

## **BIBFRAME Pilot (Phase One—Sept. 8, 2015 – March 31, 2016): Report and Assessment**

### ***Acquisitions & Bibliographic Access Directorate (ABA)***

#### ***Library of Congress***

***(June 16, 2016)***

The Network Development & MARC Standards Office (NDMSO) and the Cooperative & Instructional Programs Division (COIN) of the Acquisitions & Bibliographic Access Directorate planned and developed a pilot to test the efficacy of BIBFRAME, the Library of Congress's approach to linked data. After some time developing the BIBFRAME model and vocabulary, the ABA director was insistent that a pilot be launched no later than the beginning of fiscal 2016, i.e., October 2015. It was understood that full functionality of the BIBFRAME system would not be present. Much work was done by staff in both divisions to prepare for the Pilot. What follows is the assessment of that planning process, what was being tested, the results, and lessons learned that will assist the Library of Congress as it progresses the BIBFRAME model and vocabulary. This first Pilot was conducted in the BIBFRAME 1.0 environment. The next phase of the Pilot will be carried out in the BIBFRAME 2.0 environment, building on what was learned during the Phase One Pilot.

Approximately forty staff were identified for the Pilot. Participants were selected from among the ABA divisions and from the Collections & Services special format divisions. The chosen participants were a mix of catalogers and technicians that catalog materials in all languages, scripts and formats. The participants are responsible for processing monographs, serials, cartographic materials, music (notated), sound recordings, moving image, and two-dimensional art (prints and photographs). Participants were to process the materials they regularly receive. Because the Library must still distribute its cataloging output to its cataloging distribution service subscribers, participants were required to catalog in both the MARC 21 format and BIBFRAME. This requirement for dual data creation affected the participants' normal production. As a result there was no attempt to address the impact of BIBFRAME on production.

This report is divided into two segments, prepared respectively by COIN and NDMSO. An Appendix is included prepared by COIN that addresses training in greater detail. —*Beacher Wiggins, ABA Director*

### **PREPARATION AND TRAINING**

*(Cooperative & Instructional Programs Division—Judith Cannan, chief)*

#### **Training**

The BIBFRAME Pilot participants were true pioneers, embarking on an endeavor very few had explored and working in a system that was still under development with limited functionality.

Between late spring and September 2015, the participants attended sixteen hours of instruction on the Semantic Web, Linked Data and use of the BIBFRAME Editor. COIN staff members provided the training. The materials were designed, developed and taught by Tim Carlton (senior instructor), and Paul Frank, Les Hawkins and Hien Nguyen who form part of the Program for Cooperative Cataloging secretariat at LC. Paul Frank was also instrumental with developing profiles for the BIBFRAME Editor and designing the Editor interface.

The training materials are available from the Cataloger's Learning Workshop website <http://www.loc.gov/catworkshop/bibframe/>

The participants began using the Editor immediately after their training in its use. They entered data into both the LC ILS (Voyager) and the BIBFRAME Editor, creating MARC records in LC ILS first. Weekly 'de-briefings' were held to help the participants, instructors, and developers learn from each other about 'best practices' and what did or did not work. Participants were encouraged to make suggestions based on their experiences and many of these suggestions were implemented right away. Eventually, more formal 'refresher' training was held in December 2015 and January 2016 and the participants were instructed to enter data into the BIBFRAME Editor and then create a MARC record in LC ILS.

### **Pilot environment**

At the outset of the Pilot, searching was available to primary datasets on the LC Linked Data Service Authorities and Vocabularies web site, [id.loc.gov](http://id.loc.gov). These were the LC/NACO Authority File and *Library of Congress Subject Headings* (LCSH). Although these two datasets were fully accessible in the BIBFRAME Editor, enhancements were made to the searching mechanism to improve retrieval. For example, enhancements were made to enable searching of non-Latin scripts, special characters, and diacritics. Searching left-to-right script languages presented a particular challenge. As the Pilot progressed, and refinements were made to the BIBFRAME profiles, additional datasets from [id.loc.gov](http://id.loc.gov) were made searchable from the Editor. By the end of Phase One of the Pilot, many more datasets were searchable via the Editor, including some controlled lists from *Resource Description & Access* (RDA).

Three months into the Pilot, the ability to access previously input BIBFRAME descriptions was possible. The descriptions could not be edited, however; Pilot participants understood that the descriptions they created in BIBFRAME did not constitute a database of record and would not be formally distributed as part of the Library's cataloging distribution service.

### **Workflow**

Since the participants were still creating MARC records in the LC ILS there was no need to make changes in workflow. The Pilot participants, when entering data into BIBFRAME, were not operating in production mode. All the data they created will eventually be discarded.

## **Lessons learned**

A good understanding of RDA is essential for working in the BIBFRAME Editor as configured by COIN staff. Some staff members still rely too much on MARC tags and subfields.

Staff members need to start to converse using RDA terminology and not MARC coding.

