Introduction

This article will discuss the involvement of the Program for Cooperative Cataloging (PCC) in Machine-Readable Cataloging (MARC) to Bibliographic Framework Initiative (BIBFRAME) conversion. In many ways, this is somewhat counterintuitive given that there is currently no MARC to BIBFRAME conversion process specific to the PCC. MARC to BIBFRAME conversion processes, such as those used by the Library of Congress (LC), the Share Virtual Discovery Environment (Share-VDE), and the Swedish National Library, are each specific to a given community and correspond to different flavors of BIBFRAME. Nevertheless, the PCC has formal relationships with LC, Share-VDE, and Linked Data for Production Phase 2 (LD4P2) for work on BIBFRAME, and PCC institutions have been very active in the exploration of these conversion processes. It is timely to look into this topic further given that LD4P3: Closing the Loop includes plans to create Monographic Bibliographic Record Cooperative Program (BIBCO) and Cooperative Online Serials (CONSER) BIBFRAME data collections, and the ability to create original PCC data in BIBFRAME through the Share-VDE conversion processes and with Sinopia (a linked data creation environment). See Figure 1.1
Figure 1. Share-VDE/Sinopia dataflow conceptualization.

This article will then review the involvement of the PCC in the development of BIBFRAME. Focus will then transition to the MARC to BIBFRAME conversion work at LC, Share-VDE, and LD4P2, and how these projects intersect with efforts of the PCC. Finally, we will discuss areas for further exploration by the PCC leading up to the creation of PCC conversion specifications and PCC BIBFRAME data.

Context: PCC, LC, and BIBFRAME

Following the emergence of Resource Description and Access (RDA), the library community was keenly anticipating a shift to standards whereby data lived on the web. Yet by early 2011, it seemed that there was no clear leadership toward an infrastructure for library linked data. The end of the MARC formats had been predicted for years, but no serious alternative for- mat had risen to challenge MARC.

After the national test of RDA in 2010, some thought that RDA could be the answer, or at least create a cataclysm which would mean the demise of MARC altogether. At the 2011 American Library Association (ALA) Midwinter Meeting, the MARC Formats Interest Group held a lively panel presentation called, “Will RDA Mean the Death of MARC?” While the panelists seemed to agree that RDA data could be encoded in MARC, it was also apparent that an
alternative to MARC needed to be developed in order to enable our library data to be “of the web.” Encoding RDA in MARC was never a good fit and could be compared to attempting to pour a pot of coffee into an espresso cup. With the development of BIBFRAME, a suitable vessel was finally available, so that in recent years the evolution of MARC has increasingly been focused on conversion to BIBFRAME in Resource Description Framework (RDF).

In May 2011, LC announced that it had officially launched its BIBFRAME Initiative. This gave the PCC and the rest of the library community a tangible schema to anticipate. BIBFRAME development continued through several phases until, in April 2016, BIBFRAME 2.0 replaced BIBFRAME 1.0. The release of the LC BIBFRAME 2.0 conversion tools in 2017, along with MARC Bibliographic Conversion Specifications was incredibly helpful to the library community, allowing libraries to experiment and see what their own data looks like in BIBFRAME. While LC launched internal pilot projects to test the model and the vocabulary, formal and informal investigations were also going on in the PCC community. Key questions that followed immediately after seeing converted data were, “What should well-formed data look like?” and, “What should be included in a BIBFRAME record based on our current standards?”

Initial attempts at answering some questions began through the efforts of CONSER and BIBCO. In fall 2015, a CONSER Task Group began to review how well BIBFRAME could be mapped to the CONSER Standard Record (CSR), taking an element-by-element approach to evaluation. Similarly, a BIBCO group reviewed the BIBCO Standard Record (BSR) to BIBFRAME mapping. When BIBFRAME 2.0 was released in early 2016, the PCC Policy Committee (PoCo) charged the more formal PCC Linked Data Advisory Committee (PCC LDAC) to make recommendations on questions such as, “What role do standards like RDA, BIBFRAME, and others, play in library linked data strategies?” and “Depending on decisions made on managing linked data, what should replace the BSR and PCC CONSER Standard Record (CSR)?”

