Implications of BIBFRAME and Linked Data for Libraries and Publishers

Michele Casalini
Casalini Libri, michele@casalini.it

Michele Casalini, "Implications of BIBFRAME and Linked Data for Libraries and Publishers" (2016). *Proceedings of the Charleston Library Conference.*
http://dx.doi.org/10.5703/1288284316449

Implications of BIBFRAME and Linked Data for Libraries and Publishers

Michele Casalini, CEO, Casalini Libri

Abstract

This article considers the current situation of transition from the machine-readable cataloging (MARC) formats to the Bibliographic Framework Initiative (BIBFRAME) data model, and the further step to organize and publish catalog information in the emerging linked data technology. The definition and development of new tools to realize the required changes are discussed and an outline provided of the steps being taken by Casalini Libri to ensure the compliance of its bibliographical production and services with the new standards and offer assistance to libraries and publishers in their implementation.

BIBFRAME and Linked Data

The emerging BIBFRAME data model for the future evolvement of bibliographic formats is currently the subject of discussion and development within the library community. The new framework is intended to open up the possibilities of linked data to libraries, archives, and museums, providing greater visibility and discoverability for the resources they hold and making bibliographic information more flexible and accessible to end users across the web, rather than just library patrons. Many organizations are beginning to experiment with this framework and develop new workflow and business models to respond to the changing needs of the library world. This paper focuses on the efforts of Casalini Libri, a bibliographic agency, specialist vendor of European publications, and aggregator of digital content, to explore how BIBFRAME resources can be created and delivered in addition to traditional MARC services.

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, becoming one of the leading suppliers of European publications and related library services, specializing in monographs, series, and periodicals from Italy, France, Spain, Portugal, Belgium, Switzerland, Greece, Malta, and the Vatican City. A family-run business that is now in the hands of the children of its founder, Mario Casalini, Casalini Libri's 90-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 has always been the provision of quality bibliographic information, from details of relevant new titles to the in-house production of highly accurate catalog records. The company produces more than 40,000 original bibliographic records for romance language publications each year, all of which are accessible through the online I Libri database (www.ilibri.com). The records are created in native MARC21 according to the RDA BIBCO Standard Record (BSR) guidelines using the in-house WeCat cataloging module of the OLISuite ILS, developed by @Cult (www.atcult.it). The company's specialized cataloging team contributes new authority records to the national authority file and makes subject and classification proposals through participation in the Name Authority Cooperative Program (NACO) and Subjective Authority Cooperative Program (SACO) programs of the Program for Cooperative Cataloging (PCC).

A Changing World

Developed in the 1960s, by the early 1970s, MARC had become the international standard format for the encoding and exchange of bibliographic data, remaining so for several decades. Technological advancements over this time and the emergence of the World Wide Web, so dynamic in nature, have acted as the crucible for the re-examination of the MARC format. This has culminated in the realization of the necessity for a more flexible model for the recording and exploitation of data.

The first step in this process was the introduction of the Resource Description and Access (RDA) standard, initially released in 2010. Particularly appropriate for use by museums and archives, RDA replaces the

https://doi.org/10.5703/1288284316449

Anglo-American Cataloguing Rules, Second Edition (AACR2), providing 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.

The impact of these developments manifests itself in various aspects of library activity, not least in the need to adapt components of library systems to support the new format and update workflows both within libraries and with partners and suppliers. New challenges and necessities for libraries and librarians, linked directly to the ongoing advances, include the need for library data to be more visible, open to a higher level of interoperability with other systems, and the web as a whole, in order to support new co-operative institutional initiatives and provide concrete advantages for end users. 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 cannot easily be harnessed and linked to other, related resources. In a world that is ever more web-based, libraries risk marginalization if they cannot build links toward making their data more accessible; therefore, they must increase the discoverability of their resources.

