Submitted on: May 18th, 2018

Who Will be Our bf: Comparing techniques for conversion from MARC to BIBFRAME

Ian Bigelow
Bibliographic Services, University of Alberta Libraries, Edmonton, Canada.
E-mail address: bigelow@ualberta.ca

Danoosh Davoodi
Bibliographic Services, University of Alberta Libraries, Edmonton, Canada.
E-mail address: danoosh.davoodi@ualberta.ca

Sharon Farnel
Bibliographic Services, University of Alberta Libraries, Edmonton, Canada.
E-mail address: sharon.farnel@ualberta.ca

Abigail Sparling
Bibliographic Services, University of Alberta Libraries, Edmonton, Canada.
E-mail address: abigail.sparling@ualberta.ca

Abstract:

The University of Alberta Libraries is actively ramping up for linked data implementation through local experimentation, research and external partnerships. Though BIBFRAME is still in development, several transformation tools have already been created, and with many libraries planning for a future move to linked data it would seem timely to compare approaches to moving legacy MARC data to BIBFRAME. Setting aside the question of whether BIBFRAME should be the vehicle for libraries to move to linked data, this investigation is aimed at comparing two approaches to converting MARC to BIBFRAME 2.0:

1. *University of Alberta Libraries (UAL) locally developed process based on the Library of Congress MARC to BIBFRAME XSLT: An XSLT 1.0 application aimed at converting MARC to RDF/XML released in March 2017.*
2. *Casalini SHARE Virtual Discovery Environment: A project by Casalini Libri and @Cult to develop a linked data discovery environment, including a conversion tool for MARC to BIBFRAME 2.0 RDF.*

Keywords: BIBFRAME, Cataloguing, Linked data, MARC, Metadata.

Introduction

In her article on the development of BIBFRAME (Bibliographic Framework), McCallum (2017) reflects on the current state of the transition of standards in libraries:

> In the 1960's and 1970's the AACR [Anglo-American Cataloguing Rules] cataloguing rules and MARC [Machine-Readable Cataloging] 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 [Resource Description and Access] and BIBFRAME. (p. 84)

Forty years is a long time, yet it is also worth considering that it has been more than 16 years since Tennant (2002) wrote *MARC Must Die*, and just over 10 years since Berners Lee (2007) first pitched the idea of the semantic web (or "giant global graph"). Moreover, it has been seven years since the Library of Congress (LC) launched BIBFRAME. Progress on the development of a replacement for MARC21 has been slow, but we now appear to be approaching a tipping point, and a truly exciting time where viable non-MARC21data alternatives exist.

Library of Congress is currently working through its second BIBFRAME pilot, bridging from the first phase (August 2015 to March 2016), and as of June 2017 has been working on:

> testing of the input of non-Latin scripts for description with no corresponding romanization, testing of authority descriptions for Agents, and a fuller level of interaction with a live BIBFRAME database, consisting of a complete BIBFRAME conversion of the Library of Congress bibliographic file (Library of Congress, 2017).

Given the importance of the Library of Congress for shared efforts in bibliographic description, their commitment to transitioning from MARC21 to BIBFRAME is significant. With many libraries planning for a future move to linked data it would seem timely to compare approaches to moving legacy MARC21 data to BIBFRAME and identify both alignment and gaps with current cataloguing standards. This investigation is aimed at comparing two approaches to converting MARC21 to BIBFRAME 2.0:

1. University of Alberta Libraries (UAL) locally developed process based on the Library of Congress MARC21 to BIBFRAME2 XSLT (Extensible Stylesheet Language Transformations): An XSLT 1.0 application aimed at converting MARC21 to RDF (Resource Description Framework) XML (Extensible Markup Language) released in March 2017.
2. Casalini SHARE Virtual Discovery Environment: A project by Casalini Libri and @Cult to develop a linked data discovery environment, including a conversion tool for MARC21 to BIBFRAME 2.0 RDF.

To provide context an overview will be given to the Casalini SHARE-VDE project as well as our deployment of the LC BIBFRAME converter. Through the comparison and analysis of these transformation tools several topics will be explored. First, there will be a more general analysis of both conversion methods, examining the underpinning modelling for each, including entity reconciliation processes.

Next, we will explore how well the LC converter and the SHARE-VDE project handle the responsible conversion of legacy library data. The paper examines how well varying content standards (AACR vs. RDA for example) are dealt with through conversion, along with a discussion of findings related to appropriate conversion of data elements.

With the possibility of moving away from MARC21 on the horizon, another timely question is how much development should be put into further MARC21 enhancements. In particular, the topic of URI (Uniform Resource Identifier) enrichment of MARC21 data has become an important one as we plan for transitioning. Findings that shed light on the importance of URIs in MARC are discussed alongside determining the correct balance between pre and post MARC21 to Linked Data (LD) conversion URI enrichment.

Finally, the paper will wrap up by drawing together general findings that point towards the development of workflows that provide a way forward for libraries to truly shift from records to user friendly, web designed data.

Literature Review

There is general consensus within the library community that MARC21 must be replaced with a new encoding standard in order to propel library data onto the semantic web and break down the library-specific data silos (Dull, 2016). Linked data generally, and BIBFRAME more specifically, are lauded for their potential to accomplish this work by moving data beyond MARC21's flat record structure, more effectively identifying bibliographic entities and expressing bibliographic relationships, all while making library data visible to a larger audience on the web (El-Sherbini, 2018; Fallgren, Reynolds & Kaplan, 2014; Hardesty, 2016; Kelley, 2016).

As Library of Congress BIBFRAME 2.0 development progresses, the library community has responded by analysing the BIBFRAME model's ability to effectively represent MARC21 and other non-traditional library data. At the highest level, studies have investigated how BIBFRAME's modelling aligns with accepted library standards such as the FRBR (Functional Requirements for Bibliographic Description) conceptual model (Baker, Coyle & Petiya, 2014; Zapounidou, Sfakakis & Papatheodorou, 2017) and the RDA content standard (Taniguchi, 2017). Although the BIBFRAME 2.0 model moves "bibliographic data closer to the FRBR/RDA view of bibliographic data" (McCallum, 2017, p. 79), studies continue to call for increased granularity in the BIBFRAME model and vocabulary in order to effectively accommodate FRBR and RDA data (Baker, Coyle & Petiya, 2014; Balster, Rendall & Shrader, 2017; Shoichi, 2017; Taniguchi, 2017; Zapounidou, Sfakakis & Papatheodorou, 2017).