By 2018, as a result of the work done by LC, PCC, and other institutions, a tipping point was reached based on a number of new projects and increased work on BIBFRAME that resulted in one of the largest shifts in cataloging and metadata
standards and practice in recent memory. Sally McCallum captured this well by comparing our current state to past transitions:

“In the 1960s and 1970s, the AACR cataloguing rules and MARC format for bibliographic data were developed. Forty years later we are in the transition to new cataloguing rules and also a new carrier environment, with RDA and BIBFRAME.”9 At the 2018 ALA Annual Conference, LC confirmed that BIBFRAME will be their replacement for MARC.10 This was a major signal to the library community given the worldwide influence of the Library of Congress. Moreover, LC’s work has not been done in isolation, as demonstrated by the July 2018 announcement by the National Library of Sweden that they were the first national library to fully transition to linked data with a BIBFRAME implementation.11

MARC to BIBFRAME: LC, LD4P, and Share-VDE

The available tools, specifications, and mappings released by LC in 2016–2017 enabled many PCC libraries to begin experimenting for themselves. Being able to reference existing standards and to conduct tests with library metadata facilitated learning and identifying areas where the tools needed improvement. As a result, 2017–2018 represents a period of rapid development and analysis within PCC, testing by PCC member institutions, and work on early implementation scenarios at several national libraries.

From 2018 onward the game has truly been afoot. In the PCC strategic directions for January 2018–December 2020, there was a significant change of focus:

“It is time to move beyond knowledge and skills related to linked data at a theoretical level and into implementation … As we move to a culture of greater data sharing, it is crucial to extend our community, both by engaging a more diverse range of members in the work of the PCC and by collaborating with vendors, open-source communities, and others.”12

With ongoing development in the international BIBFRAME community, and the stated need to begin work on implementation for PCC, it became increasingly clear that PCC needed a way to work on creating or adopting best practices. The Report of the PCC BIBFRAME Task Group brought up key questions for PoCo, e.g., “What is next beyond the mappings?” and “Do we want to facilitate a pilot
using BIBFRAME tools with PCC member institutions to create bibliographic metadata in BIBFRAME RDF? A pilot project was a natural next step and has been created through LD4P2 and in collaboration with Share-VDE.

LD4P2 cohort

The creation of LD4P2 established collaborations between the LD4P Partners (Cornell University, Harvard University, Stanford University, and the University of Iowa School of Library and Information Science), an extended cohort of PCC member institutions, and the PCC. One of the key goals for this group is the creation of a growing pool of linked data expressed in BIBFRAME-based application profiles. It is worth noting that to supply institutional data for the RDF data pool, the LD4P2 Cohort used Share-VDE, “For the cleanest, most reconciled pool of collective data shared across collections.”

A primary focus of LD4P2 is the creation of a BIBFRAME cataloging editor for the description of resources to PCC standards of use in a cloud-based environment. This editor, Sinopia, is currently being refined and original resource description in BIBFRAME is underway by LD4P2 Cohort members. This presents libraries with a very concrete path forward but highlights the need for additional work and analysis, particularly in terms of the interaction of the flow of MARC data converted to BIBFRAME with what is created in Sinopia and other BIBFRAME editors.

Share-VDE

Share-VDE is a community-driven initiative to implement a linked data environment for descriptive metadata, influenced by the vision of the LD4P projects as well as an international group of national and research libraries. Share-VDE is composed of several PCC institutions as well as institutions across Europe and North America and so constitutes a wider community than the PCC alone. There is a lot of overlap between the groups discussed in this article.
Working directly with PCC, and through LD4P, Share-VDE has a significant role in the ongoing efforts for MARC to BIBFRAME conversion, reconciliation, and enrichment.