New Tools and New Implementations

Aware of the value of the RDA model to the library community and the importance of supporting its own customers in the transition to a new concept of organizing and providing data, Casalini Libri successfully implemented a smooth transition to the RDA standard, with staff members receiving professional training from key community figures. Keen not only to respond to the changing requirements of libraries in this field, but to anticipate them whilst offering prompt and effective resources, the company has been actively involved in the translation of RDA into Italian and is highly engaged in the work of the European RDA Interest Group (EURIG), whose annual seminar and meeting will be hosted by Casalini Libri in 2017.

Casalini Libri is particularly interested in linked data and the opportunities created by this method of publishing structured data. Although the technology has not yet come of age, this is a unique and fascinating period in which standards, models, and guidelines are all in the process of definition and refinement. At the same time, individual tools are being developed, tested, and shaped to re-design the emerging information chain. Given the shifting organic nature of current developments, short-term projects are a particularly useful approach that can provide tangible results, giving real answers. This information will form the foundation upon which to base future plans. The activity of Casalini Libri in this field can be divided into four main areas of focus; the first three of these are currently in progress, while the fourth and final aspect is at present in the planning stage:

- The enrichment of MARC records with uniform resource identifier (URI) to simplify their conversion into linked data/BIBFRAME;
- The use of a framework to automate the conversion of data from MARC format to resource description framework (RDF), using the BIBFRAME model;
- The creation of a publication layer in FRBR/BIBFRAME for bibliographic and authority records in order to optimize the end user experience;
- The creation of original data directly in RDF with the possibility of retro-conversion into MARC. (Planned development, not yet in progress.)

MARC Enrichment

One aspect to be taken into consideration in the application of data models is the conversion of data into an alternative structure without the loss of content. To ensure an effective passage from the MARC record to BIBFRAME, the Casalini Libri implementation plan foresees that MARC data elements must be enriched through the addition of several local and global identifiers. Once the automatic and manual processes required for this procedure are established, the implemented module will allow MARC to be converted into linked data by any entity. Separate profiling options should be made available to handle the personalization of URI sources according to the preferences of single libraries.

An essential step toward MARC enrichment is the implementation of a URI management system to organize the identifiers for each heading access point. In the case of Casalini Libri, the URI

management system has been implemented directly into the WeCat cataloging module of the OLISuite ILS by @Cult, already integral to the company's cataloging workflows. This allows catalogers to check, validate, modify, delete, or add identifiers to a single heading.

There are various ways in which URIs can be used, for instance as an $0 subfield associated to an access point within the MARC bibliographic record, or as an RDF property of a person entity. In the data export and conversion process, it must be possible to define how many URIs are available for each heading, how to associate them to the heading, and how to show them in relation to the use and format of the data.

Casalini Libri is currently working to complete enhancements to its database to allow for URI to be embedded into headings. Also in progress is the development of additional APIs to facilitate interaction with the various sources and provide simple, agile tools for export that will accommodate different approaches to URI source profiling among libraries.

Automating MARC Conversion to RDF

The conversion of data from different formats to RDF occurs within WeCat thanks to embedded micro software agents that are specialized for each specific MARC data elements to convert and export the information contained as linked data. The same conversion process can be activated independently from alternative sources that use data in different formats such as MARXML and LIDO.

The framework used in this data conversion process is based on the automatic publication under Linked Data Paradigm of Library Data (ALIADA) project. The ALIADA project was co-financed by the European Union's Research and Innovation funding program for 2007–2013. It involved five partners from Italy, Spain, and Hungary: the Spanish TECNALIA Research Institute and ARTIUM Museum, the Museum of Fine Arts in Budapest, in collaboration with the IT companies @Cult and SCANBIT. Spanning 24 months, from November 2013 to October 2015, the project results are available from www.aliada-project.eu.

