

Collocation and Hubs. Fundamental and New Version

Sally H. McCallum^(a)

a) Library of Congress, <http://orcid.org/0000-0002-6137-2129>

Contact: Sally H. McCallum, smcc@loc.gov

Received: 3 June 2021; **Accepted:** 11 July 2021; **First Published:** 15 January 2022

ABSTRACT

This paper discusses collocation as a fundamental concept of metadata description that is reinterpreted and expanded in the BIBFRAME library linked data environment via the development of “hubs”. With the MARC title authority description as a basis, the relationships that support broader collocation are examined and the affinity of the MARC title authority to a bibliographic entity is explained. The reinterpretation of the title authority as a bibliographic hub will assist the fluidity needed in today’s environment between the MARC format, used for the last 50 years, and the new BIBFRAME ontology intended to replace it for richer linked data applications.

KEYWORDS

Metadata; BIBFRAME; Library linked data; MARC.

Collocation

Collocation of information items has been a primary purpose of rules for bibliographic descriptions for a very long time. It was stated by Cutter in 1889 (Cutter 1889), well-articulated by Seymour Lubetzky in the 1950s, and then reaffirmed and refined by the Paris Principles in 1961 (Lubetzky 1963). The traditional library collocation is attained by clustering item descriptions by agent names (e.g., authors) and titles – enabling this collocation is a major contribution of the Library cataloger. These clusters are, of course, done by indexing – in the past via the card catalog, but now via machine. Authors' names may vary, work titles may vary, and work content may vary but bringing together descriptions using different criteria gives the end user the ability to find the most useful resources for their needs.

Authority files were developed to support the clustering function and they work well for names (agents), even though much can be debated (and is) about categories of names – persons, corporations, families, conferences, real, imaginary, animals, spirits, etc. They can even be distilled to what is recently called “real world objects”. Either character strings (labels) or identifiers can be associated with them so they can serve the purpose of collocation of an agent's corpus and enable end users to find content more easily.

Titles are more difficult as the precise content associated with title strings is problematic to equate. The library profession has tried to apply the names model to titles to achieve collocation of content and has worked to establish unique labels that are associated with all items having the same content. These are the uniform titles of AACR2 (Anglo-American Cataloguing Rules 1978) and earlier cataloging rules and they were entered into name authority records augmented for titles – where additional data included alternate labels (i.e., references) for the uniform title. These title authority records do not contain descriptions of the contents the titles represent, but leave that to the bibliographic records for the resources. They do, however, contain title character strings or identifiers, like name authorities, and enough information to perform the same clustering or collocation functions as names do.

With the development of FRBR (Functional Requirements for Bibliographic Records 1998), however, a very close look was taken at the data in a bibliographic description to sort out data that could be associated with the conceptual work, the expression of the work, the manifestation of the work/expression, and the item. This dissection of description has been valuable to increase understanding of the bibliographic description, even though strict designation of data elements to work, expression, manifestation, and item does not hold up with the variety found among bibliographic resources – different media, editions over time, uniqueness of expression, rareness, etc.

The FRBR work concept and the authority file uniform title need to be reconciled for a future that can employ the new analysis in a useful way. This has led to an attempt to make the title authority record in MARC (MARC 21 Formats 2020) a FRBR work record; and an attempt, initially, to literally follow FRBR (as contained in RDA 2010) in BIBFRAME (BIBFRAME, n.d.)¹. In both cases adjustment had to be made to enable fluidity between MARC and BIBFRAME.

¹ BIBFRAME is a data model and ontology for bibliographic description. It is designed to replace the MARC standards, and to use linked data principles to make bibliographic data more useful both within and outside the library community.

MARC Title Authorities

The MARC Authority format (MARC 21 2021) was developed (and has been used for over 40 years) to establish and share authoritative labels for names that could be used across a file to enable collocation of resources associated with the name – creative contributions by the named agent or a subject association of the named agent. MARC authorities focused on including alternative forms of the label (MARC 4XX fields). The ideal is/was that every name used in a bibliographic description would be represented by an authority record and that form was to be used in the bibliographic records for access points.

The authority record concept was also extended to titles. The authority records for titles are different and more complex than those for names. Also title authority records are not made for all titles in a file so they share collocation duties with MARC 245 titles on bibliographic records. Title authority records are usually made when references are needed (1-4 below) or the cataloger wants to add cataloger research information (5-6 below). Title authority records are made for the following special situations:

1. When there are likely to be multiple bibliographic resources that are judged to have the same content and different titles.
2. When there are variations in a title authority label. These may be the title in other languages or scripts, or other editions, for example.
3. When there are joint creators or other related agents. The title authority records them as “alternative titles”.
4. When catalogers needs to record related titles that have a special association with the authorized title.

