



PCC BIBFRAME Data Exchange Meeting

September 9-10, 2021

Online

Summary

The Program for Cooperative Cataloging (PCC) called a two-day meeting to discuss exchange of BIBFRAME data between systems and implementations. There are a number of different projects in the works experimenting or implementing BIBFRAME and linked data -- an essential question now is to ensure that we can all “talk to each other” in the end.

Expected outcomes:

- Better understanding of who is committed to which types of services/tooling
- Identify what exchanges are desired between which systems
- Identify challenges/showstoppers and steps to address them
- Shared ownership of challenges and support for BIBFRAME implementation
- Determine next steps to keep the conversation going

Represented groups:

British Library; EBSCO; European BIBFRAME Group; ExLibris; FOLIO community; German National Library (Observer); Index Data; ISSN International Centre; Koha; LD4P3; LD4 Steering; Library and Archives Canada; Library of Congress, Network Development and MARC Standards Office; Library of Congress, Policy, Training and Cooperative Programs Division; National Library of Medicine; National Library of Sweden; OCLC; PCC Policy Committee; PCC Standing Committee on Standards; PCC Standing Committee on Training; PCC Steering; QA; RDA Steering Committee (Observer); Share-VDE (Casalini); Sinopia

Day 1 began with a report by Beacher Wiggins (Library of Congress) on a linked data summit between Library of Congress, OCLC, Share-VDE and Stanford University Libraries that was held in August.

The issues discussed were closely related to the ones raised on the agenda for this PCC meeting. The participating groups were looking to address the questions of distributing BIBFRAME descriptions, handling serials, ensuring that locally created data will be shared, and OCLC’s infrastructure as a component of sharing linked data.

The outcomes of this meeting included working groups and plans for future summits. Two working groups have already been established: Use Cases (chaired by Philip Schreur) and Data Exchange (chaired by Marti Heyman).

Philip Schreur presented a first use case outlining the creation of an e-record. The cataloger finds a record for the print version; adapts the description within Sinopia, then pushes to Share-VDE. The full BIBFRAME description will be exported to OCLC for sharing with the community.

Working groups to come: Technical documentation/Exchange validation. Ideas for future summits include: serials, data openness (use/access), and discovery.

Discussion: Participants expressed interest to interact with the work of these groups. Questions were raised where and how the results of working groups and summits will be shared and how the larger community could be included. Future work could include people from other units and experts from other places.

[Detailed slides of this presentation are available.](#)

Agenda topic: **What does data exchange mean in a BIBFRAME environment?**

Philip Schreur (Stanford University Libraries) introduced this topic.

There was a lot of consensus around sharing of PCC data in a BIBFRAME environment:

- Our data exchange should not be limited to PCC data only which could create a new silo; but the group recognised the benefit of starting with a smaller set of data and then to expand. The PCC also offers a defined community to help test what is working and what needs to be enhanced.
- It is critical not to lose sight of data other than bibliographic data, for example id.loc.gov and other linked data services.
- PCC data should be freely available even if at present not everybody will be able to make use of it due to lack of technology.
- The question was raised how we define PCC data -- Are we just considering data coded 042 "pcc" or all data created by a BIBCO/CONSER member; or data created by other libraries that follow PCC guidelines. This topic was added to the November PCC Policy Committee meeting agenda for a decision.

The conversation around agreements covered both the need for foundational agreements (covering community, data goals, infrastructure, etc.) and more "legal" agreements or MOUs covering shared effort and responsibilities. The suggestion was made that the former could be structured following the model of a [BIBCO funnel](#).

Individual institutions do not need an agreement to use data from the PCC data pool. For instance, Share-VDE and the PCC pool are currently open for searching, download, and so on.

Though there are differences in authorization levels, for example, to directly work on Share-VDE or PCC data, using the specific tools and protocols.

Question: What do we mean by “exchange data”?

Data must be as accessible as possible. However, practically exchanging data involves access (for consumption), uploading data, as well as updates. Open issues: Where do updates take place (locally, network-level) and what triggers an update?

URIs help the data exchange. We need to continue to move away from strings.

The group agreed that our short term needs will very likely be different from our long term needs.

Local caching may be what we need in the short term to drive discovery, but the ideal/future might be different. We will need to think about use cases in terms of short and long term.

The same might apply to exchanging both description sets (aka records) and specific statements (triples). We need to develop use cases for the latter. Most use cases right now tend to focus on exchanging entire resource descriptions (or at least work or instance descriptions)

Related: Work of the [Best Practices for Authoritative Data Working Group](#).

Agenda topic: **Come to an Agreement on Core Needs for Data Interchange**

This topic was introduced by Jeremy Nelson, Stanford University, and Sally McCallum, Library of Congress.