Library data spans multiple domains, which include numerous specialized formats, and for this reason the Andrew W. Mellon Foundation funded the LD4P (Linked Data for Production) family of projects in 2016 to extend the BIBFRAME ontology to better represent domains such as art, performed music, and cartographic/geospatial materials (Linked Data for Production, 2016). While projects such as the LD4P ontology extensions demonstrate that the conversion of MARC21 data to BIBFRAME is complex for specialized domain-specific formats, recent scholarly studies and PCC task groups have also analyzed the BIBFRAME model's ability to effectively represent formats common across the library community. In these studies particular attention is paid to monographs and serials and their format-specific entities and relationships. In their study, Zapounidou, Sfakakis, and Papatheodorou (2017) investigate data interoperability and integration issues associated with converting MARC21 monograph records to BIBFRAME 1.0. The PCC task groups responsible for mapping both the CONSER and BIBCO standard records (CSR and BSR respectively) to BIBFRAME 2.0 have also been instrumental in identifying format-specific mapping drawbacks for both monographic and serial resources. In addition to mapping discrepancies and shortfalls for these formats in BIBFRAME 2.0, the final reports also highlight joint areas for improvement for BIBFRAME 2.0 such as the need for greater URI usage, additional format-specific note types, and further development around administrative metadata representation (Balster, Rendall & Shrader, 2017; BIBCO Mapping BSR to BIBFRAME 2.0 Group, 2017).

Though studies analysing BIBFRAME modelling continue to emerge, little has been published analysing conversion workflows, particularly the resulting alignment issues and gaps and the success of entity reconciliation. Xu, Hess and Akerman (2018) have begun to bridge this gap by providing a broad analysis of BIBFRAME 2.0 mappings, conversion specifications and pre/post conversion data cleanup requirements. Acknowledging the need for more studies like their own, Xu, Hess and Akerman (2018) call for the creation of a BIBFRAME development quality control program. As more libraries experiment with MARC21 to BIBFRAME 2.0 conversion a similar call has been made by Suominen and Hyvonen (2017) who advocate for increased collaboration around analysis in order to promote BIBFRAME development that is "more open, transparent and organized" (p.11).

While calls for increased collaboration on analysis grow, work on implementation has expanded beyond LC. The BIBFRAME 2.0 Implementation Register (Library of Congress, n.d.) provides quick reference to efforts towards transitioning from MARC21 to BIBFRAME, several of which are worth noting. The Finish National Bibliography has been working on an implementation of BIBFRAME with Schema.org (Suominen, 2017) and the Swedish Union Catalogue is scheduled for a full implementation of BIBFRAME 2.0 this summer (Lindström, 2018). In addition, European BIBFRAME workshops have been established to create a "forum for sharing knowledge about practice of, production with and planning of BIBFRAME implementation" (Andresen, email posted to BIBFRAME list, April 12, 2018), with further emphasis on the need for a stable, more international BIBFRAME community (Andresen, 2018). One interesting result of initial workshops was the publication of *BIBFRAME Expectations for ILS tenders* (Organizer Group, 2018). At the LD4 (Linked Data for ...) Workshops in spring 2018 it was also announced that PCC (Program for Cooperative Cataloging) and LD4 would be working on the creation of a shared, cloud-based sandbox triplestore environment for experimentation and early implementation to support the creation of original resource descriptions using BIBFRAME (Baxmeyer & Billey, 2018). This project, alongside the Casalini SHARE-VDE project which could see collections with upwards of 100 million bibliographic records and associated institutions converting to and working with BIBFRAME, points to a major shift in the nature of work from a phase of experimentation to one of early implementation.

Given that the community for BIBFRAME experimentation and implementation is growing, more work is needed to create shared spaces for BIBFRAME analysis, whether it be to report modelling issues, suggest vocabulary additions, or share improved conversion workflows.

Summary and Scope of Work

The following analysis examines both the data and overall processes from the Casalini SHARE-VDE project as well as similar processes developed at the UAL using the LC converter. This project started with initial experimentation with the conversion of MARC21 data to BIBFRAME, entity enrichment and reconciliation, but as the SHARE-VDE project developed and our own processes improved the need for wider analysis became apparent. Working towards the development of both vendor supported and internal workflows for conversion to BIBFRAME data is a very large topic and has required us to scope our analysis to focus on BSR and CSR core elements. Similarly, it is worth noting that both the Casalini SHARE-VDE project and use of the LC converter at the UAL are works in process, and this paper should only be seen as a report on work thus far.

Casalini SHARE-VDE

The Casalini SHARE-VDE project is an @Cult and Casalini Libri partnership. The project, and more specifically, the conversion tool, started as the "ALIADA project, co-financed by the European Union in 2013-2015, [which] originally applied the Linked Data paradigm using FRBRoo based ontologies" (Casalini, 2017). Bridging from this work

> a prototype of a virtual discovery environment with a three BIBFRAME layer architecture (Person/Work, Instance, Item) has been established through the individual processes of analysis, entity identification and reconciliation, conversion and publication of data from MARC21 to RDF, within the context of libraries with different systems, habits and cataloguing traditions (Casalini Libri, 2017).

Thus far the project has progressed through two phases, with participation from the Library of Congress and fifteen large research libraries in Canada and the United States. Initially, imprint data for the years 1985 and 2015 was used for processing, returned to libraries in enriched MARC21 and BIBFRAME, and included for use in the virtual discovery environment. Through the second phase, participating libraries received data back for their entire collections in BIBFRAME and enriched MARC21. A third phase is currently being defined with the intent to: publish all participants data in the SHARE-VDE platform; support batch updates through library exports, as well as editing and original cataloguing within the platform; develop return of enhanced data back to libraries; and create options for generating reports on the data.

The below diagram (Casalini, 2017) outlines the overall process for the conversion, entity reconciliation and enrichment used by the Casalini SHARE-VDE project. The Casalini process starts with entity detection/enrichment/reconciliation with the MARC21 data. The initial Authify process takes the MARC21 through a clustering tool, enriching records with URIs from a range of sources (LC, BNF (Bibliothèque nationale de France), ISNI (International Standard Name Identifier), VIAF (Virtual International Authority File), Wikidata…). Through this process an aggregation of entities takes place where clusters are formed and assigned unique identifiers. This information is then saved and an enriched MARC21 file is created.

Picking up from the Authify process, the enriched files act as the input for another process called Lodify, which translates the MARC21 into triples and uploads them to the Blazegraph triple store. As outlined by Casalini (2017), "Lodify is a module that automates the translation and publication of bibliographic data in RDF according to BIBFRAME 2.0 ontology in a linearly scalable way."

