

IFLA-Satellite-Meeting 2017
Digital Humanities. Connecting Libraries and Research
<http://dh-libraries.sciencesconf.org>

BIBFRAME and Linked Data practices for the stewardship of research knowledge

Michele Casalini
Managing Director
Casalini Libri
michele@casalini.it
www.casalini.it

 orcid.org/0000-0003-4643-8895

Abstract

This article considers need for more visible, available, accessible, innovative and shared bibliographic data in the internet age and the subsequent benefits of these transformations for galleries, libraries, archives and museums. Recent and ongoing research and development activities in the following fields are explored: entity identification, reconciliation, data enrichment, MARC records enriched with URIs, conversion to RDF, creation of relationship criteria for the improved identification of entities and a knowledge base of clusters that uses the paradigms of the semantic web. These improvements are discussed in the context of the BIBFRAME (Bibliographic Framework Initiative) data model and associated projects such as SHARE-VDE. The aim of the article is to outline current and future research and development activities in collaboration with the library community concerning the dissemination and discoverability of bibliographic data and research knowledge.

BIBFRAME and Linked Data practices for the stewardship of research knowledge

1. BIBFRAME and Linked Data

The emerging BIBFRAME (Bibliographic Framework Initiative) data model for the future advancement of bibliographic formats is currently the subject of discussion and development within the galleries, libraries, archives and museums (GLAMs) community. The new framework is intended to open up the possibilities of Linked Data, providing greater visibility and discoverability of all resource types, embracing various scripts, making bibliographic information more flexible and accessible to end users across the web rather than just library, archive or museum patrons. Many organisations are beginning to experiment with this framework and develop new workflow and business models to respond to changing needs.

The aim of this paper is very practical, focusing on recent and ongoing research and development activities: entity identification, reconciliation, data enrichment, with URIs (Uniform Resource Identifier) enhanced MARC records, conversion into RDF (Resource Description Framework), creation of relationship criteria useful to increase the effectiveness of entity identification, knowledge base of clusters that uses the paradigms of the semantic web, the BIBFRAME three layer architecture portal will be addressed. The SHARE Virtual Discovery Environment in Linked Data project encompassing a group of North American institutions with a range of different systems, habits and cataloguing traditions will be described (www.share-vde.org).

The theoretical context of these recent developments is:

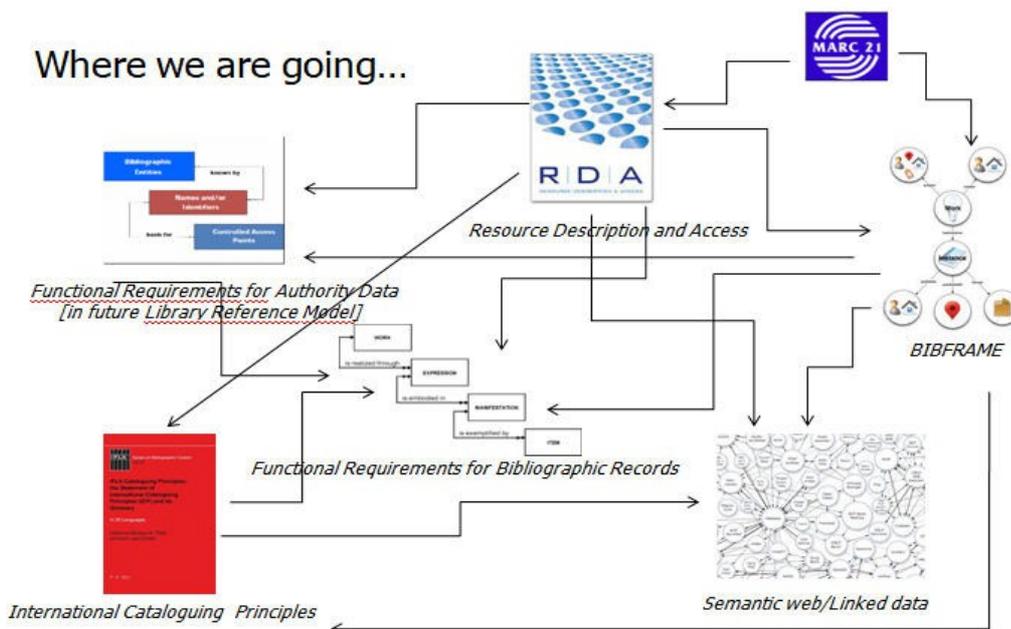


Figure 1: Brief theoretical context

- RDA (Resource Description and Access), initially released in 2010 and particularly appropriate for use by libraries, archives and museums replaces the Anglo-American Cataloguing Rules, Second Edition (AACR2). It provides a new structure for the organisation 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, currently under final formal approval by the IFLA committees), replacing the Functional Requirements family of models.
- BIBFRAME, 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 (www.loc.gov/bibframe) proposes three core levels: Work, Instance, Item; Persons or Corporate bodies are within an Agent relationship with the Work in the data model. While libraries hold a wealth of well organised information, the MARC format is not suited to the Semantic Web as the linear and static nature of the information it contains cannot easily be harnessed and linked to other, related resources. Version 2.0 of BIBFRAME was released by the Library of Congress in November 2016 and updates, inclusive of community input, are ongoing.

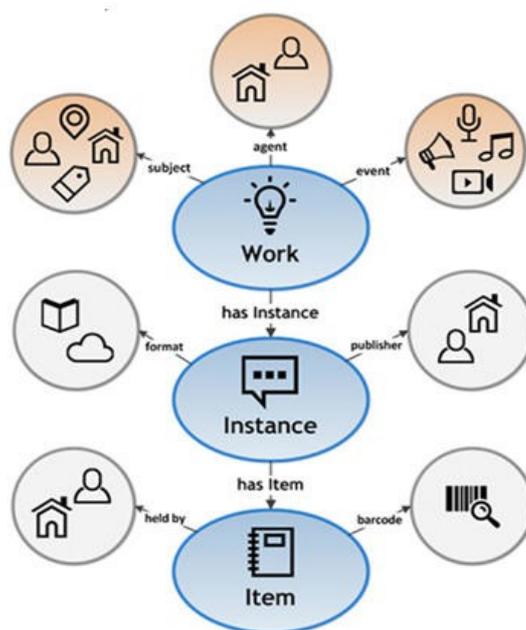


Figure 2: BIBFRAME 2.0 data model

2. Experience

Casalini Libri was established in 1958 with the dual purpose of advancing the profile of Italian culture and learning across the globe as well as providing a first-class bibliographic search and supply service for academic libraries. The company has grown considerably since its foundation by Mario Casalini, becoming one of the leading suppliers of European publications and related library services. Still a family-run business, Casalini Libri's ninety-strong team remains both faithful to the traditions of the business and committed to innovation, facilitating selection, acquisition and processing workflows working with thousands of publishers and libraries.

One of the priorities of Casalini Libri (www.casalini.it) has always been the provision of quality bibliographic information. The company produces more than 40,000 original bibliographic records for Romance Language publications each year, all of which are accessible through the online ilibri database (www.ilibri.com). Casalini Libri contributes new authority records to the national authority file as well as maintaining existing records, makes subject and classification proposals through participation in the NACO and SACO programs of the Program for Cooperative Cataloguing (PCC), and became recently ISNI Registration Agency. The records are created in native MARC21 according to the RDA BIBCO Standard Record (BSR) guidelines using the in-house WeCat cataloguing module of the OLISuite ILS, developed by @Cult.

A consulting and software development company established in 2001, @Cult (www.atcult.it) delivers effective and innovative tech solutions to improve information management and knowledge sharing. Casalini Libri's IT Department and @Cult have a longstanding collaboration and have been working together for over 10 years. One of the several projects in the field of the semantic web for cultural heritage institutions which @Cult has been involved in is ALIADA.

The ALIADA (Automatic publication under Linked Data Paradigm of Library Data) project was co-financed by the European Union's Research and Innovation funding programme and ran from 2013 to 2015. It involved five partners from Italy, Spain and Hungary. The project originally applied the Linked Data paradigm using FRBRoo based ontologies. The project was developed with the aim of supporting the entire process, from conversion to the publication and linking of data. The results of the project are available from www.aliada-project.eu.

3. BIBFRAME and Linked Data practices: the SHARE Virtual Discovery Environment in Linked Data project (SHARE-VDE)

Following the success of ALIADA and after input received from the library community, in 2014 Casalini Libri embarked upon a strategy for the progressive implementation of the BIBFRAME data model in close collaboration with @Cult. After a year of initial study and a feasibility analysis Casalini Libri presented the first results from the joint venture and a plan for action at the BIBFRAME LC Forum in Boston January 2016. In the subsequent months, following discussions with the library community, a research and development project, so far involving sixteen North-

American institutions, was established: SHARE Virtual Discovery Environment in Linked Data (www.share-vde.org).