[Presentation](#) (Jeremy Nelson)

[Presentation](#) (Sally McCallum)

Both presentations highlighted challenges in interchange of BIBFRAME data -- these problems can occur both when different choices are made in original data creation (e.g. LC BIBFRAME editor versus Sinopia) and in data conversion (primarily from MARC but also from MODS). Some of the different implementation decisions of the same ontology may seem minor, but they do add up.

Jeremy's presentation shows examples of title statements and data formation differences in Share-VDE, LC, and Sinopia.

Sally McCallum's presentation made the case that simplified and coordinated MARC to BIBFRAME and BIBFRAME to MARC is one of the core needs for BIBFRAME data exchange.

Questions raised for the BIBFRAME to MARC conversion included: What constitutes a useful metadata package? What elements go into work, instance, item description? What is a MARC record/unit?

Discussion on MARC/BIBFRAME:

For the purpose of defining what is needed in a functioning MARC description: MARC and linked data have somewhat different goals: The former is a surrogate that lets us find the resource; linked data descriptions can be a carrier of information in itself -- a way to explore relationships through metadata.

The ISSN Center started work on this and is eager to contribute.

Libraries need to share their data as a group. If it is shared individually this will be a disservice to the user. Therefore, best practices need to be developed and hosted at a shared space. We need to define a subset of BIBFRAME data to create the operational MARC record. This approach would help with interoperability into other data sets.

Discussion on different implementation decisions:

Follow up group is needed with the goal of defining the shape of a BIBFRAME entity and to identify the reasons for modeling differences.

Application profiles answering some of these questions are in process for PCC. The Task Group charge is available [here](#).

SHACL and other ways to define profiles can be used to define a floor (as we do now in MARC) and be open to additional information.

An unresolved issue continues to be "the data of record" -- creating and updating data upstream versus downstream. This is not as clear in a BIBFRAME environment as it was in MARC.

Agenda topic: **Agreement on use cases and best practices on when and why to link out and when to copy data and cache locally**

Topic introduced by Steven Folsom, Cornell University Library

Status updates by vendors:

ExLibris (Jane Burke)

Background:

Vendors (including ExLibris) have been providers of local systems (both resource management and discovery), but linked data is an ecosystem (graph cataloging, interoperability, discoverability).

Interoperability is considered a requirement and a goal for ExLibris' participation.

To improve discoverability, the network can be used to enrich what is presented to the end-user. This benefits data reusability, higher quality data, access to rare information, accommodation of different points of view and multi-lingual access.

Challenges are diversity of POV, multiple sources (de-dupe, combining data). Lack of best practices on when to cache, when to copy and when the link out, and data packaging -- what part of data should be packaged for distribution?

Current state:

Alma is using the Library of Congress conversion of MARC21 to BIBFRAME and publishes those as BIBFRAME (several years in place).

ExLibris is in the process of integrating 3rd party editors and is saving BIBFRAME data internally in Alma.

Record enrichment: Alma Refine CloudApp (look-ups against vocabularies), API endpoints

A prototype discovery environment that leverages linked data exists.

Roadmap:

Cataloging with existing editor in 2022

Future: Cataloging with metadata editor in Alma; managing authorities as linked open data;

Look-up for LoV

Share-VDE (Anna Lionetti (Casalini))

[Share-VDE Presentation](#)

The LOD platform (technological framework) consists of multiple layers:

Data layer: Datasets are converted to BIBFRAME and MARC datasets are enhanced.

Applications layer: JCricket editor is being built, Discovery portal.

Services: Triple store indexing, integration with other systems, authority services.

Technology layer: APIs, advanced entity model, tenant infrastructure (allows for institutional config).

Anna believes that Share-VDE can support the workflows suggested in this session, including linked data publishing (new entity-based discovery environment launching later in September and multi-tenancy); aggregation/Look-up services (clustering process for reconciliation; REST APIs, GraphQL APIs and triple store for queries); conversation to/from linked data - PCC data pool; already converting to MARC from Library of Congress linked data sources in the context of Share-VDE authority control services; RDF-native entity-editing using JCricket, and authority services now implemented for MARC and evolution to authority control based on linked data. Different groups of libraries have different requirements and so have different tenants in Share-VDE. Data are stored separately but connected via the Share Family Index. A separate PCC discovery interface is possible, following the Kubikat model.

Index Data (Sebastian Hammer & Wayne Schneider)

[Linked Data in FOLIO](#)

FOLIO is an API-based library service platform built on microservices with an open community developing open source code. There is no linked data in FOLIO right now; instead Index Data is waiting to see what tooling develops to connect data in the cloud to local data for libraries.

Tooling is on the roadmap but not under development.