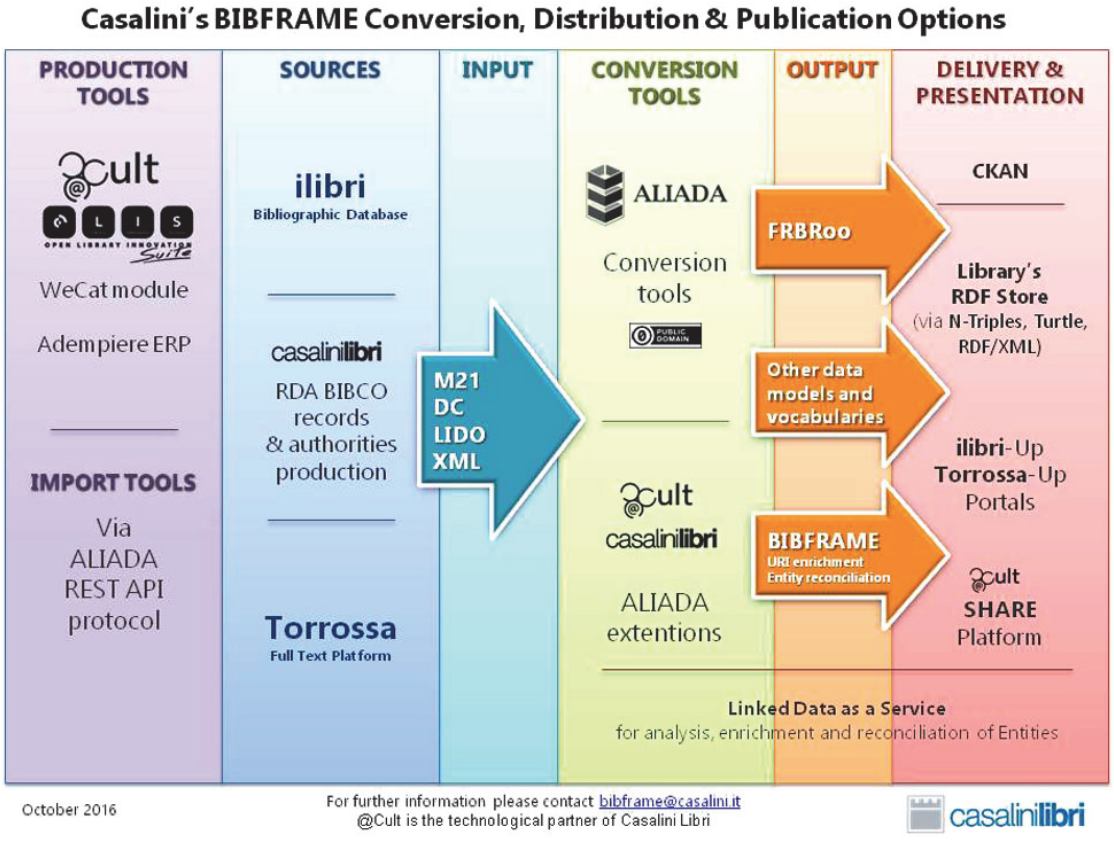

Figure 1. The scheme illustrates the conversion, distribution, and publication options of the described BIBFRAME projects, aimed at providing also linked data as a service.

ALIADA originally applied the linked data paradigm using FRBR-object oriented-based (FRBRoo) ontologies. The project was developed with the aim of supporting the entire process, from conversion to the publication and linking of data from public bodies such as museums and libraries through an open source plugin for management systems already in use. It makes an increased degree of interoperability possible among libraries and museums, allowing not only for the sharing of collection data but creating new opportunities for interaction with the general public. The tool provides a single point of access to the collection of datasets published through a representational state transfer (RESTful) interface, retrievable from multidevice platforms able to use the linked data to develop innovative applications. The main components of ALIADA are:

- A user interface designed specifically to facilitate the use of the tool by nonspecialist operators required to carry out administrative tasks, content selection, and publication;
- Conversion: Translation of the content selected in RDF datasets;
- Publication of RDF datasets that have been validated in the linked data cloud;
- Linking RDF datasets in the linked data cloud through common entities or concepts.

