IFLA WLIC 2018
KUALA LUMPUR

Submitted on: 13.07.2018

The Evolution of BIBFRAME: from MARC Surrogate to Web Conformant Data Model

Philip Schreur
Technical Services, Stanford University, Stanford, USA.
E-mail address: pschreur@stanford.edu

Abstract:

Initiated in 2011 by the Library of Congress, BIBFRAME is intended to become the international standard for encoding library metadata so as to make it understandable to the semantic web. The development process has been an evolutionary one, however, as the standard itself needed to be revised to meet best practices of the Resource Description Framework (RDF) and extended to encompass non-book materials such as performed music, art objects, or cartographic materials. BIBFRAME 2.0 has now been released and has been put into practice in both the United States, Europe, and Asia.

Keywords: linked data, BIBFRAME, MARC, LD4P.

Background

Ever since its inception in 1989, more and more of our world has been moving to the Web. Everything from social media to dating services, tax forms to government services, groceries to commercial services, find their home on the Web. It is not surprising, then, that our library patrons often look to the Web for their information services. However, the library's information ecosphere has its roots in an earlier time, the 1960s, and a form of communication called MARC, or, MAchine Readable Cataloging ("MARC Standards Homepage").

Libraries worldwide rely upon MAchine-Readable Cataloging (MARC)-based systems for the communication, storage, and expression of the majority of their bibliographic data. MARC, however, is a communication format developed in the 1960s to enable machine manipulation of the bibliographic data previously recorded on catalog cards. Connections between various data elements within a single catalog record, such as between a subject

heading and a specific work or a performer and piece performed, are not easily, and therefore not usually, expressed as it is assumed that a human being will be examining the record as a whole and making the associations between the elements for themselves. MARC itself was a great achievement, eliminating libraries' dependence on card catalogs and moving them into a much needed online environment. It allowed for the development of the Integrated Library System, or ILS, and great economy in the acquisition, cataloging, and discovery of library resources. However, as libraries transition to a linked-data based architecture that derives its power from extensive machine linking of individual data elements, this former reliance on human interpretation at the record level to make correct associations between individual data elements becomes a critical issue. Although MARC metadata can be converted to linked data, many human-inferred relationships are left unexpressed in the new environment. It is functional, but incomplete. With each day of routine processing, libraries add to the backlog of MARC data that they will want to convert and enhance as linked data.

In the last ten years, computer science has embraced the LOD pathway that demands more semantic expression of data (that supports machine inferencing). It has developed approaches to data and international standards that support the new environment in the form of the use of identifiers to link data and the international standard, Resource Description Framework ("RDF Homepage"), or RDF, for recording it. Redevelopment of the platform for expressing and communicating bibliographic data is needed to move libraries more firmly into the internet and web environment.

The development of the digital library, often based upon a digital repository, has further complicated the library environment. In addition to their MARC data, libraries have become curators of rapidly expanding collections of digital objects, data sets, and metadata in other schemas such as the Metadata Object Description Schema (MODS) ("MODS Homepage") or Dublin Core. These resources and their metadata are typically stored in digital repositories and become a parallel, yet separate, database of record. This lack of integration has caused great difficulties in consistency and maintenance as the long held concept of a single database of record has broken down. And even beyond these two repositories (the ILS and the Digital Repository), as academic libraries look to the future, they will be asked to step outside these more traditional materials to become the curators of the vast knowledge the university creates, in all its richness and diversity. Interactive scholarly works, unpublished data sets, information about faculty contained in profiling systems, metadata about learning objects, once integrated with more traditional library resources, will allow our faculty and students to explore our information resources and make associations that are impossible today.

In 2011, the Library of Congress (LC) began a project to end libraries isolation from the semantic web through the creation of a new communication format, called BIBFRAME "BIBFRAME Framework as a Web of Data," 2012), as a successor to the MARC formats. The development of BIBFRAME has been a complex one as its creators try to balance the need to capture the data encoded in MARC, the constraints of RDF, and input from the community it hopes to serve. In addition, there are other schemas available for libraries' use, such as Schema.org ("Schema.org Homepage"), the CIDOC Conceptual Reference Model (CIDOC-CRM) ("CIDOC_CRM Homepage"), and the Europeana Data Model (EDM) ("Europeana Data Model Documentation"). Although not designed as replacements for MARC, these other schemas are used by important information communities, such as Europeana ("Europeana Portal") or Museums, with which libraries interact. The resultant metadata ecosystem has created a very complex environment.

