=true&l=en.