ALIADA's conversion and publication processes are built on the basis of an asynchronous pipeline, whereby each component is responsible for a fractional part of the overall task. Each processor can act as a splitter or aggregator and can achieve content manipulation of the incoming message.

Casalini Libri's implementation of ALIADA technology is well under way. Relying on the Casalini Libri and @Cult partnership an ALIADA extension for BIBFRAME was developed. Following the publication in May 2016 of the set of BIBFRAME 1.0 test data in N-Triples, Turtle and RDF/XML formats for the data produced in one month for six North American research libraries, BIBFRAME 2.0 test data was released in October 2016. The data is also available online in the Blazegraph database that supports SPARQL queries, at http://lod.casalini.it:9999/.

Creating an FRBR / BIBFRAME Publication Layer

Current catalog data predominantly contains descriptions of manifestations/instances. The objective is now to respond to the need to redesign this data model with a system that derives data from existing records to produce a new person/work layer.

The process creates for each person entity a cluster of possible variant forms and does the same for each associated work. This is a crucial step in the process that retrieves and reconciles data: A controlled name access point is created for person names and the cluster consequently associated to relevant works. In order to harness all variants, the system must interrogate external authority files, such as ISNI and VIAF, among others. Forms of data found in an authority record will have the greatest weight in the reconciliation algorithm. Each work cluster is then linked to the instance titles and, finally, to the items at the level of individual online public access catalogs (OPACs).

One concrete example of how this process can work in collaborative data use among libraries, museums and archives is the SHARE platform, created and developed by @Cult. SHARE is a common discovery layer based on linked data that applies the BIBFRAME data model to identification, reconciliation, and clustering processes. The first project, SHARE catalog, which was initiated in 2014 before going into production in spring 2016, involved seven Italian university libraries and is accessible at http://catalogo.share-cat.unina.it/sharecat/clusters?l=en. Data from the seven distinct catalogs were integrated with semantic web technology to construct a single portal for the shared use of catalog data throughout the authority and bibliographic entries of the participating universities, thereby facilitating interaction, exchange and the re-use of information from the ILS of each single institution to the librarian and the end user. Each library maintains its own local system and continues to work according to its own local rules and guidelines.

To reach this scope, data are converted according to the reference model indicated by the World Wide Web Consortium (W3C) standards for linked data, RDF, and structured in compliance with BIBFRAME. The data sets are subsequently enriched through connections to authority files and other external

sources before being published on the shared portal for consultation. The portal is, therefore, 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 catalogs 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.

Following the success of this project, May 2016 saw the creation of the enriched knowledge base I Libri-up, an enhancement of Casalini Libri's existing ilibri bibliographic database, which is currently under beta test.

In October of this year, plans were made for work on a new project, SHARE Virtual Discovery Environment in Linked Data, with a group of North American research libraries that aims to develop a prototype in the context of institutions with different systems, habits, and cataloging traditions. The project will create, in addition to the three BIBFRAME portal layer, a database of relationships and a common knowledge base of clusters. The realization of this initiative is based on the Casalini Libri and @Cult partnership.

Conclusions

All work on the potential exploitation and diffusion of library and museum data to a wider audience, enriching the World Wide Web with valuable information that until now has remained hidden in archives and catalogs, promotes a culture of openness toward knowledge that has multiple advantages for all components of the information chain. Libraries and museums benefit from the opportunity to provide more comprehensive tools, while end users can access a wealth of information. Publishers also benefit from faster and greater visibility for their content.

The shift to the RDA cataloguing standard is a first step in the transition process, helping libraries share information through the creation of more searchable data. Much exploration and experimentation is still necessary to decide the best ways of converting data into links and create tools that will facilitate the coexistence of BIBFRAME and MARC for a long period of transition. Casalini Libri, building on a solid foundation and experience as a cataloging agency, aims to work closely both with the library community and the developers to create services that will support the information chain in making a smooth transition to the new standards and technology.