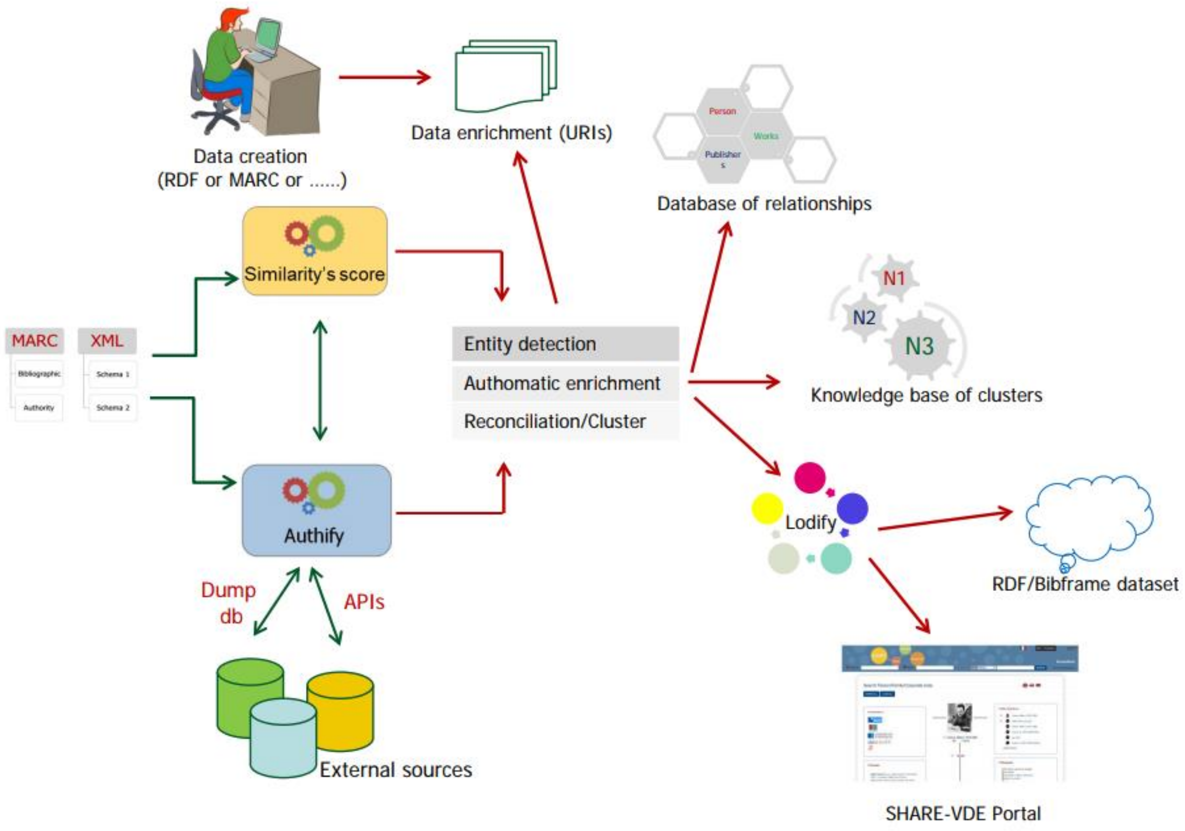


Figure 1. The SHARE-VDE processes (Casalini, 2017).

As a vendor supported workflow the SHARE-VDE project is very interesting. Existing systems (ILS, Discovery) as well as workflows for working with existing data have been long established and transitioning will take time and significant resources. Having an external agent, driven by community values, that can proof new discovery concepts for linked data and help with some of the heavy lifting for data conversion, reconciliation and enrichment is very useful. Given the scale of the SHARE-VDE project, with a large number of major research libraries working towards linked data implementation, the importance of work to analyze associated processes and metadata become obvious.

Local Conversion, Reconciliation and Enrichment Process

Prior to participation in the SHARE-VDE project, UAL was experimenting with the LC conversion tool. As the previous section highlights, SHARE-VDE is far more than an XSLT converter for BIBFRAME. This difference ignited curiosity about how practical it would be to create a more complete workflow around the LC converter without vendor support. The development of these local processes to date are outlined below.

UAL's local process starts with "*.mrc*" files. The .mrc file is converted to MARCXML format using the pymarc library (Summers, n.d.) with python. The pymarc library was modified to facilitate local requirements, such as error handling for CJK (Chinese, Japanese, Korean) encodings in the MARC21 data. The text from 245$a was used as filename when writing the "*.xml*" files.

From this point on, we experimented with two different approaches. The initial approach used Oxygen XML Editor (SyncRO Soft, n.d.) and OpenRefine (OpenRefine Community, n.d.) third party software to perform XSLT transformations, data clean-up and reconciliation, while the revised approach used only python native libraries and processes to perform all of the mentioned tasks.

Initial process

Using the marc2bibframe2 XSLT converter package (Library of Congress, n.d.), MARC/XML files were transformed into BIBFRAME format. With the use of another XSL stylesheet all names from the BIBFRAME.xml that are in the tag “bf:Agent“ and has a “rdf:about” attribute were extracted into a tab separated file, with each “rdf:about” attribute’s value saved as a unique key. This unique “example.org” URI would allow us to easily find the position of a name when we were trying to ingest the enriched URI. Finally, the stylesheet also extracted name types from “rdf:resource” attribute of “rdf:type” within the “bf:Agent”. The name types would allow us to send specific queries to the API (Application Programming Interface) for better results (e.g. Corporate vs Personal names).

```
<bf:Agent rdf:about="http://example.org/6815285#Agent100-13">
  <rdf:type rdf:resource="http://id.loc.gov/ontologies/bibframe/Person"/>
   ...
  <rdfs:label>Veksner, Simon,</rdfs:label>
</bf:Agent>
```

Veksner, Simon,	**http://id.loc.gov/ontologies/bibframe/Person**	http://example.org/6815285#Agent100-13

All the extracted names were saved to a “*.tsv*” file and imported to OpenRefine for text clean-up and reconciliation. For each name an attempt was made to retrieve the identifier from LC name authorities database as well as from VIAF.

The results returned from the APIs were analyzed and ordered (based on their similarity score) and the result with the highest score was selected. Once this process was completed, the data was exported from OpenRefine in “.tsv” format. This file was then fed to another XSL stylesheet which used the “example.org” URI to ingest the retrieved URI back into the BIBFRAME.xml file.