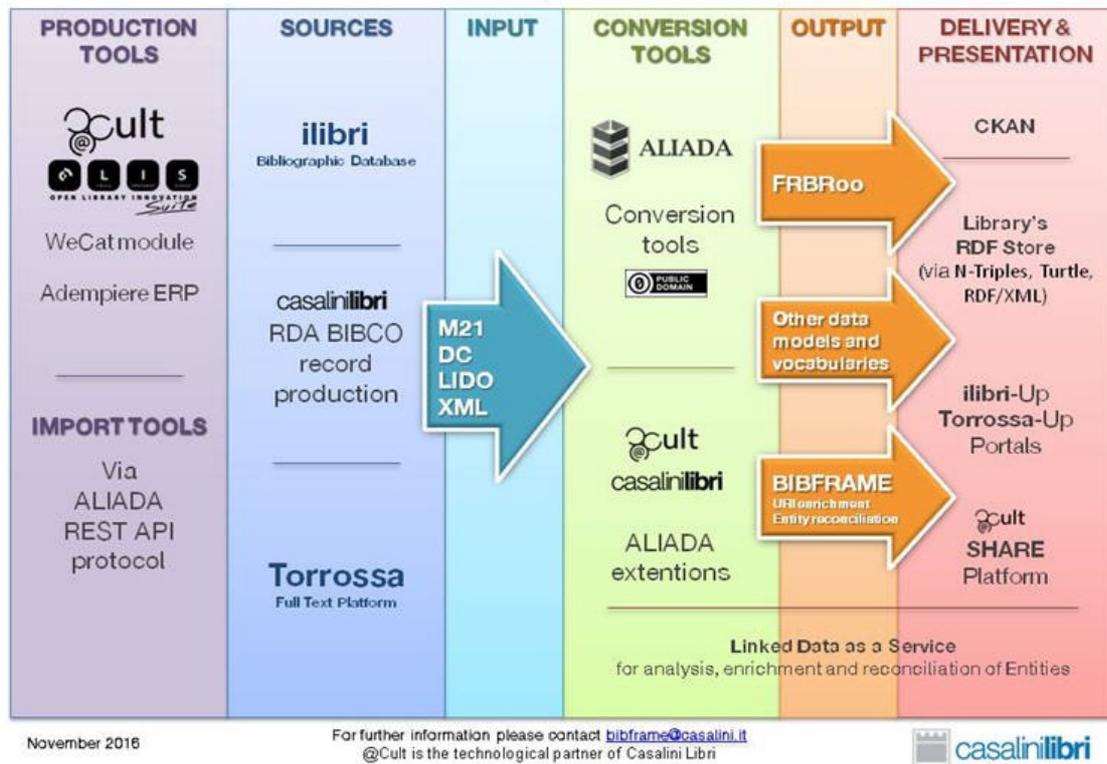


Figure 3: Casalini Libri's BIBFRAME Conversion, Distribution & Publication Options

3.1 Project highlights

SHARE-VDE is a library community driven research and development project to establish entity identification, reconciliation and conversion processes as well as a prototype of a virtual discovery environment with a BIBFRAME three layered architecture (Person/Work, Instance, Item). Furthermore, the project will create a database of relationships that is open to the community and a common knowledge base of clusters that uses the paradigms of the semantic web but also allows the libraries to continue to handle their data as independently as possible.

In terms of outcomes for partner libraries the project aims to create an environment that is useful for both library patrons, empowering them with advanced discovery interfaces, and librarians, incrementally providing them with cataloguing functions in native semantic web standards, integrating processes with the local systems, and implementing tools in a collaborative environment. SHARE-VDE also hopes to help to reveal a richness within the data of existing collections, often hidden or unexpressed in a traditional catalogue. This will be carried out through discussion, experimentation and configuration of the options for the future data creation, enhancement and sharing of all type of resources with the library community. Emphasis is placed on the use of short phases, yielding tangible results, on which institutions can base future decisions and further steps.

Among the guiding principles of the initiative are the need for independence from the different local systems (ILS), habits and cataloguing traditions already in place; for the components to be available individually in as flexible as possible configuration as expected by the needs of various library - but also archive and museum – community groups. Scalability of the tools is highly important and the project aims to test these new tools using over 100 million traditional bibliographic and authority records. Alongside the ability to address issues and problems related to new information management processes, and taking into account the complexity of the long transition time that will see the coexistence of native MARC and RDF data.

The project is divided in three phases and is directed by Tiziana Possemato, Chief Information Officer of Casalini Libri and Director of @Cult.