Schema.org itself deserves a special mention in this complex environment. Sponsored by Google, Microsoft, Yahoo, and Yandex, "Schema.org is a collaborative, community activity with a mission to create, maintain, and promote schemas for structured data on the Internet, on web pages, in email messages, and beyond." ("Schema.org Homepage") It has been designed for the broadest possible use and focuses upon the semantic understanding of Web search engines. Because of this focus, it is of great interest to libraries and library-related organizations, such as OCLC, for embedding library data into the semantic web. It was never designed, however, to capture even the full richness of the data contained in MARC. Rather, its focus is on broad integration into the Web. BIBFRAME has been designed to fill that gap so that, as libraries move to the semantic web, the richness and detail of their metadata can be reflected there.

Libraries have survived in their current environment by adhering to structural and data quality standards to facilitate the easy exchange of metadata for commonly held resources. These standards also allowed metadata from various institutions to be quickly combined into large discovery interfaces. As libraries transition from their current environment to a much more complex one based in LOD, these standards must be rethought and re-envisioned. Their need is still as strong but their expression is unclear.

Since its inception, BIBFRAME has been used in a number of individual projects both within the United States and internationally. For instance, the University College London Department of Information Studies has been awarded a grant to develop a Linked Open Data bibliographic dataset based on BIBFRAME ("Linked Open Bibliographic Data"). The Library of Alexandria will focus on the conversion process for data in the Arabic language ("The Standard Model "BIBFRAME" for Resource Description and Access in Web Environment"). The National Library of Medicine has experimented with more modular approach to the BIBFRAME vocabulary by paring down the existing vocabulary to its core concepts (BIBFRAME-Lite) ("NLM BIBFRAME Update").

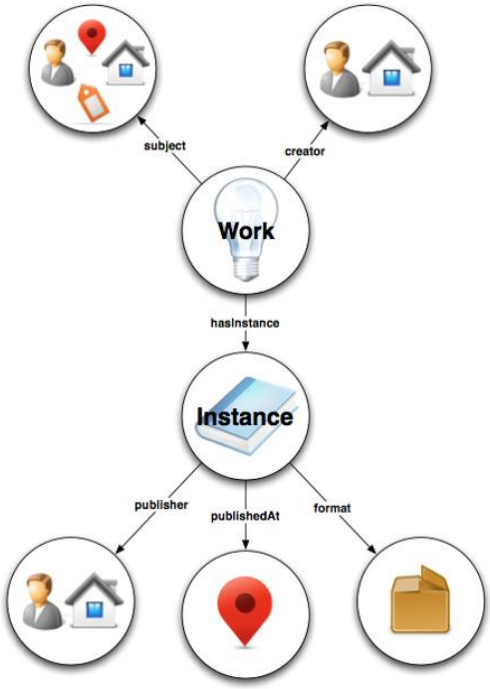
But why linked data specifically? It is apparent that library patrons have preferred searching for information on the Web for quite some time. By integrating library data into the Web in a semantic way, our patrons can find well-formed library data there as well as in library catalogs. By taking advantage of the semantic web, library patrons can directly benefit from other important data sources on the Web. A third advantage is that the Web is an international environment. By shifting to linked data, libraries worldwide can take advantage of the bibliographic and authoritative data many national libraries create and make available now as linked data. And last, the Web is a continually evolving environment. Without a doubt, linked data will evolve into some other standard with time. But in order to move along with this evolution, libraries will need to make that first important step in the transition to a Web environment.