Name	**Ingest_key**	**LC**	**VIAF**
Veksner, Simon,	http://example.org/6815285#Agent100-13	no2011039513	169080997

```
<bf:Agent rdf:about="http://id.loc.gov/authorities/names/no2011039513">
   ...
   <rdfs:label>Veksner, Simon,</rdfs:label>
   <bf:identifiedBy>
      <bf:Identifier>
```

```
            <rdf:value rdf:about="http://viaf.org/viaf/169080997"/>
        </bf:Identifier>
    </bf:identifiedBy>
</bf:Agent>
```

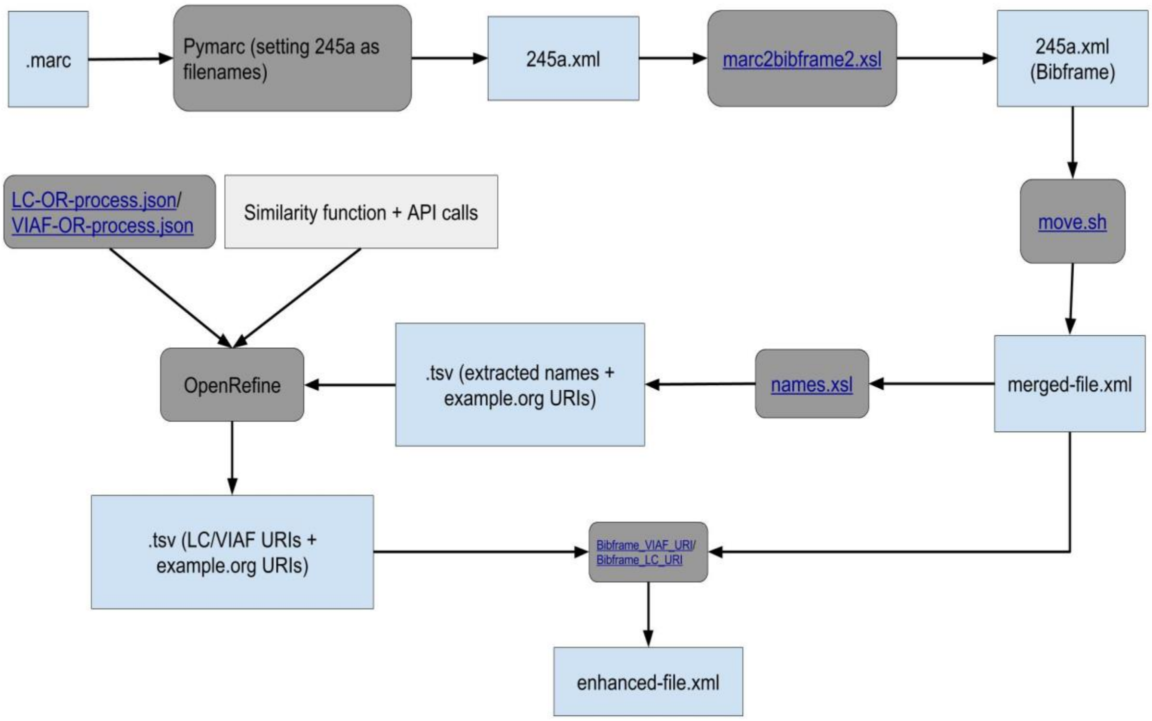

.marc
Pymarc (setting 245a as filenames)
245a.xml
marc2bibframe2.xsl
245a.xml (Bibframe)
LC-OR-process.json/ VIAF-OR-process.json
Similarity function + API calls
move.sh
OpenRefine
.tsv (extracted names + example.org URIs)
names.xsl
merged-file.xml
.tsv (LC/VIAF URIs + example.org URIs)
Bibframe_VIAF_URI/ Bibframe_LC_URI
enhanced-file.xml

Figure 2. Initial process flow chart.

Revised process

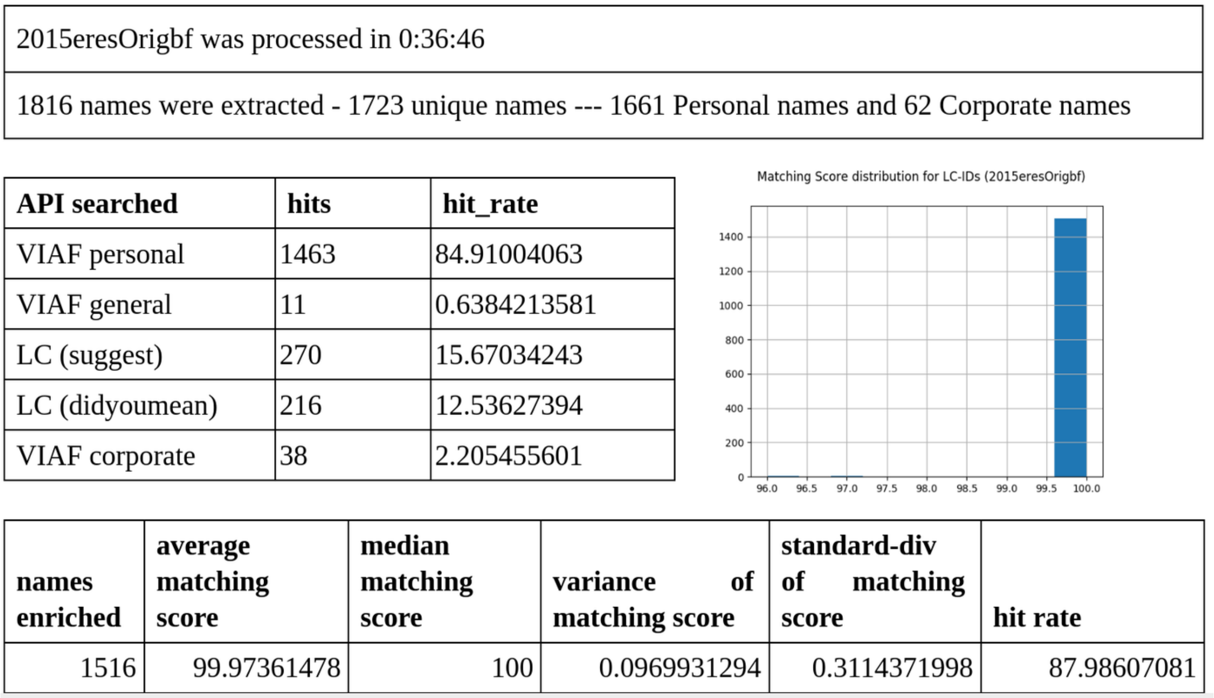
This process (Davoodi, 2018) was developed fully in a python environment. The modified pymarc library was used to transform MARC21 to MARC/XML, then using a python native library for working with the XML file (Python Software Foundation, n.d.) the marc2bibframe2 stylesheet was applied to the MARCXML files to transform them to BIBFRAME. Names (name, name type, URI) were then extracted from the BIBFRAME files using the same python library. For each extracted name the program built queries and sent them to the corresponding API (2 from LC and 3 from VIAF):

This revised process gave us more control over API calls. In the initial process, if a certain name was repeated in the BIBFRAME file, multiple API calls were sent, all of which returned identical results. We were able to improve our processing time substantially in the revised process by creating a checksum for names extracted from the BIBFRAME file and sending only one API call for a repeated name.