Phase 1 (October 2016 – January 2017): two sets of data for each participating library were anticipated, consisting of the titles with imprint year 1985 and 2015 within each library's system. The advantage of having two data sets was the ability to test the processes in a complex environment with a significant number of cases that come from different stages of the library catalogue, including various type of resources and scripts. This phase resulted in a total of 2,308,204 bibliographical records with 3,601,327 authority records being reconciled, enriched, converted into BIBFRAME 2.0 and published on the SHARE-VDE portal.

Phase 2 (March – December 2017): the complete library catalogue of each participating institution are converted into BIBFRAME 2.0 and returned to each library following an overall reconciliation process and applying enhanced tools based on the feedback from phase 1. Over 100 million records and associated datasets are on course to be processed, undergoing the multiple coherent steps to conversion.

Phase 3 (2018 onward): this phase will be implemented in production based on input and use cases collected from the interested institutions.

The institutions participating in phase 1, 2, or both phases of the project are the following: Stanford University, University California Berkeley, Yale University, Library of Congress, University of Chicago, University of Michigan Ann Arbor, Harvard University, Massachusetts Institute of Technology, Duke University, Cornell University, Columbia University, University of Pennsylvania, Pennsylvania State University, Texas A&M University, University of Alberta, University of Toronto.

In the following sections the four major components of the project are described.

3.2 Entity identification, reconciliation and data enrichment

Entity identification is a highly relevant component of the pathways for researching and locating resources. This is why it has traditionally been considered an integral aspect of cataloging. However, the use of attributes to uniquely identify a person or a work has not previously been widely used.

With the presence online of different catalogues and authority files available in various formats, where possible in open mode, the concepts of authority control and of catalogue unification have 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. This entity may be known by different names deriving from cultural differences, disparity in cataloguing rules, linguistic variations, and simple typographical errors. Reconciliation accepts this variety, turning it to the advantage of both the cataloguer and the end user.

The enrichment of records, derived through connections to authority files (centralised, distributed or local), 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 specialised 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.

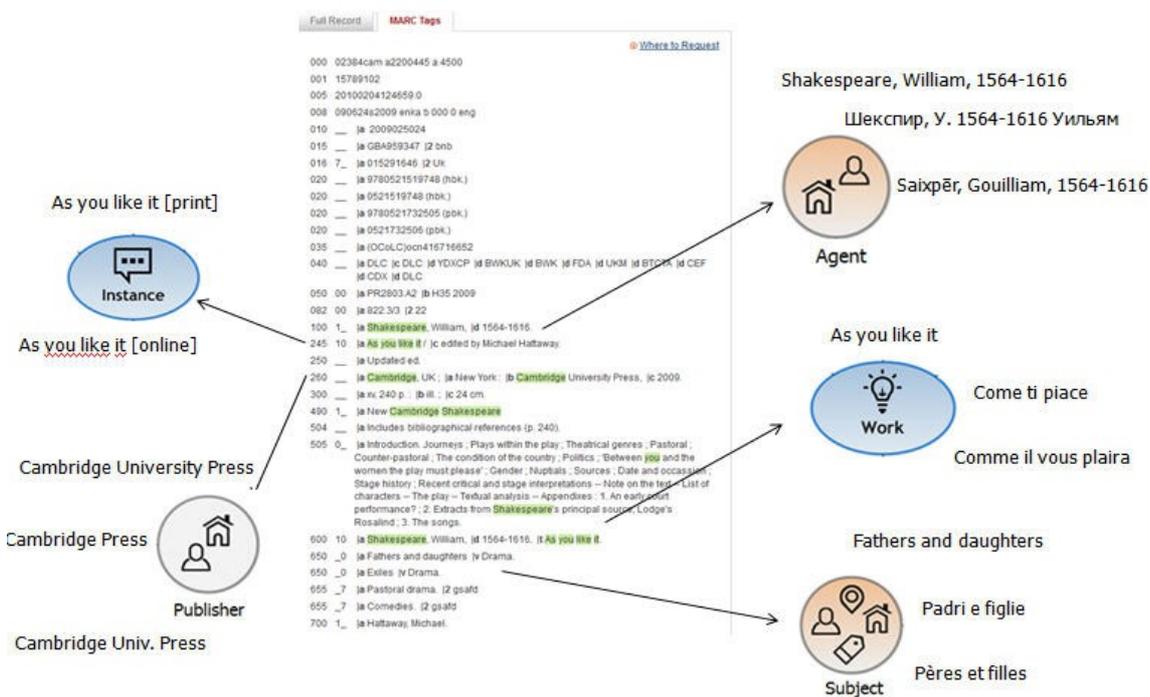


Figure 4: The new revolution: from record to entity

Data reconciliation and enrichment is obtained by means of complex logic 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.