In addition, over time notes were added to record:

5. Supporting information for the formulation of the title label.
6. General notes about the title.

At the Library of Congress, title authority records are also generally made for titles for which the Library does not hold the resource but the title is needed in a MARC bibliographic record as an added entry or as a subject. Since the Library of Congress does not have the related resource, there is no bibliographic record for it in the Library of Congress files so the MARC title authority record is a stand-in for a MARC bibliographic record for the related title.

With these “rules” for when a title authority is made, only a small number of title authority records are made. At the Library of Congress while there are over 21 million titles in the bibliographic file, there are only 1.5 million title authority records. It should be noted that title authority records are not made for many cases where a relationship is expressed by a simple added entry. In those cases the bibliographic record serves the authority record role.

Recently attempts have been made in the community to make the MARC title authority serve as a FRBR/RDA work record, which has resulted in proposals to add many elements from the MARC bibliographic format to the MARC authority format to accommodate the additional FRBR work

elements – effectively making the MARC authority an authority/bibliographic record. This is not easy to do, however, as the tag groups in the MARC authority format are not compatible with those in the MARC bibliographic format.

The Library of Congress undertook an internal study in 2018 to map the MARC title authority record elements used for title authorities to a MARC bibliographic record to see if it was feasible and less disruptive to simply use the MARC bibliographic format for the title authoritative label records. This would have the advantage of enabling libraries to use additional elements for the bibliographic description of a work if an institution wants to add them, rather than using inappropriate fields in the MARC authority format for the data. It would also avoid a massive undertaking to add the missing elements to the MARC authority format. The study found a good fit for the title authorities with only a few adjustments.

The Library of Congress could also see that this would enable a more fluid transformation between formats – with, of course, BIBFRAME being a primary consideration.

BIBFRAME Hubs

When the first pilot for BIBFRAME began at the Library of Congress an attempt was made to use the FRBR/RDA model. BIBFRAME took a slightly simplified approach to FRBR and combined work and expression. The FRBR manifestation was called an “instance” to keep it from being mistaken for equivalence to a FRBR manifestation, although the two were closely aligned. While simplified, the BIBFRAME work/expression and instance shared many of the characteristics of the FRBR/RDA model entities. The Library of Congress began testing this RDF-based ontology with a pilot program, Pilot 1.

Sorting data elements and collecting relationships

For Pilot 1 an attempt was made to identify the data elements in MARC bibliographic records that FRBR/RDA associated with a work/expression and those it associated with an instance. When converting MARC records to BIBFRAME descriptions this allocation of data was made by machine. However, “well curated” as Library of Congress data is it has a long history that includes different sets of cataloging guidelines (ALA, AACR, AACR2, RDA to name a few)², community practices, and internal Library of Congress policies that affected consistency across a file of 21 million records. Those records describe resources from text to maps, audio-visuals, music, and still images – in print and various electronic forms. The files of records have been continuously added to for the last half century – with large numbers of records being added from retrospective conversion of catalog cards carried out 40 years ago using minimal record guidelines and then massaged in various projects to improve them.

² The primary rules used by the Library of Congress since 1908 include: *Catalog Rules: Author and Title Entries*, 1908; *American Library Association rules: A.L.A. Cataloging Rules for Author and Title Entries*, 1949 (ALA); *Library of Congress rules: Rules for Descriptive Cataloging in the Library of Congress*, 1949; *Anglo-American Cataloguing Rules*, 1967 (AACR); *Anglo-American Cataloguing Rules, 2nd ed.*, 1978 (AACR2); *Resource Description and Access*, 2010 (RDA).

Yet, the BIBFRAME system had to rely heavily on label matching to establish relationships and identify the proper URIs for data found in the MARC record. The system exploited some relationships that originated in the MARC bibliographic linking entries in the MARC 76X-78X, that sometimes have slightly more data to identify links. Many others came from added entries in MARC 700-740. And, of course, the prime MARC links in bibliographic records, the MARC 130 and 240 uniform titles were used. Series entries in the MARC 800-830 produced additional relationships between bibliographic resource descriptions as did 6XX subject entries. These relationships created collocation in the catalog so they were a key focus in the conversion to BIBFRAME. The relationships were collected into “hubs” and it was quickly realized that the hub provided additional power to the BIBFRAME file in support of collocation.