Work on implementing a webapp version for this process which would allow users to upload a .marc (or MARC/XML or BIBFRAME) file and get a BIBFRAME file enriched with URI from a selectable number of API(s) is currently underway.

Benefits of the revised process include:

- Start/Stop the process with one command
- Reduced processing time (a 10,000 record BIBFRAME file takes approximately 3h 50m)
- Flexibility to choose the API(s) to be searched (e.g. only search VIAF for personal names)
- Generate statistical data and visual graphs



2015eresOrigbf was processed in 0:36:46
1816 names were extracted - 1723 unique names --- 1661 Personal names and 62 Corporate names

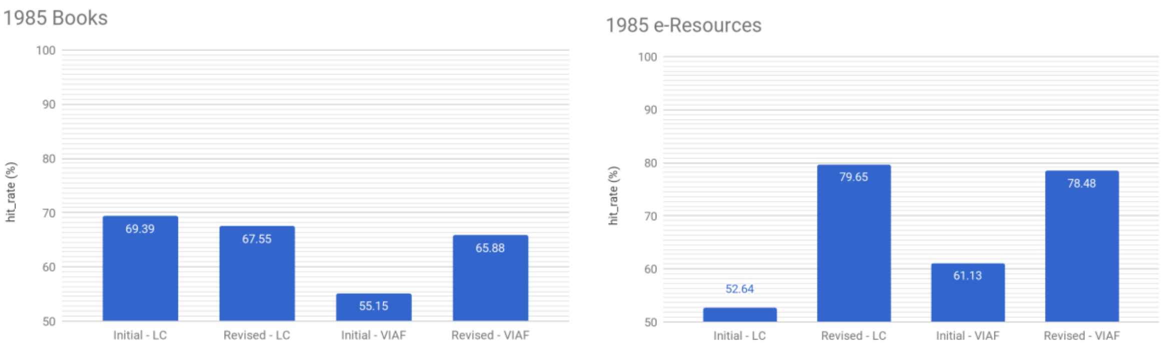
API searched	hits	hit_rate
VIAF personal	1463	84.91004063
VIAF general	11	0.6384213581
LC (suggest)	270	15.67034243
LC (didyoumean)	216	12.53627394
VIAF corporate	38	2.205455601

names enriched	average matching score	median matching score	variance of matching score	standard-div of matching score	hit rate
1516	99.97361478	100	0.0969931294	0.3114371998	87.98607081

Figure 3. Screenshot of statistical output.

Comparison

Figure 4 represents success rates of the initial and revised process on six sample dataset.



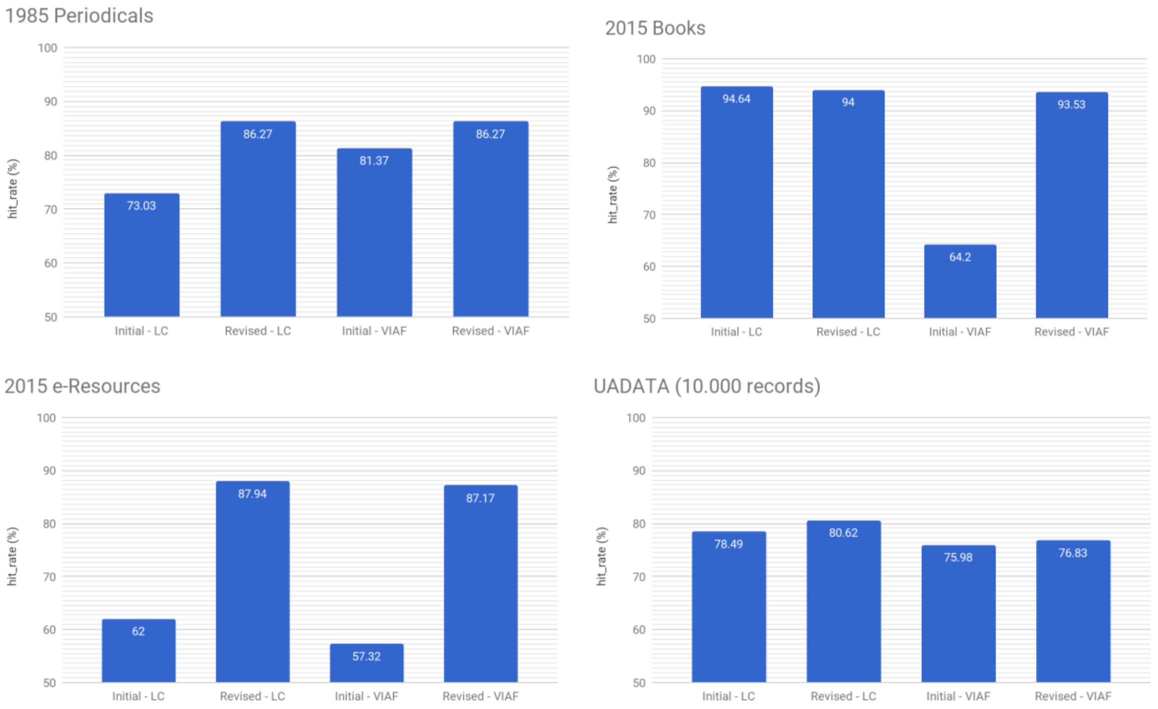
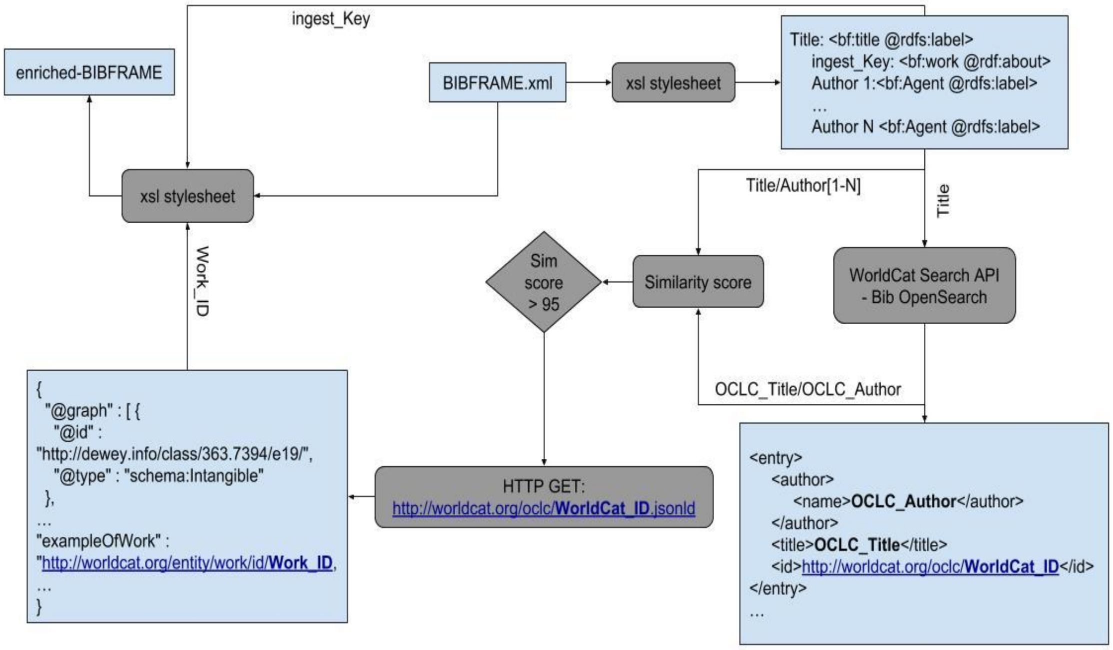