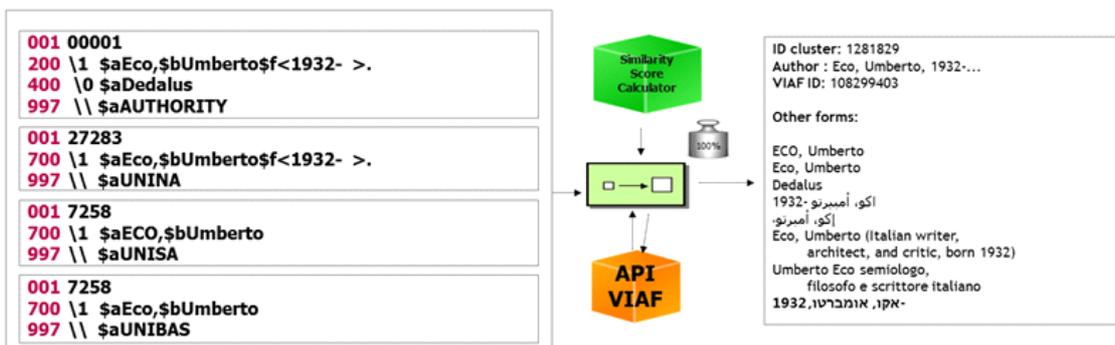
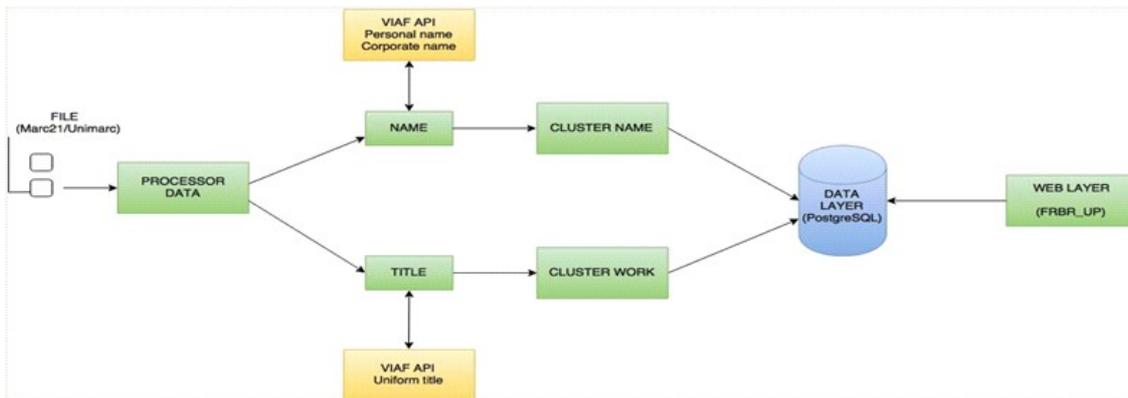


Figure 5 and 6: Authify tool in SHARE-VDE to obtain more comprehensive and precise URI retrieval and automated process of cluster creation for Person- and Work-type entities.

Selected heading: Kafka, Franz, 1883-1924

Source	Http Uri	Validated	Options
NAF	http://id.loc.gov/authorities/names/n81063091	<input type="checkbox"/>	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>
ISNI	http://isni.org/isni/0000000012280370X	<input type="checkbox"/>	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>
VIAF	http://viaf.org/viaf/56611857/	<input type="checkbox"/>	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>

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

In SHARE-VDE 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 suite.

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

- I. 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.) with the final goal of these procedures being to provide a more effective identification of the entities of interest.
- II. 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.

3.3 Enhanced MARC records with URIs

In the recent months important steps were achieved by the PCC Program for Cooperative Cataloging (PCC) Task Groups on URIs in MARC & BIBFRAME, in conjunction with the MARC Advisory Committee. These decisions embraced the redefinition of subfield \$0 for recording URIs which represent objects in RDF triple statements, and of subfield 4 to be used for recording URIs which represent predicates in RDF triple statements.

One aspect always to be taken into consideration in the application of data models is the conversion of data into alternative structures without the loss of content. To ensure an effective transition from the MARC record to BIBFRAME, the implementation plan must ensure that MARC data elements are enriched through the addition of the necessary local and global identifiers. Once the automatic and manual processes required for this procedure are established, MARC can be converted into Linked Data by any entity.