History of BIBFRAME – BIBRAME 1.0

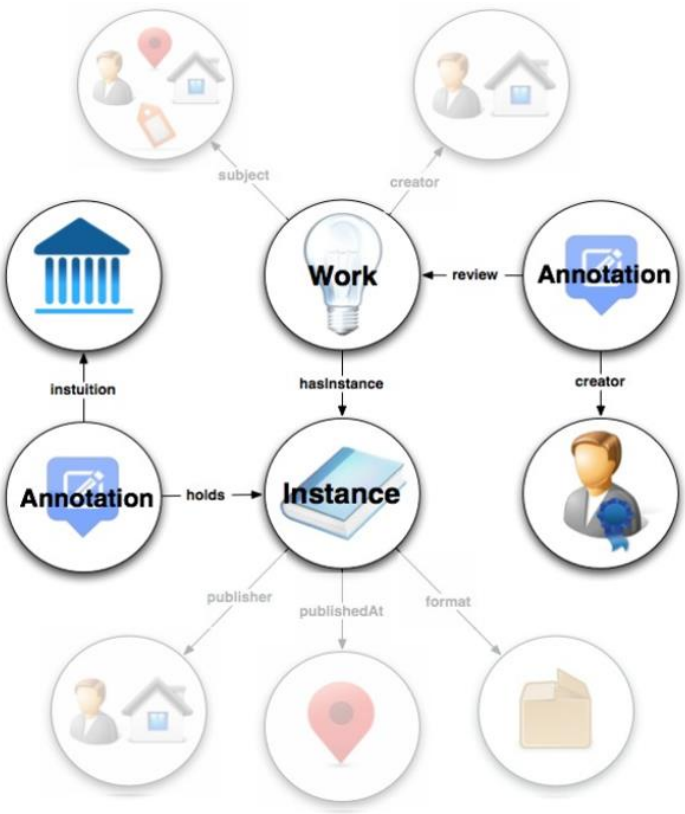
BIBFRAME, short for the Bibliographic Framework Initiative, was officially launched in May of 2011. It was developed in partnership between the Library of Congress and Eric Miller of Zepheira ("Zepheira Homepage") and was designed as a replacement for the MARC formats, that is, as a communication format. BIBFRAME "aims to re-envision and implement a new bibliographic environment for libraries that makes the network central and interconnectedness commonplace." It does this by allowing us both to transform our legacy data and to create new metadata as linked data, the language of the semantic web.

BIBFRAME is also designed to integrate with and engage in the wider information community that is developing on the Web while serving the very specific needs of its maintenance community, that is, libraries.

The original model itself (now BIBFRAME 1.0) was quite simple being divided into four classes: Works, Instances, Authorities, and Annotations.



The two key divisions in BIBFRAME 1.0 are Work and Instance. The Work is the conceptual resource that is being cataloged. It is an abstract entity composed of the FRBR ("Functional Requirements of Bibliographic Records") concepts of Work and Expression. The Instance is the material embodiment of a BIBFRAME work. Each BIBFRAME instance represents one and only one BIBFRAME work and is parallel to the FRBR Manifestation. Authorities are key concepts such as people, places, topics, or organizations that may be associated with a Work or Instance.



An Annotation represents concepts other than Authorities that can be associated with a Work or Instance. For instance, the review of a Work can be considered an Annotation to the Work and the fact that an institution holds a particular resource can be considered an Annotation of the Instance. The strength of this elegantly simple model is the ease with which it can be extended to cover the needs of more complex formats such as music or audio-visual materials.

As early adopters began experimenting with BIBFRAME, however, two key issues began to surface. The first related to the structure of the model itself. BIBFRAME was conceived of as a neutral communication format, that is, it was meant to transmit a variety of information, not just library metadata. One of the concerns with the MARC formats was that they were used exclusively by libraries and understood only by them. By making BIBFRAME a more neutral communication format, it was hoped that this problem could be avoided in the future. However, this meant that the model itself did not reflect the FRBR-like nature of many library resources as the FRBR Work and Expression were combined in the BIBFRAME Work and the FRBR Item was represented as one of many Annotations possible to a BIBFRAME Instance. It was not clear, then, how metadata created according to the cataloging standard Resource Description and Access (RDA) ("RDA Toolkit Homepage") could best be represented in BIBFRAME.