Figure 4. Initial vs. revised process success rate.

OCLC Work IDs

We are currently working on developing a process for extracting Work IDs form OCLC. This process uses a new XSL stylesheet that extracts title-author pairs from BIBFRAME files. The pairs are then used to send a query to WorldCat's Search API (Bib OpenSearch). The results returned are then processed to extract the "id" field for the record that best matches our title-author pair. This "id", which is an OCLC Record ID, is sent to OCLC's open API to retrieve "jsonld" (JSON for linking data) results which contain the Work ID. At the time of writing this process is a proof of concept. We intend to further develop elements used for searching APIs (e.g. date of publication) which will improve the accuracy of matching Work IDs.



```
Title: <bf:title @rdfs:label>
   ingest_Key: <bf:work @rdf:about>
   Author 1:<bf:Agent @rdfs:label>
   ...
   Author N <bf:Agent @rdfs:label>
```

```
{
  "@graph" : [ {
    "@id" :
"http://dewey.info/class/363.7394/e19/",
    "@type" : "schema:Intangible"
  },
...
"exampleOfWork" :
"http://worldcat.org/entity/work/id/Work_ID,
...
}
```

```
<entry>
   <author>
      <name>OCLC_Author</author>
   </author>
   <title>OCLC_Title</title>
   <id>http://worldcat.org/oclc/WorldCat_ID</id>
</entry>
...
```

Figure 5. OCLC Work ID enhancement process.

There are certainly aspects of SHARE-VDE and the processes at UAL that aim to accomplish the same thing. Still, our hope is that community and/or vendor supported processes, and local development can work in concert to assist libraries through this transition.

Analysis of Data Through Conversion

Through this section it is worth remembering that this is an analysis of data through conversion rather than an investigation of how well the BIBFRAME model fits with a descriptive standard (e.g. RDA). In addition, because our analysis is still underway what is reported here is not comprehensive, but rather serves to highlight some interesting or surprising results.

Clarification about the scope of analysis is also important here. For this project the focus was on BSR and CSR elements and highlights of this are included below. While this is an important starting point in the identification of alignment issues and gaps between core elements of RDA description and BIBFRAME conversion, further analysis work will be required.

General points of interest

The UAL database, like many others, includes RDA records alongside hybrids, AACR2 and pre-AACR2 records. Creating mappings that take various standards into account poses challenges and this was evident in the analysis of both conversion processes.

Using the LC converter, place of publication is mapped as text from the 260/264, with the corresponding URI generated from the location code in the 008. Publisher name is mapped as text from the 260/264, and date of publication is mapped from the 260/264/008. The Casalini process assigns a SHARE-specific URI for place of publication and publisher name based on the text in the 260/264, and date of publication is mapped from the 260/264. Data for statements of manufacture, production, and distribution are mapped in a similar fashion to date of publication. Of note is that both the LC and Casalini processes are able to map data in pre RDA records to publication statements only, based on the data in the 260.

Also of note, the LC process strips brackets and other punctuation when mapping to places, agents, and dates, resulting in loss of important nuance for users working with the data. In contrast, the Casalini process retains brackets and other punctuation, and also generates a SHARE-specific URI for the activity that clusters similar date expressions. For example: “1958.”, “[1958]” and “1958]” are all clustered.

Relationship designators are important components of RDA records that provide additional information for users of that data. With the LC process, records that do not have these designators map agents appropriately (from the MARC21 1xx and 7xx) but with each assigned the role of contributor (with the role URI coming from LC). If the MARC21 records have been enhanced with URIs (such as from LC or VIAF) for the agent prior to conversion, the URI is mapped appropriately.

With the Casalini process, agents are similarly mapped from the MARC21 1xx and 7xx, and the records that have relationship designators are mapped with the appropriate URI (from LC). Agents are mapped with URIs from LC and/or VIAF as appropriate. In cases where there are no relationships designators, Casalini uses a relator term detection process to attempt to determine that role. “Starting from a Marc21 record (whatever is the specific dialect) the system analyses all (configured) tags that contain a name and, for each of them, tries to figure out (using the statements of responsibility of the input record or other parts of the record) what is the corresponding role within the work represented by the given record.” (Casalini Libri, 2017). Interestingly, in such cases the agent is not assigned a URI from LC and/or VIAF but rather a SHARE-specific one. However, the Casalini process does incorporate some modeling using owl#sameAs to enhance potential linkages between agent URIs. For example:

<http://share-vde.org/sharevde/rdfBibframe/Agent/78151>
<http://www.w3.org/2002/07/owl#sameAs> <http://www.wikidata.org/entity/Q200580>

It was also noted that while most core elements are converted (though in varying methods), several alignment issues were found. In both the LC and Casalini processes, mapping of content and carrier types depends on the existence of MARC21 336 and 338, respectively. When present, these are mapped as both text and URIs. In cases where these fields are not present (i.e., pre RDA records), these are absent in the converted data. This absence raises important questions about the usability of the converted data.