Pilot participants were much more curious about RDF serializations than the trainers initially thought. There was wide interest in seeing and analyzing the BIBFRAME RDF serializations created during the Pilot, reinforcing the training objectives on the Semantic Web and Linked Data presented in Modules 1 and 2. This lesson was learned in an unexpected way – the inability to save and retrieve BIBFRAME descriptions at the outset of the Pilot resulted in the need for participants to check their work by consulting the RDF serializations. There was lively discussion of the RDF data elements and the Linked Data triples that participants identified in the RDF serializations they created.

## **Best practices**

Pilot participants developed their personal “best practices” in describing resources in BIBFRAME. These best practices were shared with other participants at the informational sessions held after the initial training.

For example, some participants discovered that “grouping” like resources reduced the amount of time spent in description, enabling re-use of data across multiple resources.

Participants working with non-Latin data developed a best practice to pair non-Latin native scripts with Romanized data.

## **Goals met**

Participants were vigilant about cataloging and meeting regularly with instructors to provide input for consideration to improve BIBFRAME functionality. During the course of the Pilot, searching was greatly improved and the profiles considerably enhanced.

In December, Pilot participants started to enter metadata into BIBFRAME first and then create a record in the LC ILS. This critical change, which reflects how catalogers will work in production, was possible because of the participants’ active involvement in the development of the BIBFRAME Editor.

Metadata created using the BIBFRAME Editor version 1.0 are available to a world-wide audience to explore, study, analyze, and use for experiments to increase their knowledge of the linked open data environment. Phase One of the LC BIBFRAME Pilot lasted six months (October 1, 2015 – March 31, 2016), but Pilot participants are continuing to catalog in the BIBFRAME Editor to retain their skills, so BIBFRAME data continue to be created and analyzed. After the LC

Pilot BIBFRAME 2.0 is underway, the data created using version 1 of the BIBFRAME model will be discarded.

### **PILOT PHASE ONE SYSTEM**

*(Network Development and MARC Standards Office—Sally McCallum, chief)*

#### **Scope of the BIBFRAME description creation system**

The Network Development and MARC Standards Office created the technical components that supported the Pilot. The System included most of the Library of Congress's MARC bibliographic records transformed into BIBFRAME descriptions, controlled authority and term lists with URIs, and an input editor for the participants to use. NDMSO staff divided the technology tasks as follows: Nate Trail handled the infrastructure and MARC conversions; Kirk Hess was responsible for the BIBFRAME Editor and Profile Editor; and Qi Tong worked with the linked data resources in the LC Linked Data Service.

The Pilot's focus was on input of data and impact on catalogers. The important function of end user access was not studied. The system also did not support the recording of holdings, acquisitions processes, or description distribution functions, which are not central to the MARC/BIBFRAME vocabulary. Over 2,000 records created in the Pilot, however, were made available in a bulk download file.

The following "questions for exploration" were articulated before the start of the Pilot.

#### **Can catalogers input BIBFRAME descriptions into a BIBFRAME oriented system?**

Using the v1.0 Editor, the Pilot participants submitted over 2,000 descriptions to the system. Eight profiles for different resource types were established to assist with input: monograph, notated music, Audio CD, serials, cartographic, BluRay DVD, 35mm Feature Film, and prints/photographs.

#### **Is the Work/Instance dichotomy clear and useful for catalogers?**

The modeling of Works and Instances was clear, with certain properties in Instances and others specified for Works. However, the Pilot participants generally just looked for the RDA rule and viewed it or put in the value, so how it was packaged by the model was not that important to know.

#### **Do type a-head and drop downs make work easier?**

Dropdowns and lookups were popular features. As expected they improved the accuracy of data strings, provided the data linking URIs without keying them, and made input more efficient. Several term lists were created with RDF descriptions and assigned URIs in support of the drop downs and type a-heads.

### **Is the labeling on the editor clear and useful?**

The labeling on the interface of the BIBFRAME editor used to input the BIBFRAME bibliographic records was designed by the bibliographic specialist, Paul Frank, in collaboration with the Pilot catalogers and the technical implementer of the interface, Kirk Hess. It used labels closely synchronized with RDA and also links to key RDA rules for an element, where possible.

The Pilot participants found the labels and RDA rule links very helpful. The treatment of Expressions in the BIBFRAME model required additional explanation since RDA is keen on Expressions and the BIBFRAME model considers an Expression a Work with links between the RDA Work and RDA Expression.

### **Can adequate searching be implemented?**

Searching as implemented was adequate but could be improved. The look a-head fields were very useful for known item searching, since as one typed one got closer to what was wanted, which also sped up data entry. There was also some “what do you have like this” searching but since the cataloger had the item in hand and was looking for it or it’s Work, known item searching usually sufficed.

On a few occasions the cataloger was used to searching for unknown items, such as subject headings or names, and/or browsing a hierarchy rather than searching for a specific result. In those cases it was recommended that they use an external tool first and then come back to the Editor with the known item that was identified. NDMSO will work with COIN to identify and add additional searching and browsing features in Pilot 2 to meet cataloger needs.