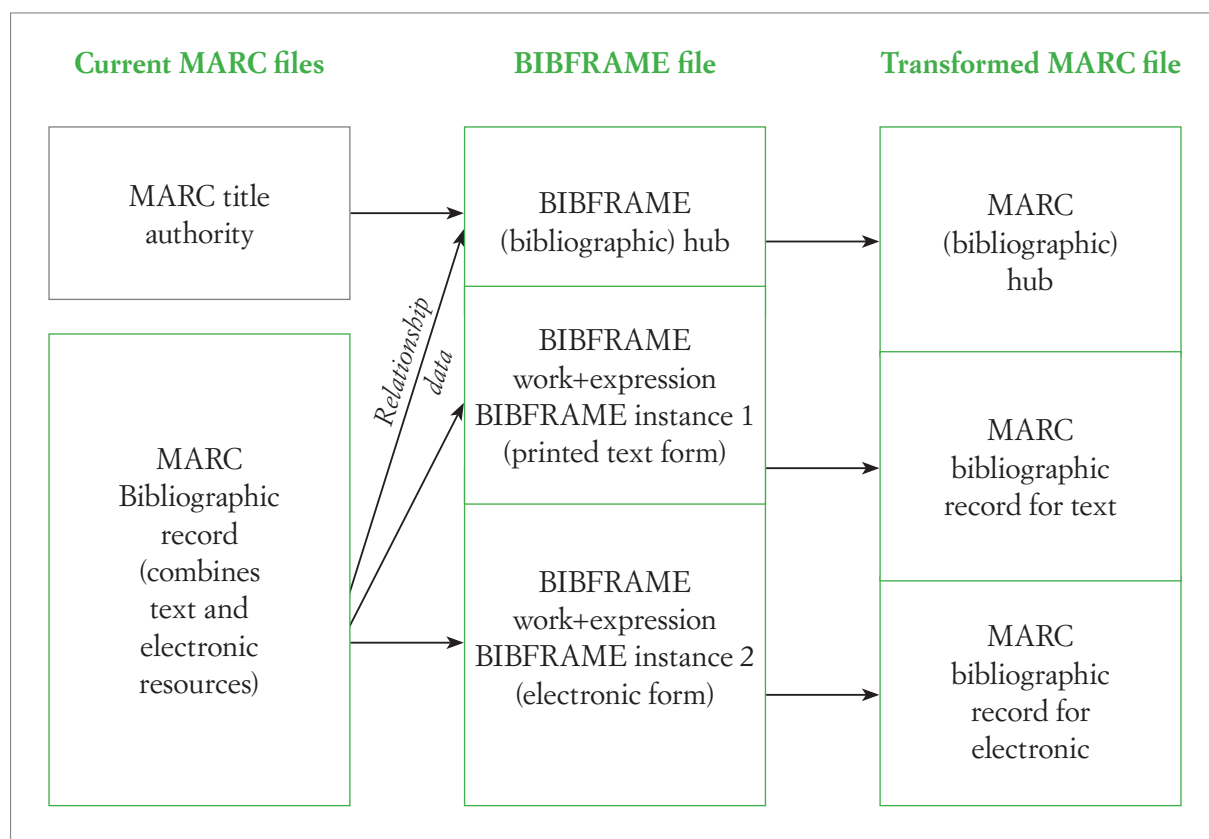


Fig. 1. Current MARC files, BIBFRAME file, transformed MARC file

Despite this exploratory effort creating hubs, Pilot 1 focused on merging, or trying to merge like bibliographic descriptions, or records, when the same resource was described. A difficult aspect of this merging was bringing together subject headings when multiple MARC bibliographic records merged to create one BIBFRAME work description. The subjects were considered part of the work description according to the FRBR/RDA model, not instance properties. Thus, when several MARC records collapsed into one BIBFRAME work, an attempt was made to reconcile the subjects. The merging of subjects proved to be especially difficult.

Pilot 2 and Hubs

So, when the Library of Congress started its second pilot, Pilot 2, it was based on lessons learned in Pilot 1 (BIBFRAME 2016) which included augmentation of the BIBFRAME ontology to better reflect aspects of RDA. But more importantly the project moved to a more realistic model that used “hubs” for collocation based on experience from Pilot 1, allowing the pilot to realize or take advantage of the collocation that had been provided in the MARC environment with the title authority records. The MARC title authorities were converted to BIBFRAME bibliographic work descriptions and called hubs, providing a solid foundation for hubs. Those 1.5 million hubs were then added to when hubs and relationships were created from the MARC bibliographic records as described above, bringing the total to more than 2.3 million. The BIBFRAME hub is a BIBFRAME bibliographic entity, not an authority description, and our current direction is – starting from the point of view of a BIBFRAME hub – to align the BIBFRAME hub with the MARC bibliographic format, not the MARC authority format as has been library practice. Our work with hubs has clarified a long-standing issue: the title authority is really a bibliographic record in authority clothing! This is a step toward the fluidity needed between BIBFRAME and MARC.