FOLIO architecture:

The Stripes user interface (UI) is a modular Web application built in Javascript that has a number of different apps (e.g.: check-in, Inventory, Data Import, etc.). The UI communicates with the Okapi API Gateway, managing the back-end module exchange (business logic and storage modules).

For example: Inventory, an abstraction layer that allows for managing resources; conceptually connected to the BIBFRAME model.

The Harvester/Edge Module is being used by Chalmers (Sweden) and GBV (Germany) to manage bibliographic data externally with a harvest into FOLIO.

Longer-term goals:

Entity Management in FOLIO -- a vision document is available: <https://wiki.folio.org/x/OSsuAg>.

Apps would provide storage, extensible data model, abstraction layer/cache and allow more cataloging by reference similar to what ExLibris mentioned re: graph cataloging.

The LD4P3 grant is working on a connection between Sinopia and FOLIO (aka Sinolio) and Share-VDE have started an investigation to develop APIs to integrate SVDE and other tenants data pools in FOLIO.

Discussion:

Can diversity created by different points of view be seen as a benefit of linked data rather than a challenge? Vendors agreed that this is a challenge but also an opportunity. Share-VDE is trying to address this with shared discovery environments; also the idea of having multi-tenant infrastructure with each tenant connected is another way of keeping together entities and resources stored in many places.

It was pointed out that we need to distinguish caching from cloning: A cache in a typical IT environment is just a performance enhancer, not a persistent store.

Examples of ongoing work on system integrations with FOLIO:

- Stanford is currently working on a proof-of-concept for Sinopia/FOLIO API integration. Catalogers will be able to create a BIBFRAME instance description in Sinopia and have a portion of that description appear as a FOLIO inventory record.
- The Chalmers Technical University in Sweden uses FOLIO in conjunction with a national BIBFRAME-based catalog, so all “cataloging” happens centrally and is merely cached/mirrored in FOLIO.
- The Share family community has started an investigation to develop APIs to integrate Share-VDE and other tenants data pools in FOLIO.

Agenda topic: **Ideas for a testing interchange**

This topic was introduced by Christie Thomas, University of Chicago Library, and Michele Casalini, Casalini Libri.

[Presentation](#) (FOLIO, Christie Thomas)

[Presentation](#) (Share-VDE, Michele Casalini)

Christie Thomas presented the perspective of a local institution, University of Chicago, a soon-to-be FOLIO implementor. This included process, types of integration in production and those wanted. The EMWG (entity management working group) defined use cases. This group was led by Jason Kovari.

The inventory model was designed so that the local record representing the bibliographic entity is not dependent on any specific source or format of bibliographic data. Instance records can be derived from the MARC bibliographic store (internal to FOLIO) or the data can be surfaced from a variety of external sources. German consortia manage Pica data externally but integrate into FOLIO; Chalmers is doing something similar with BIBFRAME.

EMWG User Stories available at: <https://wiki.folio.org/display/MM/Use+Cases+-+EMWG>

Conceptual modeling illustrating data flow and FOLIO infrastructure can be found in the vision document: <https://wiki.folio.org/x/OSsuAg>

Jira issues are available for each of the work areas outlined in EMWG; available in issues.folio.org (and search for entity management.)

Michele Casalini (Share-VDE / Casalini Libri)

Share-VDE is working on two layers.

- The existing workflow for PCC data pool fetches data from OCLC, enriches it with URIs for external sources, reconciles entities through clusterization process, and converts to BIBFRAME. The BIBFRAME datasets are openly available and sent back to OCLC. Share-VDE has a long term commitment for housing the PCC data that contributes to form the SVDE BIBFRAME node.
- Identification of further parts of the workflow based on outcomes of the use cases working group. Possible model: outline use case, identify functional/tech components; development of components; testing.
Tests can already be done for the existing part of the workflows on the data pool exchanged between PCC - OCLC - SVDE, but there is a need to identify tests that are not already in place in the various institutions and take advantage of LOD in order to “close the loop” goal of LD4P3.

Discussion:

Question: BIBFRAME datasets are sent back to OCLC -- what is OCLC doing with the data then?

OCLC is storing the PCC BIBFRAME data while finishing the work on the [entity infrastructure](#). After the grant, OCLC will work on two projects: A tooling system in the product area that would allow people to edit entities (building on grant work). The infrastructure behind this will help manage data sources and bring in external data (libraries' decentralized BIBFRAME, VIAF and other non-library data), also to produce BIBFRAME. This work is slated to start in January 2022 and include the PCC data.

Question: What comprises a dataset? A graph of a single bibliographic description?

In Share-VDE the PCC dataset comprises the whole catalog of PCC libraries, converted to BIBFRAME: the output is both the “bibliographic” data in BIBFRAME and the Cluster Knowledge Base of entities. This consists of two graphs: one for bibliographic data set (converted and

enriched to CKB). The vision for SVDE 2.0 is that the graph will include both entities (Works, authors, places, etc.) and Instances - one graph with all possible entities.