Flavours of BIBFRAME

With the LC process, preferred title is generated as one would expect from MARC21 130/240, or 245 when these are not present. The Casalini process utilizes MARC 130/240 information along with dates to create an equivalent of “mainTitle” clusters in the form of http://share-vde.org/sharevde/rdfBibframe2/Title. This is a significantly different approach from LC.

With the LC process, form of work is mapped appropriately from MARC21 008 positions 24-27; text and LC URI are generated based on the code. The Casalini process also maps form of work from MARC21 008 positions 24-27. However, as with other elements, the URIs generated are SHARE-specific rather than taken from an existing vocabulary such as LC.

Mapping of subject relationships differs between the LC and Casalini processes, and is somewhat idiosyncratic in both, but in different ways. In the LC process, subjects are given unique URIs but of an 'example.org' pattern. If the MARC21 records have been enriched with URIs prior to conversion, the URIs map appropriately. With the Casalini process, URIs are generated appropriately, but they are SHARE-specific rather than those from existing vocabularies such as LC or VIAF. In addition, Casalini generates two triples, one using the subject literal and the other the subject URI.

The differences between the modeling for LC BIBFRAME and the SHARE-VDE project raise questions about standardization and interchange, especially as other implementations such as the Swedish Union Catalogue (National Library of Sweden, 2008) develop. It would be very helpful to have vocabularies and modeling published for SHARE to facilitate further analysis.

Pre vs Post MARC21 to BIBFRAME entity URI enrichment

The PCC Task Group on URI in MARC21 has been working to develop ways to incorporate URIs into MARC21 data. At the American Library Association (ALA) Annual Conference in June 2017, use of $0, $1, $4 and 758 were approved by the MARC Advisory Committee (MARC Advisory Committee, 2017).

There are tools and vendor services which can help libraries populate URIs in MARC21 data based on some of these specifications, but one question that came up during our investigation was whether effort was best spent including this data in the MARC21 format or enriching post conversion to BIBFRAME.

Many URIs that can be included in the MARC21 format are related to controlled vocabularies. In both the UAL processes and the Casalini SHARE-VDE project, the efficacy of populating BIBFRAME with URIs based on built-in mappings for controlled vocabulary lists would suggest that it only makes sense to include these in the MARC21 data if needed for current processes (use of MARC21), rather than for work towards linked data conversion.

In contrast, if one has a workflow for accurately populating MARC21 with URI for authorized headings, this could be a beneficial way to enhance conversion when compared to post conversion enrichment methods using automated matching algorithms based on degrees of confidence for matching.

With respect to the analysis of URIs through conversion, at the time of our analysis the LC converter did not make use of $1, $4 or 758, highlighting the need the need for an updated XSLT. In contrast, the MARC21 returned from Casalini makes use of all the new fields and subfields for URIs in MARC21, and the BIBFRAME data also includes both work URIs and additional relationship between URIs built into the data.

Of particular interest is Casalini's use of work identifiers. As we begin thinking about updates to parallel MARC21 and BIBFRAME data sets, the question about how to cluster work information through conversion and facilitate ongoing updates to BIBFRAME data arises. MARC21 758 was created to give a place to record primary resource identifiers, including work URI. Until we can move past the need for our MARC21 data, it will be important to include work URI in MARC21 to support ongoing conversions to BIBFRAME data. Given that there are few work URI vocabularies to leverage, the development of these for the SHARE-VDE project is significant on its own.

Overview of data analysis

Through the analysis for this project, several key observations were made about the conversion processes: there are notable issues with conversion preserving RDA core elements; multiple previous standards (pre AACR2, AACR2, RDA) presents challenges for the conversion; the Casalini conversion tool represents a significantly different flavour of BIBFRAME than the LC version; there is variable usage of existing vocabularies; and processes highlight the need for further development of URI vocabularies and best practices for entity enrichment in MARC21 versus post BIBFRAME conversion.

Both the LC and the Casalini processes overall do quite well with mapping, and mapping accurately, with respect to RDA Core elements. However, some potential problematic areas do exist and should be addressed to improve the conversion processes and the resulting data. For example, important punctuation associated with production, publication, manufacture, and distribution not being retained in converted data; reliance on the presence of MARC21 336 and 338 for content and carrier type; lack of handling of data in MARC21 758. The LC process could be enhanced to better handle data that lacks relationship designators, and the Casalini process could be enhanced by making greater use of existing vocabularies in addition to SHARE-specific ones.

The conversion processes highlight the challenges faced when dealing with data that has been created and encoded according to older standards (e.g., AACR vs. RDA). This poses the question of how much remediation of the data, and of what kind, should be carried out prior to conversion. For both the LC and Casalini processes, no work was done with the data prior to the process, and so the analysis described above will be essential information in making decisions about remediation efforts.

Conclusion

While BIBFRAME data created through the LC converter and the Casalini processes is functional, further analysis is needed and some areas should be addressed. At the same time, both workflows have their own strengths. SHARE-VDE offers the support of a vendor relationship as well as community development, has an experimental discovery tool established, and involves more complex text analysis and clustering. By contrast, the UAL process has the potential to create a single tool for conversion, reconciliation and enrichment that could be open to the library community, and demonstrates that capabilities to develop such processes exist outside of vendor contexts.

Many of the large scale projects being discussed in this paper are also reaching a point where continued work means a shift away from experimentation to implementation with incremental change. With implementation, work must be bridged beyond explorations by a few staff with limited resources to institution (and community) wide projects involving the commitment of staff and development of production technology.

As more libraries implement BIBFRAME or other linked data applications, and more data exists in alternate formats, the pace of change will likely increase. Framed by this, the data and workflows from both processes may need to be measured as outlined by Lindström (2018), not by whether they are perfect or finished, but by whether they are real and workable.

It seems apparent that major changes are in store for resource description in libraries. Those involved with cataloguing and metadata will need to maintain expertise in existing standards and formats, while simultaneously becoming familiar with linked data tools and workflows. Waiting until we have no choice to transition will not foster the desired community collaboration around BIBFRAME development or support a smooth implementation and it is worth considering training and development for cataloguers and other library staff now.

References

Andresen, L. (2018) European BIBFRAME Community. LD4 Workshop, Stanford, May 1, 2018. Retrieved from https://drive.google.com/drive/folders/18h0ijn_ZAln9QihQFdAbi5ViTAhH-Olq

Baker, T., Coyle, K., & Petiya, S. (2014). Multi-entity models of resource description in the Semantic Web. *Library Hi Tech*, *32*(4), 562-582. doi:10.1108/LHT-08-2014-0081.