As was the case at the University of Alberta Library and Duke University Libraries, additional support for testing BIBFRAME conversion at scale and, “Having an external agent, driven by community values, that can proof new discovery concepts for linked data and help with some of the heavy liftings for data conversion, reconciliation and enrichment” was, and continues to be, very helpful for many institutions. Moreover, “the scale of the Share-VDE project, with a large number of major research libraries working toward linked data implementation, the importance of work to analyze associated processes and metadata become obvious.” Doing so through community collaboration has been critical due to the scale and complexity of work which is larger than any one institution.

The Share-VDE project includes a virtual discovery platform (http://www. share-vde.org) with a four-layered adaptation of the BIBFRAME data model which was developed to provide a linked data discovery option. From the start in October 2016, and unlike other efforts in the BIBFRAME community which focused on creating new metadata in BIBFRAME, the Share-VDE project began with a focus on the conversion of metadata from MARC to RDF using the BIBFRAME vocabulary and other additional ontologies as needed.

In August 2018, based on discussion at the Share-VDE workshop, the Share-VDE Transformation Council was formed based on the need for analysis of MARC to BIBFRAME conversion, and more organized recommendations from the Share-VDE membership; this group was recently renamed the Share-VDE Advisory Council (Share-VDE AC). Due to the active participation among the community, a more formal structure was needed in order to vet which recommendations would be actively pursued. Almost immediately, the need for sub-groups was recognized, so charters and use-cases were developed for four working groups:

a) Work Identification Working Group (WIDWG)
b) Authority/Identifier Management Services Working Group (AIMS)
c) Cluster Knowledge Base Interaction/Editor Working Group (CKB WG)
d) User experience/User Interface Working Group (UX/UI WG)
The work of the Advisory Council and the nature of the working groups highlight key challenges for Share-VDE, but also may inform future work for the PCC. Because of the overlap between the Share-VDE members with LD4P and PCC membership, there are formal connections built into the charges to work with groups involved with linked data and BIBFRAME, including the PCC.

The Share-VDE AC and Working groups analyzed LC’s conversion specifications, reports from analysis done by member libraries, the converted data available in the Share-VDE portal, and the LC BIBFRAME Comparison Tool. While the time frame was short, the number of quads (triples plus provenance) in the cluster knowledgebase quickly reached into the billions as the data from each member library were converted and added to the data-store. The conversion of just over 100 million MARC records resulted in approximately 24 billion quads. Following an iterative approach to testing with this data, the Share-VDE AC released several recommendations for Share-VDE which were finalized in December 2018.

The four general recommendations are supportive of collaboration and openness in order to promote shared community development. These are:

a) BIBFRAME properties should be used instead of other ontologies whenever possible in order to promote consistency and sharing of data.
b) Share-VDE conversion specifications, based on LC Conversion Specifications, need to be published to support community development and ongoing analysis.
c) The library community should be prepared for the need to iterate the conversion process in 1–2 years to support implementation strategies. Share-VDE AC should continue to communicate with library community members.
d) Share-VDE developers should coordinate with LD4P developers to ensure that supported vocabularies for Sinopia are utilized by Share-VDE.

The remaining 10 Specific Recommendations are less general and include topics such as: establishing preferred sources for Uniform Resource Identifiers (URIs), the definition of the Share-VDE Opus, previously called SuperWork, which was subsequently developed as Hubs by LC, and improved conversion of the series added title metadata. To avoid confusion with
other definitions of “superwork,” the Share-VDE SuperWork was renamed the Share-VDE Opus in February 2020.

PCC BIBFRAME data: key considerations

1. International BIBFRAME community and BIBFRAME as an international standard

The work of LD4P and Share-VDE also adds to wider collaborative efforts and the creation of a community. In 2017, the first European BIBFRAME Workshop was held at the German National Library in Frankfurt with the goal to “be a forum for sharing knowledge about the practice of, production with, and planning of, BIBFRAME implementation.”18 This was followed by another European BIBFRAME Workshop in Fiesole in 2018 with international participation of eighty participants from twenty libraries.19 In Fiesole there were calls for, and discussions of, the need for an international community to support BIBFRAME as an international standard.20 One of the results of these discussions was a shift to make the group less centered on Europe, and thus the “BIBFRAME Workshop in Europe” was held in 2019 in Stockholm. While this is not a formal governing body, the organizing group has representation from the Royal Danish Library, Casalini Libri, the German National Library, the Library of Congress, Stanford University, the National Library of Finland, and the National Library of Sweden. This group has been key in coordinating discussion on an international scale and supporting the development of best practices, and the workshops have quickly become a forum for the community to come together.