For an original cataloguing data producer and provider, such as Casalini Libri, it's important to embed the ability to parameterise workflows and foresee profiling options to handle, for example the personalisation of URI sources according to the preferences of each individual institution.

As the transition from MARC to the RDF environment will occur over a substantial period, heterogeneous systems will coexist for long time, probably also within the same institution.

3.4 Conversion of authority and bibliographical data in BIBFRAME

In SHARE-VDE Lodify is the tool in command of the conversion to RDF. It contains an asynchronous pipeline where the process is split into pieces (processors), each of these responsible for a small part of the overall task. Each processor can act as a splitter or aggregator and can achieve content manipulation of the incoming message.

Following the ALIADA methodology, Lodify converts each incoming record by means of conversion templates. Each template associates a MARC record from the incoming data-stream with a set of (conversion) rules associated with one or more ontologies.

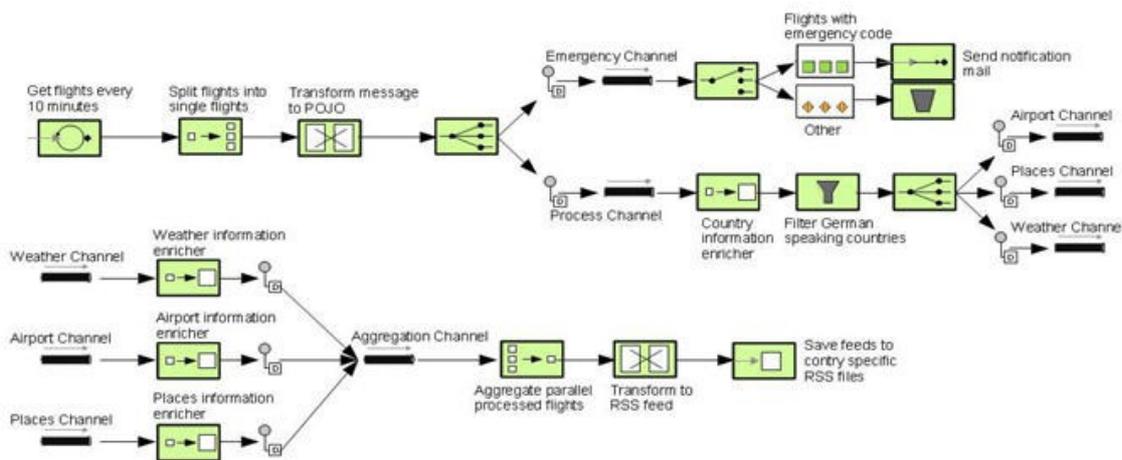


Figure 8: Lodify tool in SHARE-VDE for the conversion into RDF.

SHARE-VDE uses the triple-store Blazegraph, in conjunction with Solr search engine and PostgreSQL database.

3.5 Publication of a BIBFRAME three layered platform prototype

Current catalogue data predominantly contains descriptions of Manifestations (in FRBR) / Instances (in BIBFRAME). 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.

- 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 Instance 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 for 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. It involved seven Italian university libraries that used and continue to use different local systems: some based on MARC21, others on UNIMARC, also applying different cataloguing codes. The platform can be viewed at <http://catalogo.share-cat.unina.it/sharecat/clusters?l=en>.

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.

The SHARE-VDE platform is accessible at <http://www.share-vde.org>.

The following series of figures depicts an overview of the SHARE-VDE processes as well as a series of examples.