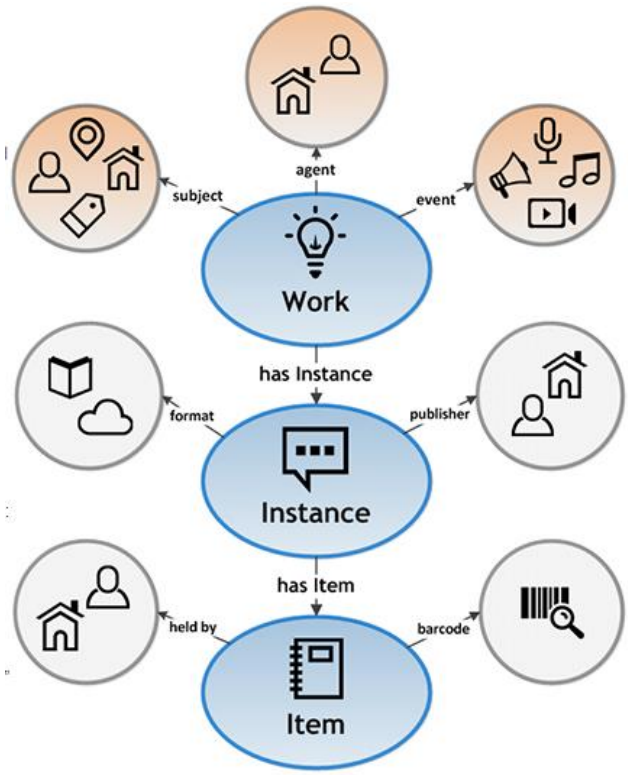
A second issue concerned how best to represent the enormous wealth of detail contained in the MARC formats within BIBFRAME. There is a wonderful quote by Sally McCallum of the Library if Congress saying that if MARC ever died it would be from obesity. As more and more of the detailed information encoded within MARC was mapped into BIBFRAME, the once elegantly simple model became muddled and began to stray from the best practices of RDF, the W3C schema that provides the data-modelling vocabulary.

BIBFRAME 2.0

In 2014, Cornell successfully applied to the Mellon Foundation for support for a project called Linked Data for Libraries ("Linked Data for Libraries Homepage"), or LD4L. This project was a collaboration between three institutions (Cornell University, Harvard University and Stanford University) with a goal to create a Scholarly Resource Semantic Information Store (SRSIS) model to capture the intellectual value that librarians and other domain experts and scholars add to information resources when they describe, annotate, organize, select, and use those resources, together with the social value evident from patterns of usage. As part of this grant, the partners agreed to use the BIBFRAME ontology as the model for the bibliographic data from their collections. As the project included the conversion of all MARC bibliographic data from the libraries holdings, it became one of the first tests of BIBFRAME at scale.

The mass conversion of MARC data, and the desire to use it in a linked data context, exposed many of the shortcomings in the development of BIBFRAME 1.0. In response to this, the Office of Network Standards of the Library of Congress commissioned a review of BIBFRAME 1.0 by Robert Sanderson (Sanderson, 2015). Mr. Sanderson, along with his team of reviewers, examined BIBFRAME 1.0 in the context of current best practices within the Linked Data domain. The report resulted in a series of recommendations to the Library of Congress attempting to preserve the value and semantics of existing data in the new ontology while balancing current best practices against the completeness of transformation of historical data.

The recommendations in the report cover a wide variety of issues such as the merging of predicates, the use of authorities, the structure of annotations, the representation of holdings, the role of events, and many others. The Library of Congress accepted many of the recommendations in the Sanderson report in its release of BIBFRAME 2.0.



The most obvious changes in BIBFRAME 2.0 from the earlier BIBFRAME 1.0 model is the inclusion of Item to better reflect the FRBR model, and the elimination of the Authority and Annotation classes. Other changes included fewer data properties through reification of data elements, as well as a conceptual restructuring of events and contributions.