Considering current MARC usage, and the worldwide reliance on shared cataloging practices by libraries, the move toward internationalization for the discussion of MARC to BIBFRAME conversion is of critical importance. Other factors aside, we need to be able to use BIBFRAME for data interchange. Just using LC, Share-VDE, and the National Library of Sweden as examples, these represent three flavors of BIBFRAME so having an international community is important to ensure that where differentiation exists it does not inhibit the sharing and reuse of data. Given the PCC’s focus on cooperative cataloging at an international level, engagement on this front is critical. One
example could include working alongside the wider library community to build the infrastructure to maintain standard BIBFRAME practice. While there has been some reluctance to slow the pace of development by adding such a formal structure, it will soon become important to have a body similar to the MARC Advisory Committee to assure BIBFRAME metadata interoperability and reuse, and to support BIBFRAME as an international standard.

2. Clustering, identifier use, and identifier management

A key element of conversion includes enrichment with URIs. If text strings from MARC fields are solely converted to the same in RDF then BIBFRAME would remain relatively flat, which obviously is not ideal. Entification, the process of identifying entities in data to describe them using appropriate vocabularies and relationships, is an important aspect of the overall transformation in order to support URI enrichment. URI enrichment can also happen within the MARC data to support the MARC to RDF conversion process. Whether enriching in MARC or through the process of conversion, URI enrichment comes with challenges, both technical and in terms of deciding the appropriate URI vocabulary for the right element. URI vocabularies are those which have been updated to include URI data.

The PCC Task Group on URIs in MARC created the foundation for this work, with one output being Formulating and Obtaining URIs. This document is now being used as a key source for the new PCC URI Pilot launched in September 2019. The PCC LDAC has provided additional guidance on assessing URIs for use in MARC data.

While these guidelines are intended for the enrichment of MARC data, the principles are equally valid for enrichment through conversion. However, thought is needed to match vocabulary usage in the original description and other update processes to those of conversion for data consistency. Ultimately for PCC, data choices will need to be made about which URI vocabularies can be supported and maintained in shared systems. Best practices for entification and associated URI vocabulary usage will be informed by the Questioning Authorities project.
(QA) of LD4P. The QA project has involved the creation of entity-based searching, or lookups, for URI vocabularies utilized in Sinopia for description with BIBFRAME. Given that enriched BIBFRAME data is also available in Sinopia from Share-VDE, comparative analysis of enrichment through conversion and entity use in description has been ongoing.

With multiple URI vocabulary options, it also becomes important to build relationships between them. At the same time, there are elements that lack identifiers and are critical for resource discovery or the general structure of BIBFRAME data (Opus, Work, etc.). Stemming from these needs, Share-VDE has processes to cluster data elements and create URIs for them. The various cluster identifiers in Share-VDE are then stored in the cluster knowledge base, called Sapientia, or, the CKB. Sapientia is a triplestore containing Opus, Work, Agent, Publisher, and other elements along with associated relationships. Creating these processes and having such a valuable tool for the library community, then requires mechanisms for identity and identifier management in the platform. Share-VDE will soon have an editor for these clusters, named J. Cricket. As these tools and initiatives mature, it will be very interesting to see how current methods of authority control evolve.