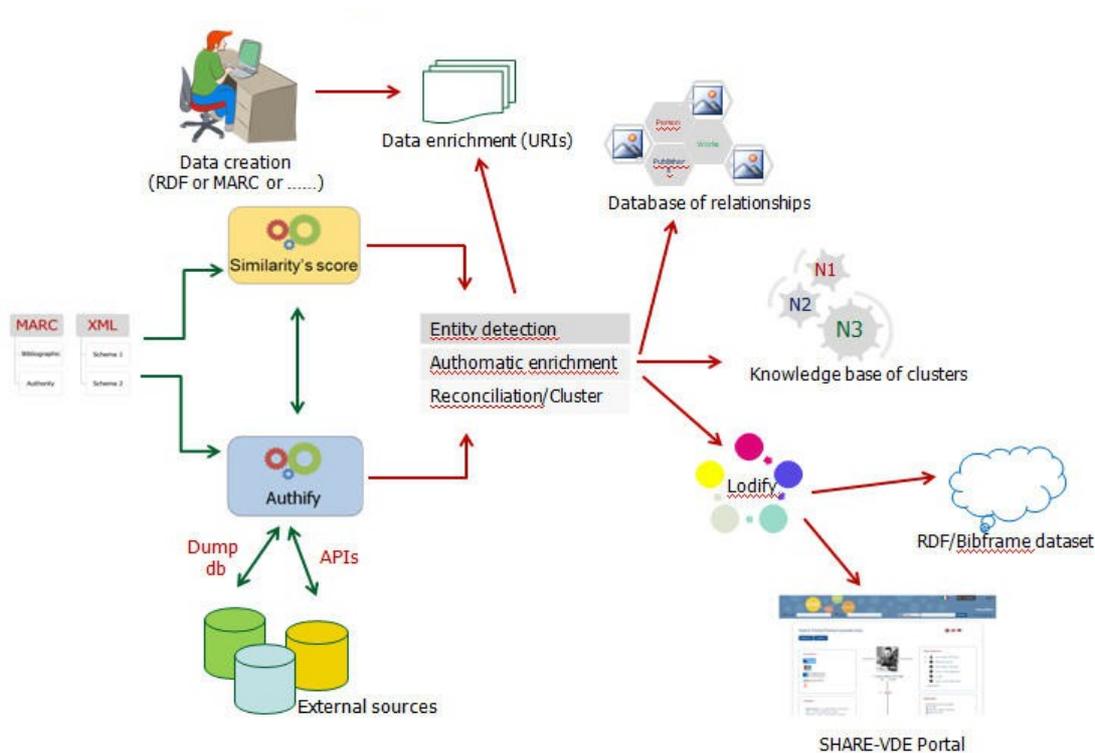


Figure 9: SHARE-VDE overall processes



Persona/Ente/Famiglia

ESPANDI RIDUCI

Questo autore in

- isni
- LIBRARY OF CONGRESS
- data.bnf.fr
- VZ
- AF

Wikipedia

Albert Camus (IPA: [al bœ ka my]) (Mondovi, 7 novembre 1913 – Villeblevin, 4 gennaio 1960) è stato uno scrittore, filosofo, saggista, drammaturgo e attivista francese. Con la sua multiforme opera è stato in grado di descrivere e comprendere la tragicità di una delle



Camus, Albert, 1913-1960
ID: 133656

Opere

Altre forme del nome

- Camus, Albert, 1913-1960
- كامور, ألبرت, 1913-1960
- Camus, Albert, 1913-1960
- Camus, A. 1913-1960 Albert
- كامور, ألبرت
- Камю, А. 1913-1960 Альбер

...[altre forme]

Bibliografia

(Clicca per cercare su google)

- Actualités
- Actualités - English, Selections
- Adapté ha-rishon
- Albert Camus vous parle

...[altri titoli]

http://share-vde.org/sharevde/searchNames?n_cluster_id=133656

Figure 10: Albert Camus on the SHARE-VDE platform (http://www.share-vde.org/sharevde/searchNames?n_cluster_id=133656)



Vivaldi, Antonio, 1678-1741
ID: 37154

Questo autore in

- LIBRARY OF CONGRESS
- data.bnf.fr
- AF

Altre forme del nome

- Vivaldi, Antonio, 1678-1741
- 1678-1741, ڤيڤالدي, 1741
- Vivaldi, Antonio, 1678-1741
- Vivaldi, Antonio
- Vivaldi, Antonio, sac., 1678-1741
- Вивальди, А. 1678-1741, Антонио
- Вивальди, Антонио, 1678-1741
- Vivaldi, Antonio, 1680-1741
- 1741-1678 - ڤيڤالدي, 1678
- Antonio Vivaldi compositore e violinista italiano esponente di spicco del tardo barocco veneziano
- Vivaldi, Antonio, ca.1678-1741
- Vivaldi, Antonio (Italian composer and musician, 1678-1741)
- Prete rosso, 1678-1741
- Vivaldi, A., 1678-1741
- Vivaldi, A. (Antonio), 1678-1741
- Vivarudi, Antonio, 1678-1741
- Vivaldi, Antonio

...[altre forme]

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



Figure 11: Entities in cluster: an example of collaboration and sharing (http://www.share-vde.org/sharevde/searchNames?n_cluster_id=37154&l=en)

The result of a reconciliation of the entity *Antonio Vivaldi* in the Share VDE project, with data from different sources and projects:

- the authorized form from a local authority file
- the variant forms originating from the references on the local authority records
- the variant forms originating from the VIAF
- the forms of the name used in the bibliographic records.