In 2016, the Library of Congress became a partner in a new Mellon grant proposed by Stanford University called Linked Data for Production ("Linked Data for Production Homepage") (LD4P). LD4P is a collaboration between six institutions (Columbia, Cornell, Harvard, Library of Congress, Princeton, and Stanford University) to begin the transition of technical services production workflows to ones based in Linked Open Data (LOD). This first phase of the transition focuses on the development of the ability to produce metadata as LOD communally, the enhancement of the BIBFRAME ontology ("BIBFRAME Homepage") to encompass the multiple resource formats that academic libraries must process, and the engagement of the broader academic library community to ensure a sustainable and extensible environment.

LD4P has engaged with BIBFRAME 2.0 in two main areas. The first is the extension of the ontology itself. BIBFRAME should be seen as a core ontology for describing the most common library resources. It is meant to be simple and easy to use. Specialized domains, however, have special needs and these needs need to be expressed as extensions to the basic BIBFRAME ontology. LD4P worked to extend BIBFRAME 2.0 in four key areas: art, rare books, performed music, and cartographic resources. Although many libraries hold art objects, and many museums hold books, the ontologies used to describe these materials are quite different. The art extension working group looked to see if and how BIBFRAME could be enhanced to better described objects within library collections. The rare materials working group focused on extending BIBFRAME at the Item level to better cover issues such as provenance and binding. The performed music working group focused on issues such as extensions of medium of performance, work, and event modeling for musical recordings. The cartographic working group examined extensions needed to cover printed maps, atlases, and geospatial data sets.

A more fundamental engagement involved the creation of a derivation of BIBFRAME called bibliotek-o ("Biblioitek-o"). Bibliotek-o is both an extension and a deviation from BIBFRAME 2.0 in key modelling areas such as activities; content type, carrier type, and media type; identifiers; titles; and object versus datatype properties, among others. It is meant more as an opportunity to reflect on BIBFRAME's structure than as a foil to BIBFRAME. By creating a concrete alternative to key structural BIBFRAME elements, bibliotek-o allows for a practical discussion of alternatives.

The Future of BIBFRAME

The first European BIBFRAME Workshop was held at the Deutsche Nationalbibliothek in September of 2017. Its aim was to create a forum for the library communities in Europe concerning the implementation of BIBFRAME, the exchange of knowledge about current projects, and to create an open dialogue with members of the Library of Congress. The hope of the group is that BIBFRAME might become an international standard for the interchange of bibliographic metadata with the potential of becoming more global than MARC21. In

order for this implementation to take place, however, there needs to be more awareness of BIBFRAME itself and the opening of BIBFRAME to become a more community-driven, international standard.

The conference was attended by representatives from twenty-four institutions representing sixteen European countries plus the United States. Presentations were made on such topics as BIBFRAME considerations from an RDA-implementation perspective, preserving bibliographic relationships in mapping from FRBR to BIBFRAME 2.0, and BIBFRAME at Springer Nature ("Documents and Results of the European BIBFRAME Workshop 2017"). Breakout sessions included areas such as barriers for implementation of BIBFRAME, community building, vendor support, and documentation. Members at the workshop were very encouraged by the discussions that took place. It was clear from the many lightening talks presented that there was a good deal of BIBFRAME experimentation going on in Europe at the current time. The group agreed to meet in Florence in September of 2018 to continue the discussions.

Similarly, the second phase of LD4P, called Linked Data for Production: Pathway to Implementation, will begin in July of 2018. This second phase of LD4P builds upon the foundational work of LD4P Phase 1 to begin the implementation of the cataloging community's shift to linked data for the creation and manipulation of their metadata. A collaborative project among four institutions (Cornell, Harvard, Stanford, and the University of Iowa) and the Program for Cooperative Cataloging (PCC), this phase of LD4P has seven goals: the creation of a continuously fed pool of linked data expressed in BIBFRAME; development of a cloud-based sandbox editing environment to create and reuse linked data; the development of policies, techniques and workflows for the automated enhancement of MARC data with identifiers to make its conversion to linked data as clean as possible; the development of policies, techniques, and workflows for the creation and reuse of linked data and its supporting identifiers as libraries' core metadata; better integration of library metadata and identifiers with the Web through collaboration with Wikidata; the enhancement of a widely-adopted library discovery environment (Blacklight) with linked-data based discovery techniques; and the orchestration of continued community collaboration through the development of an organizational framework called LD4. Collaboration will be key in this phase of LD4P. No successful implementation of linked data can be accomplished without the cooperation of the large partnership of libraries represented by the PCC. And as the Web is a truly international environment, this partnership should include representatives from libraries worldwide, including those represented at the European BIBFRAME Workshop.