3. Cataloging standards, conceptual models, and application profiles

In the early phases of the Share-VDE project, MARC data from imprint years 1985 and 2015 were utilized to compare how the conversion processes handled MARC data reflecting different cataloging standards, such as differences between AACR2 and RDA. When discussing the conversion of bibliographic data from MARC to BIBFRAME, we must remind ourselves that the encoding standards are being converted, but not necessarily the content standards. In the past, many people conflated AACR with MARC and referred to it as MARC cataloging. RDA is a successor to the AACR family of cataloging guidelines; BIBFRAME is a successor to MARC. Given past confflation between content and encoding standards, conversion mappings from MARC to BIBFRAME pose several challenges. More generally, given the legacy of MARC cataloging, and variations of cataloging practice across time and institutions, conversion processes need to be flexible enough to handle differentiation.
The other side of the coin, though, is that the BIBFRAME from the conversion outputs still need to match with the resource templates for associated application profiles for cataloging native to BIBFRAME. In close collaboration with PCC, the LD4P Cohort has been working on original and copy cataloging with BIBFRAME, and significant effort has gone into the creation of BIBFRAME application profiles. One observation from this work is that when cloning external BIBFRAME data, differentiation in BIBFRAME, say between LC and Share-VDE, poses challenges for mapping elements appropriately. These challenges point to the need for standardization for use of BIBFRAME for data interchange. BIBFRAME converters must be flexible enough to handle varied MARC, but also create BIBFRAME and associated identifiers in a standardized way. This means that as changes are made to cataloging standards, conversion processes of the BIBFRAME model and/or application profiles, must also adapt. Knowing this and that many changes are coming to RDA, means the PCC will need to develop BIBFRAME application profiles in a clear, purposeful manner, and then adapt conversion specifications accordingly.

Bringing together conversion issues related to identifier creation and use, data models, and application profiles, there have been several significant updates to BIBFRAME through Share-VDE. At the LD4 Workshop at Stanford University in May of 2018 conversations identified the need for a group to work on refining the creation of work identifiers for the Share-VDE project and BIBFRAME. The need and prioritization of efforts to create primary resource identifiers were based on the lack of entities and associated URIs to describe the work-level in BIBFRAME. Without universal identifiers for works, data updates and any new additions to the dataset would result in significant duplication and the inability to cluster resources together. In August 2018, the Share-VDE AC WIDWG began its charge to:

1. Review the Share-VDE work clustering outline and submit feedback on potential improvements or optimizations
2. Review the use of primary resource identifiers in the Share-VDE data set and provide feedback as appropriate
3. Engage with the PCC to identify and/or develop best practices for use of these identifiers in BIBFRAME and MARC data
Participants brought significant experience in the analysis of BIBFRAME data, and the conversion from MARC to BIBFRAME, in relation to both the LC converter and the Share-VDE process.

Initial analysis focused on the examination of the existing Share-VDE Work Clustering Outline utilized for the creation of phase 2 data. These initial algorithms, developed by Casalini Libri and @Cult, gave an excellent starting point for utilizing elements of MARC records to identify works, but the scale of the challenge quickly became evident. For BIBFRAME to be viable, a practical and sustainable method for the creation of work identifiers was needed quickly; and, the WIDWG, along with the Share-VDE AC, had just a few months to achieve this analysis in addition to completing other conversion processes for the project. All the analyses needed to be completed prior to the conversions required for the LD4P2 Cohort project.

Analysis quickly revealed a few guiding principles:

1. Other work URI vocabularies did not lend themselves to practical application in Share-VDE.
2. The existing clustering algorithms needed to be refined and adjusted to take MARC practice into account and cover the identification of more works and work relationships.
3. We wanted to create work identifiers that would be useful within Share-VDE, but also for the wider library community.
4. We wanted to keep an open mind about the definition of work, and how to work identifiers are created for Share-VDE.

Analysis of MARC to BIBFRAME conversions through the Share-VDE AC reinforced the effectiveness of the BIBFRAME work as currently outlined, which includes elements corresponding to the work and expression level elements in IFLA-LRM.