The cluster is completed and enriched with identifiers for the same entity, Antonio Vivaldi, from sources such as:

- [Wikidata](#)
- [Library of Congress Name Authority File](#)
- [Data.bnf.fr](#)
- [VIAF](#)

Grouping under a single work title of the many publication titles in the catalogue for *Cimento dell'armonia e dell'invenzione*

Single work title

Brings together different publications/resources present in different catalogues.

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



Figure 12: An example of Work/Instances reconciliation (http://www.share-vde.org/sharevde/searchTitles?t_cluster_id=11287&l=en)

http://www.share-vde.org/sharevde/search?q=The+storm+and+other+things&&h=any_bc&s=10&o=scores&v=ll&dls=true&l=en



Figure 13: Example of same Instances reconciliation for titles present in different library catalogues (http://www.share.vde.org/sharevde/searchq=The+storm+and+other+things&&h=any_bc&s=10&o=scores&v=ll&dls=true&l=en)

3.6 Further connected topics

A number of further connected topics and enhancements are addressed throughout the project, among them:

- User interface:
 - For both Person/Work and Instance layers: search-box for relationship presentation (e.g. author to subject, author to publisher);
 - First edition identification for the chronological positioning of Entities;
 - “Work of” or “Work concerning” a Person.
- Additions of classes and properties originating from ontologies other than BIBFRAME as needed.
- Subject URI enrichment; content, media, carrier enrichment.
- Analysis for the creation of relationships among subject terms and strings in different languages.
- Provenance declarations that can become the fourth element added to every triple.
- Update management and URI Registry.

In order to address the many involved aspects particularly useful have been the brainstorming opportunities among Linked Data for Production (LD4P), IMLS Shareable Authorities Forum, Linked Open Data in Libraries Archives and Museums (LODLAM), Program for Cooperative Cataloguing (PCC) and European RDA Interest Group (EURIG) contacts and meetings.

4. Conclusions

The scope of the briefly described project has been chosen so that decisions for subsequent steps for a production scenario for the GLAM communities can be based on concrete evidence from a conspicuous data sets. This can then contribute to building a realistic model of the activities, the problems to be addressed as well as the potential advantages of moving toward the Linked Data environment. Functionality also dominated the design of the project: providing various environments and user interfaces for data creation, enrichment and supply workflows to the diverse group of participating librarians, professionals, scholars, researchers and students encompassing a wide range of needs.

Promoting a culture of openness towards knowledge has multiple advantages for all of the links in the information chain. The exploitation and diffusion of library, archive and museum data to a wider audience, enriching the World Wide Web with valuable information that until now has remained hidden in archives, collections and catalogues, allows libraries and museums to benefit from the opportunity to provide more comprehensive tools, while end users are able to access a wealth of information.

All research, but perhaps the Humanities in particular, needs now more than ever to be visible, available, accessible and innovative. By using the increased discoverability offered by the projects discussed, perceptions of research fields not foremost in the mind of the public at large can be changed and their importance for society acknowledged, also reducing the risk of niche subject areas being marginalised. For this to occur, increased access to knowledge is vital and key to achieving this is collaboration among all stakeholders. Linked Data practices give vital support to the stewardship of research and introduce an invaluable opportunity that can contribute to taking forward cultural heritage for future generations.

5. Resources and links

Bibliographic Framework Initiative homepage

<http://www.loc.gov/bibframe>

RDA Steering Committee (RSC)

<http://www.rda-rsc.org>

PCC Task Group on URIs in MARC homepage

<http://www.loc.gov/aba/pcc/bibframe/TaskGroups/URI-TaskGroup.html>

Linked Data for Production (LD4P) homepage

<http://wiki.duraspace.org/pages/viewpage.action?pageId=74515029>

IMLS Shareable Authorities Forum homepage

<http://confluence.cornell.edu/display/sharedauth/IMLS+Shareable+Authorities+Forum+Home>

Linked Open Data in Libraries Archives and Museums (LODLAM) homepage

<http://www.lodlam.net>

European RDA Interest Group - 2017 meeting conference material

<http://www.casalini.it/eurig2017>

SHARE-Virtual Discovery Environment and the Casalini experience and roadmap for supplying BIBFRAME data. [Presentation delivered at the Program for Cooperative Cataloguing Operations Committee Meeting, Library of Congress, May 5th 2017]

http://www.loc.gov/aba/pcc/documents/OpCo-2017/PCC-OpCo-2017_SHARE-VDE_Casalini-Possemato.pdf

SHARE Virtual Discovery Environment in Linked Data homepage

<http://www.share-vde.org>