Question: How clean is the data we're exchanging? It is important to figure out whether we are making data better OR introducing bad data. There are systems where an unqualified name is associated with the wrong authority record because of string matches. In thinking about exchanging data, we need to ensure that we do identity management once. Can we identify best practices to do identity management in one place?

Question: The agenda listed that we need to develop a plan for testing. Michele stated that we can do this with the PCC data pool. Where do other contributors to the data pool see themselves in this testing picture?

ExLibris is happy to partner with anyone who wants to partner on this. E.g. Share-VDE because of their richness of data. They are also testing the Sinopia editor. We need to take multiple pathways to sort out what is best.

Agenda topic: **Discuss establishment of an “International BIBFRAME Standardization and Exchange Group”**

to move ahead on action items agreed upon at this meeting, and to continue the conversation for areas we still need to address. This group could be PCC led but open to others as well.

Discussion:

Scope of an International BIBFRAME Data Exchange Group:

This meeting highlighted that there are already many use cases stored in many different projects and groups. Bringing this together in one central place would allow for mitigation of duplicate work.

Question: MARC is managed by the MARC Advisory Committee. Is there a parallel? Thinking about BIBFRAME as an international standard, do we need an international body to guide development? E.g. core data that needs to be present across systems (minimum viable description) and keeping future changes in sync.

The Library of Congress has been working on having a process for keeping the ontology up-to-date. The focus of the early years was on stability to encourage experimentation.

BIBFRAME news: <https://www.loc.gov/bibframe/news/bibframe-update-an2021.html>

The point was made that, while all appreciate stability, a community led group could also contribute to that goal. If others can contribute to the development of the ontology alongside best practices for use this could lead to better standardization. It would be useful in such an environment for people to test their proposals before requesting a change to the ontology.

The issue of the BIBFRAME Extension ontologies was brought up, how those would be approved and implemented. Ideally, the process for Ontology engagement would be extended to the preliminary Application Profiles and extensions.

This group can be the place to have the ontology work and the exchange needs discussed. We could speak with vendors and developers in confidence that we have a model and need these data sets to interact. This group would understand the ecosystem and be able to help libraries engage in this work, as well as be a switchboard for where the conversations are happening (i.e. OCLC and Share-VDE, etc.)

It was noted that there is a difference between the ontology and how we want to exchange it (implementation decisions). We don't want to confuse these two.

Stanford is interested in linked data, but not exclusively BIBFRAME. How do we deal with different ontologies in the same system? How much do we need to restrict the structure of BIBFRAME, and is it really necessary? Flexibility between and within platforms for including multiple ontologies is important.

Building ontology crosswalks could be a useful area for collaboration, so is nailing down the graph shape and introducing shared tools for validation.

There was an agreement that the ontology and best practices on how to use it are separate parts. Catalogers working in Sinopia discover issues that relate to the ontology and others that relate to the current lack of documented best practices. We need a place to present decisions that are made.

Approach:

A lot of consensus developed around the suggestion that we need a timeline to develop support for BIBFRAME exchange. A rough timeline would allow us to work back from that. The BIBFRAME Standardization and Exchange Group could work on this, identify key organizations and early implementers for testing. Eventually this group could become more formalized like in MARC, but earlier on focused on those working towards initial setup and implementation.

Next steps: The PCC Policy Committee will discuss at its November meeting if/how we can develop a framework and bring it back to this group and to the community.

Note that this should not be limited to PCC, but rather the PCC taking the lead and then coordinating and including other groups and the European community. One option is to add liaisons to other groups like FOLIO, European community, etc.

Agenda topic: Planning of follow up meeting covering BIBFRAME interchange with other standards

Regarding other standards, please look also at the session "What role can RDA/RDF play in the transition to linked library data?" at <https://2021.bfwe.eu> conference (third day) and then the following session on Data Exchange.

Question: Do we need a separate group to discuss MARC/BIBFRAME as well?

There is interest in trimming MARC. Many fields are VERY low use and OCLC did some research on this: "MARC Usage in WorldCat", archived at <http://roytenant.com/proto/groundtruthing/>

We need to keep in mind that some fields may be of lower use because they are either newer or only important to a specific community.

The PCC has not yet looked in-depth into the BIBFRAME to MARC conversion.

Next steps: This will be brought back to the PCC Policy Committee as well.

Sharing of outcomes:

This summary will be posted widely.

A report of this meeting is on the agenda of the [BIBFRAME Workshop in Europe](#)

Information about the PCC and BIBFRAME is available here:

<https://www.loc.gov/aba/pcc/bibframe/bibframe-and-pcc.html>