The expanded hub contains data that would have resided in a MARC title authority. It contains the title variations, author/title labels when there are multiple creators, and cataloger notes that support the hub content. And it has some characteristics of a work description. But it will not contain subject information allocated to the FRBR work, which will remain in the BIBFRAME work description, thus avoiding the merger issue. However, the BIBFRAME bibliographic ontology that is used for hubs can easily support further development of the hub description. Because of their similarity to a BIBFRAME work, currently hub descriptions are being expressed as BIBFRAME works with a special “rdf:type” of hub, which will allow the extension of hub content as needed to include new differentiating elements.

Hubs function as authoritative resources designed to serve as a common denominator, control point, and collocation mechanism, but that is not to say that they are “authorities” and should live separately from the larger bibliographic file. That is what happens now in the MARC files because the format it resides in is the MARC authority format. The format of its storage has dictated how they are seen and where they live. What is being proposed here is not to make these resources any less authoritative and representative than they are today, but to merge them with like data – all bibliographic – to improve their efficacy. The association of the hub with the bibliographic concept is working well thus far in the BIBFRAME environment.

As catalogers can originate more descriptions in BIBFRAME, the hub concept no doubt will continue to develop, but that development will be in a new environment that understands and exploits linking.

Hubs, SuperWorks, Opuses

There are currently several major projects carrying out extensive implementations of BIBFRAME in an Open Linked Data environment. Several have realized similar needs to those the Library of Congress sought with its hubs. Prominent among the projects is one called Share Virtual Discovery Environment (Share-VDE), a collaborative endeavor of the international bibliographic agency Casalini Libri and @CULT, together with library groups in the United States, Canada and Europe (Share-VDE, n.d.). Share-VDE uses a concept similar to the hub, which they call the “Opus”. Another is the University of Alberta’s LD4P project³ where the concept was also given the name “opus”. It is meaningful that several projects in the linked data space wrestling with the same problems have developed more or less the same solution.

Going Forward

This paper has discussed some fundamental concepts in bibliographic control in relation to widespread practices in bibliographic description. As the bibliographic environment shifts to take increasing advantage of linked data opportunities, flexibility and fluidity are going to be important. Movement between system environments rooted in MARC and those based in BIBFRAME are essential so narrowing selected differences are important. Discussion will be needed for the community to shift MARC title authorities to MARC bibliographic hubs in synch with BIBFRAME hubs, but in keeping with its commitment to cooperation in the bibliographic world the Library of Congress will pursue that discussion.

³ The University of Alberta is a cohort in the Linked Data for Production (LD4P) project. LD4P is a family of successive grant funded (Mellon) projects that provided foundational work and continued with implementation phases in support of the library cataloging community’s shift to linked data for the creation and manipulation of their metadata.

References

Anglo-American Cataloguing Rules. 1978. 2nd ed. Chicago: ALA.

BIBFRAME. 2016. "BIBFRAME Pilot (Phase One—Sept. 8, 2015 - March 31, 2016): Report and Assessment." <https://www.loc.gov/bibframe/docs/pdf/bibframe-pilot-phase1-analysis.pdf>.

BIBFRAME. n.d. "Bibliographic Framework Initiative." Accessed 29 November 2021. <https://www.loc.gov/bibframe/>.

Cutter, Charles A. 1889. *Rules for a Dictionary Catalogue*. 2nd ed., with corrections and additions. Washington, D.C.: Government Printing Office.

Functional Requirements for Bibliographic Records. Final Report. 1998. Munich: K.G. Saur. <https://www.ifla.org/publications/functional-requirements-for-bibliographic-records>.

Lubetzky, Seymour. 1963. "The Function of the Main Entry in the Alphabetical Catalogue: One Approach." In *International Conference on Cataloguing Principles, Paris, 9th-18th October, 1961. Report*. London: International Federation of Library Associations. 139-143.

"MARC 21 Formats." 2020. Washington: Library of Congress. Last modified 13 March, 2020. <https://www.loc.gov/marc/>.

"MARC 21 Format for Authority Data." 2021. Washington: Library of Congress. Last modified 24 November, 2021. <https://www.loc.gov/marc/authority>.

RDA: resource description and access. 2010 Chicago: American Library Association.

Share-VDE. n.d. "Share-VDE virtual discovery environment." Accessed 29 November 2021. <https://share-vde.org/sharevde/info.vm>.