One last step in the evolution of BIBFRAME is its opening to the world in a first step to become a truly community supported and maintained international standard. Recently, on the BIBFRAME discussion list, Sally McCallum announced:

> The Library of Congress is beginning a process for maintaining the BIBFRAME ontology that will enable implementers to suggest corrections and changes. Various types of ontology adjustments can be suggested via the Issues tab in the github repository for the ontology https://github.com/lcnetdev/bibframe-ontology . Adjustments may be typos, definition adjustments, or issues that may affect the semantics of the ontology. Semantic adjustment suggestions should originate with implementers and describe their BIBFRAME experience with the issue. The Library of Congress will review the changes. Those with semantic

impact will be vetted in collaboration with other implementers who have large systems that use the BIBFRAME ontology (McCallum, 2018).

This announcement represents a great step forward in the evolution of BIBFRAME. As BIBFRAME evolves to become both more conformant to linked data best practices and more responsive to the community it serves, it fulfils its promise to become a truly international standard for the replacement of the MARC formats.

References

BIBFRAME Homepage, http://www.loc.gov/bibframe/, last accessed 2018/6/14.

Bibliographic Framework as a Web of Data (2012). Washington, D.C.: The Library of Congress.

Bibliotek-o, https://wiki.duraspace.org/display/LD4P/bibliotek-o, last accessed 2018/6/14.

CIDOC-CRM Homepage, http://www.cidoc-crm.org/, last accessed 2018/6/14.

Documents and Results of the European BIBFRAME Workshop 2017, https://wiki.dnb.de/display/EBW/Documents+and+Results, last accessed 2018/6/14.

Europeana Data Model Documentation, http://pro.europeana.eu/page/edm-documentation, last accessed 2018/6/14.

Europeana Portal, http://www.europeana.eu/portal/, last accessed 2018/6/14.

Functional Requirements for Bibliographic Records, https://www.ifla.org/publications/functional-requirements-for-bibliographic-records, last accessed 2018/6/14.

Linked Data for Libraries Homepage, https://www.ld4l.org/ld4l-2014/overview, last accessed 2018/6/14.

Linked Data for Production Homepage, https://wiki.duraspace.org/pages/viewpage.action?pageId=74515029, last accessed 2018/6/14.

McCallum, Sally. "[BIBFRAME] BIBFRAME Consultation." Message to the BIBFRAME LISTSERV. 19 April 2018. E-mail.

Linked Open Bibliographic Data, https://www.ucl.ac.uk/dis/research/collaborativeprojects/lobd, last accessed 2018/6/14.

MARC Standards Homepage, http://www.loc.gov/marc/, last accessed 2018/6/14.

MODS Homepage, http://www.loc.gov/standards/mods/, last accessed 2018/6/14.

NLM BIBFRAME Update, https://www.nlm.nih.gov/pubs/techbull/mj15/mj15_bibframe.html, last accessed 2018/6/14.

RDA Toolkit Homepage, http://www.rdatoolkit.org/, last accessed 2018/6/14.

RDF Homepage, http://www.w3.org/RDF/, last accessed 2018/6/14.

Sanderson, Rob (2015). Analysis of the BIBFRAME ontology for Linked Data Best Practices, https://docs.google.com/document/d/1dIy-FgQsH67Ay0T0O0ulhyRiKjpf_I0AVQ9v8FLmPNo/edit#heading=h.310o1a8282cm, last accessed 2018/6/14.

Schema.org Homepage, https://schema.org/, last accessed 2018/6/14.

The Standard Model "BIBFRAME" for Resource Description and Access in Web Environment, http://www.cybrarians.org/files/bibframe/rania_osman.pdf, last accessed 2018/6/14.

Zepheira Homepage, https://zepheira.com/, last accessed 2018/6/14.