Through analysis of clustering like works, the need for an additional level of abstraction in the data model was identified. In January 2019, a new Opus class was introduced in Share-VDE and LD4P data. Just prior to the ALA Annual Conference in 2019, LC introduced the Hub to their data model; stemming, at least in part, from the realization that they were “trying to do too much with bf:Work” (see Figure 2). While further analysis and refinement of practice for these related elements are needed; ultimately, they both serve the same general function in
BIBFRAME; both the Opus and the Hub are very close to the use of the Work in IFLA-LRM (International Federation of Library Associations and Institutions-Library Reference Model). These three entities are a very close match and represent a new family or type of bibliographic element. At the same time, justification for testing the application of Opus is based on two key arguments:

1. While Work to Expression relationships can currently be expressed in BIBFRAME, these are ultimately Work-Work relationships, and determining the initial or primary work, or hierarchical relationships between works may prove difficult with this structure.
2. Through conversion from MARC to BIBFRAME, or automatic work ID generation based on BIBFRAME elements, unless we can define a difference (a fingerprint for each cluster or constellation) between Work and Opus elements, then these relationships (work-expression) cannot be captured through conversion or automated processing. With the scale of data conversion underway, not doing this would seem like a missed opportunity. Once a separate fingerprint is defined for this primary work, it needs a name, thus the creation of Opus.

The Share-VDE Opus is intended to group BIBFRAME works of functional and/or near equivalency. The Opus is defined by the elements representing commonality of content between BIBFRAME works. There are also one-to-one or one-to-many relationships between Opus and Work, where for each Opus at least one BIBFRAME Work exists; each Work has a corresponding Opus; and elements included in Opus are also included in the definition of bf:Work.
As a result of this model update, based on primary resource entities, BIBFRAME can be seen as a practical application of RDA. Where IFLA- LRM outlines the fundamental theory for a conceptual model, BIBFRAME now has a very close parallel for its data model. This is in line with the position of the PCC to “continue to treat RDA as a loose content standard.”

With the inclusion of the Opus in BIBFRAME, entity definitions need to be adjusted. This change presents an opportune time to reflect on differences between the conceptual model of IFLA- LRM and BIBFRAME. Complete alignment between a conceptual, entity, and data models cannot be expected, but these are an extremely close match.

Figure 2. Illustration of the Opus concept.
The distinction between types of Opus, the Share-VDE Opus and the LC Hub, and the relationship between the updated model to IFLA-LRM warrant further discussion within the PCC, the international BIBFRAME community, and IFLA. This case highlights an area in which the PCC may wish to invest further effort, because more identifiers for opus, work, and instance are emerging, best practices and standards will be required (see Figure 3).

Figure 3. Proposed Types of Opus.

It seems clear that building best practices and standards for the use of primary resource identifiers will be an area of need for the PCC in the near future. Work on model design for Share-VDE may be of considerable benefit for this discussion (see Figure 4).

Figure 4. Share-VDE Model.
Further discussion

Through participation in projects like LD4P2 and Share-VDE, the PCC has a pathway to identify and apply best practices and re-tool for work in linked data. In particular, the noted recommendations of the Share-VDE AC and working groups indicate a growing maturity for BIBFRAME as it is used in practice. As examples of BIBFRAME maturity are the inclusion of the Opus concept and the efforts to harmonize processes for original data creation in BIBFRAME with data created through conversion. Each of these cases could easily have been discussed at length for years without resolution; with the ability to test principles and processes with real-world data harmonization is developing quickly. Development of BIBFRAME continues beyond the environments of LC, LD4P, Share-VDE, and the PCC, but these cases are illustrative of our current state.

The PCC will have many challenges and opportunities ahead as libraries update processes and workflows to transition from MARC to BIBFRAME. In his presentation on the implementation of Libris XL, Niklas Lindström discussed the balance of working versus perfect, and that the project was “not done, just real.” Similar statements can be made about MARC to BIBFRAME conversion, and the work of PCC on BIBFRAME in general. LD4P2 is still in process, in many ways Share-VDE is just getting going, and national and large research libraries still have much to consider, yet with the advent of these projects, we now have a very real path forward with tools that work. We now have a viable replacement for MARC that will allow libraries to move forward with linked data. Moreover, with BIBFRAME we finally have an encoding standard that can appropriately represent RDA.
References