Balster, K., Rendall, R., & Shrader, T. (2018) Linked Serial Data: Mapping the CONSER Standard Record to BIBFRAME, *Cataloging & Classification Quarterly*, 56:2-3, 251-261, DOI: 10.1080/01639374.2017.1364316

Baxmeyer, J., & Billey, A. (2018) The Role of the Program for Cooperative Cataloging (PCC) in Linked Data Implementation. LD4 Workshop, Stanford, May 1, 2018. Retrieved from https://docs.google.com/presentation/d/19k-EghUOysI38I7Xq3TJ_JLzfELRw6bHOCaboLklAz4/edit#slide=id.p

Berners Lee, T. (2007, November 21). Giant global graph [web log]. Retrieves from https://web.archive.org/web/20170426160611/http://dig.csail.mit.edu/breadcrumbs/node/215

BIBCO Mapping BSR to BIBFRAME 2.0 Group. (2017). BIBCO BF Mapping Spreadsheet. Retrieved from http://www.loc.gov/aba/pcc/bibframe/TaskGroups/BSR-PDF/BSRtoBIBFRAMEMapping.pdf

Casalini, M. (2017). BIBFRAME and linked data practices for the stewardship of research knowledge. *IFLA-Satellite-Meeting*, Berlin, August 16, 2017. Retrieved from https://www.ifla.org/files/assets/academic-and-research-libraries/conferences/casalini_bibframe_and_linked_data_practices.pdf

Casalini Libri. (2017). *The SHARE-VDE project*. Retrieved from http://share-vde.org/sharevde/clusters?l=en

Chiu, J. (n.d.). refine.codefork.com [computer software]. Retrieved from http://refine.codefork.com/

Davoodi, D. (2018). BIBFRAME Convertor [computer software]. Retrieved from https://github.com/ualbertalib/metadata/tree/master/metadata-wrangling/BIBFRAME/Convertor

Dull, M. E. (2016). Moving Metadata Forward with BIBFRAME: An Interview with Rebecca Guenther. *Serials Review, 42*(1), 65-69. doi:10.1080/00987913.2016.1141032

El-Sherbini, M. e. (2018). RDA implementation and the emergence of BIBFRAME. *JLIS.It, Italian Journal Of Library, Archives & Information Science, 9*(1), 66-82.

Fallgren, N., Lauruhn, M., Reynolds, R. R., & Kaplan, L. (2014). The Missing Link: The Evolving Current State of Linked Data for Serials. *Serials Librarian, 66*(1-4), 123-138.

Hardesty, J.L. (2016). Transitioning from XML to RDF: Considerations for an effective move towards Linked Data and the Semantic Web. *Information Technology And Libraries, Vol 35, Iss 1, Pp 51-64 (2016),* (1), 51. doi:10.6017/ital.v35i1.9182

Kelley, S. k. (2016). The Smaller Library Staff 's Perspective on BIBFRAME. *Technicalities, 36*(6), 8-11.

Library of Congress. (n.d.). *BIBFRAME 2.0 implementation register*. Retrieved from https://www.loc.gov/bibframe/implementation/register.html

Library of Congress. (n.d.). marc2bibframe2 [computer software]. Retrieved from https://github.com/lcnetdev/marc2bibframe2

Library of Congress. (2017). *BIBFRAME training at the Library of Congress*. Retrieved from https://www.loc.gov/catworkshop/bibframe/

Lindström, N. (2018). Gaining traction: Going into production with BIBFRAME 2 in the Swedish Union Catalogue. LD4 Workshop, Stanford, May 1, 2018. Retrieved from https://docs.google.com/presentation/d/1VJCU00jHcYptYciFrSgNJBROI9B3SiJ1cGm6NqV2FEM/edit#slide=id.p

Linked Data for Production. (2016). Linked Data for Production (LD4P). Retrieved from https://wiki.duraspace.org/pages/viewpage.action?pageId=74515029

MARC Advisory Committee. (2017). MAC meeting minutes. ALA Annual Meeting, Chicago, ILm June 24-25, 2017. Retrieved from https://www.loc.gov/marc/mac/minutes/an-17.html

McCallum, S. s. (2017). BIBFRAME Development. *JLIS.It, Italian Journal Of Library, Archives & Information Science, 8*(3), 71-85.

National Library of Sweden. (2008). LIBRIS. Retrieved from http://libris.kb.se/

OpenRefine Community. (n.d.). OpenRefine [computer software].

Organizer Group. (2018). BIBFRAME expectations for ILS tenders. Retrieved from https://www.landskerfi.is/sites/default/files/bibframe_expectations_for_ils_tenders.pdf

Python Software Foundation. (n.d.). The ElementTree XML API [computer software]. Retrieved from https://docs.python.org/2/library/xml.etree.elementtree.html

Shoichi, T. (2017) Is BIBFRAME 2.0 a Suitable Schema for Exchanging and Sharing Diverse Descriptive Metadata about Bibliographic Resources?,*Cataloging & Classification Quarterly*, *56*(1), 40-61, DOI: 10.1080/01639374.2017.1382643

Summers, E. (n.d.). pymarc [computer software]. Retrieved from https://github.com/edsu/pymarc

Suominen, O. (2017). Finnish National Bibliography Fennica as linked data. Semantic Web in Libraries, Hamburg, December 6, 2017. Retrieved from https://swib.org/swib17/slides/suominen_fennica.pdf

Suominen, O, & Hyvonen, N (2017). From MARC silos to Linked Data silos?. *O-Bib. Das Offene Bibliotheksjournal, Vol 4, Iss 2, Pp 1-13 (2017)*, (2), 1. doi:10.5282/o-bib/2017H2S1-13

SyncRO Soft. (n.d.) oXygen XML Editor [computer software].

Taniguchi, S. (2017). Examining BIBFRAME 2.0 from the Viewpoint of RDA Metadata Schema. *Cataloging & Classification Quarterly*, *55*(6), 387-412.

Tennant, R. (2012). MARC must die. *Library Journal*, 127(17), 26. Retrieved from https://lj.libraryjournal.com/2002/10/ljarchives/marc-must-die/#_

Xu, A., Hess, K. & Akerman, L. (2017) From MARC to BIBFRAME 2.0: Crosswalks, *Cataloging & Classification Quarterly*, *56*(2-3), 224-250, DOI: 10.1080/01639374.2017.1388326

Zapounidou, S., Sfakakis, M., & Papatheodorou, C. (2017). Representing and integrating bibliographic information into the Semantic Web: A comparison of four conceptual models. *Journal Of Information Science*, *43*(4), 525. doi:10.1177/0165551516650410