### **Can the MARC records be transformed adequately for cataloger use?**

A decision was made to try to simulate a total BIBFRAME environment which required conversion of the whole LC file of 18 million MARC bibliographic records in order to provide a BIBFRAME modeled back file to catalog against. This requirement was largely met with 13.5 million records converted. The 13.5 million MARC records were split into Work and Instance records, producing 13.4 million Work records and 13.85 Instance records. The transformation was credible, but was always a work in progress. It was good enough to illustrate the Work/Instance separation, although this was not thoroughly tested in the Pilot. A difficult part of the transformation was the transfer of the title and name/title MARC Authority records into the BIBFRAME framework as BIBFRAME Work descriptions and then their merger with related MARC Bibliographic record data, requiring transfer of the subjects and classifications to the new Work description.

The MARC authority records needed by the catalogers were already converted to RDF and loaded into the LC Linked Data Service (LDS) where they have been available for 5 years. For the Pilot the Library of Congress name authorities were changed from weekly load to daily load

in LDS to provide up-to-date authority lookup. Providing input of new authority descriptions into the BIBFRAME system was desirable but could not be met in the timeframe.

## **Conclusion**

The Pilot achieved its aim and is considered a success. The input from catalogers participating in testing the system enabled those developing BIBFRAME to make considerable strides in its development. BIBFRAME 2.0 model and vocabulary have been released and will form the basis of the next phase of a pilot in fall 2016. The data created under 2.0 will be different than data created in Pilot 1.

## **APPENDIX A: Training Plan (COIN)**

### **Training materials design and development**

Modules 1 and 2

- Module 1: Introduction to the Semantic Web and Linked Data was four and a half hours long
- Module 2: Introduction to BIBFRAME Tools was two and half hours long
- Taught in June 2015 using PowerPoint slides, Quizzes, and Exercises

Module 3 consisted of two Units:

- Unit 1 was a recap of the major concepts of the Semantic Web and Linked Data. This was considered necessary because of the significant time gap since Pilot participants were first exposed to these concepts, and because some found the concepts themselves difficult to understand.
- Unit 2 had the primary goal to provide hands-on training on the use of the LC BIBFRAME Editor to create BIBFRAME “records.” Secondary goals were to explain the ‘ground rules’ of the Pilot, and to prepare participants to be effective testers and provide helpful feedback.

For Unit 1, the resulting product was a 40-slide PowerPoint presentation.

For Unit 2, the resulting product was a 51-page manual, with plentiful screen captures to show Pilot participants what they should see at the various stages of working in the Editor. This manual was rapidly produced in a matter of weeks, during which time the Editor was undergoing constant revision.

Unit 2 covered: accessing the Editor; an overview of the interface; terminology (instructors had to first work with developers to decide what to call all of the Editor elements); navigating in the Editor; working with the different ‘profiles’; using various ways to enter data in the Editor

'fields'; understanding BIBFRAME 'responses' and 'behavior'; and interpreting a Preview of the resulting 'record.'

After the initial training, it was determined that Pilot participants would benefit from weekly discussions and several hands-on refreshers to suggest new techniques, announce policy decisions, and help address questions. To this purpose, COIN instructors developed refreshers for the classroom, combining exercises and discussion; a major focus of these was to help Pilot participants stop 'thinking in MARC' and start 'thinking in RDA.'

## Teaching experiences

Some features of the intended training were not possible in this Pilot – notably, a full searching capability and the ability to retrieve a ‘record’ created earlier to make revisions. Even without these important routine capabilities that were considered essential to catalogers, the participants were extremely cooperative, flexible, and ready to experiment. By and large, they understood that they were learning to use a work *very much* ‘in progress,’ and that indeed one of their functions as Pilot participants was to provide help and advice to the developers. The instructors appreciated this and acknowledge their professionalism. The duration of the training for Modules 1 and 2 of seven hours and Module 3 nine hours was about right. Any less would have left them under-trained.

Some important conclusions can be drawn about the “BIBFRAME literacy’ of the target audience:

- The theoretical concepts of Linked Data and the Semantic Web are not easily assimilated by some catalogers. Sufficient time needs to be given to this topic and care taken to evaluate comprehension.
- It was apparent that some of the participants had not yet fully mastered RDA (including the FRBR conceptual underpinnings) -- e.g., title proper vs. preferred title, creator vs. contributor, the names of RDA elements, even using the RDA Toolkit. COIN is considering the need for “RDA refreshers” focused on issues that relate specifically to BIBFRAME.
- The participants’ mastery of MARC may in some ways have hindered their initial ability to use the Editor effectively. They are accustomed to identifying cataloging elements by their MARC tags and subfields; but the Editor was configured by COIN to use *RDA* terminology rather than *MARC* terminology; i.e., they need to think of the title as the “title”, not “the 245.” Accordingly, several refreshers devoted to helping the Pilot participants better understand how to stop ‘thinking in MARC’ and focus on RDA were delivered.
- The need to exercise cataloger judgment seems still to be uncomfortable for some catalogers. Exercising cataloger judgement will become even more important as catalogers become more comfortable with Linked Data, and can ask, “what is the best way to link these data for this resource?” rather than “what MARC tag and subfield codes am I limited to encoding for these data?”