BIBFRAME Transformation for Enhanced Discovery

Qiang Jin, Jim Hahn, and Gretchen Croll

With support from an internal innovation grant from the University of Illinois Library at Urbana-Champaign, researchers transformed and enriched nearly 300,000 e-book records in their library catalog from Machine-Readable Cataloging (MARC) records to Bibliographic Framework (BIBFRAME) linked data resources. Researchers indexed the BIBFRAME resources online, and created two search interfaces for the discovery of BIBFRAME linked data. One result of the grant was the incorporation of BIBFRAME resources within an experimental Bento view of the linked library data for e-books. The end goal of this project is to provide enhanced discovery of library data, bringing like sets of content together in contemporary and easy to understand views assisting users in locating sets of associated bibliographic metadata.

The BIBFRAME model, the potential successor to the MARC data model, is an effort to transition the MARC 21 format to linked data. It was first introduced in the Library of Congress (LC) report, "Bibliographic Framework as a Web of Data: Linked Data Model and Supporting Services" in 2012.[1] BIBRAME can be situated within the context of semantic technologies that make possible contextual and interlinked resources on the broader web. The development of BIBFRAME is a response to the effects of online networked information, leveraging search engines, their impact on discovery of library collections, and the need for standardization of bibliographic resources as those resources move into linked data environments.

Background on BIBFRAME Development

To understand the BIBFRAME model, one must first explore common information modeling terminology, particularly the fundamental entity-relationship (ER) model. The BIBFRAME model is based on the ER model developed by Peter Chen in 1976.[2] There are three basic elements in the ER model: entities, attributes, and relationships. According to Chen, an entity is a "thing" that can be distinctly identified. Entities are the "things" about which we seek information. A specific person, company, or event is an example of an entity. A relationship is an association between instances of entities. Attributes are the data that we collect about the entities. For example, attributes of a person entity may include a first name, last name, birth date, and title. Relationships illustrate how instances of entities are related to one another. These broad concepts make up the conceptual underpinnings of the BIBFRAME model. The LC project page introducing BIBFRAME gives the following motivation for the model: "BIBFAME provides

Qiang Jin (qiangjin@illinois.edu) is the Authority Control Team Leader and Senior Coordinating Cataloger, University of Illinois at Urbana-Champaign. **Jim Hahn** (jimfhahn@gmail.com) is Orientation Services and Environments Librarian, University of Illinois at Urbana-Champaign. **Gretchen Croll** (gacroll34@gmail.com) is R&I Operations Specialist I, Orrick, Herrington & Sutcliffe LLP.

Manuscript submitted August 23, 2015; returned to authors November 6, 2015 for revision; revised manuscript submitted December 18, 2015; manuscript returned to authors on February 29, 2016 for additional revision; revised manuscript submitted March 29, 2016; paper accepted for publication April 1, 2016.

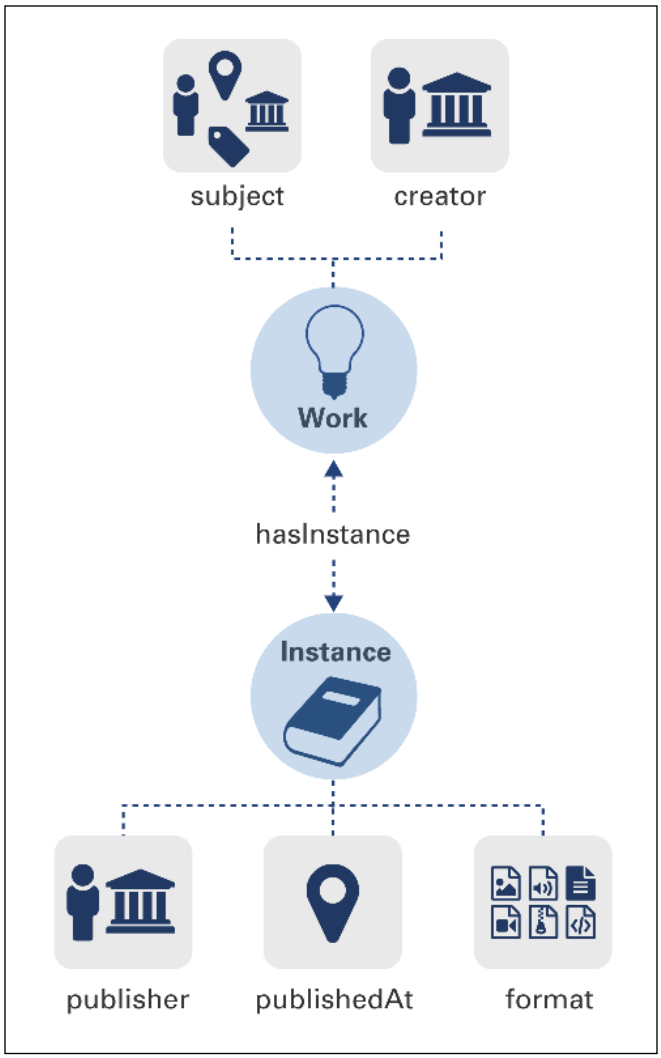

Figure 1. The BIBFRAME Model

a foundation for the future of bibliographic description, both on the web and in the broader networked world. BIBFRAME serves as a general model for expressing and connecting bibliographic data."[3] Figure 1 is an illustration of the BIBFRAME model.

The BIBFRAME data model descends from the Functional Requirements for Bibliographic Records (FRBR) conceptual model, but is not an exact implementation of that conceptual model. FRBR has four entity sets: work, expression, manifestation, and item. The FRBR entity sets work and expression are known in BIBFRAME as the entity *work*. FRBR entities manifestation and item are known in BIBFRAME as the entity *instance*.[4]

The BIBFRAME entity *work* is a resource reflecting the conceptual nature of the resource being cataloged. A BIBFRAME entity *instance* is a resource reflecting an individual, material embodiment of the work. The third BIBFRAME entity is *authority*. It includes FRBR group 2 entities for person, family, and corporate body, and FRBR group 3 entities for concept, object, event, and place. According to the report "Bibliographic Framework as a Web of Data," BIBFRAME authorities are not designed to replace existing authority efforts but rather provide a common abstraction layer over various different web based authority efforts to make them even more effective.[5] The fourth BIBFRAME entity is *annotation*. It is used to identify library holdings, cover art and reviews. BIBFRAME aims to publish and share library bibliographic and authority data via the web. It provides links to connect different pieces of information or resources and aspires to be a replacement for MARC. A key difference between MARC and BIBFRAME is that MARC presents bibliographic information as catalog records, which duplicates information across multiple records. As an example of this duplication, consider that many MARC records contain the same author's name, a repetition that is not a part of BIBFRAME since BIBFRAME emphasizes relationships between resources and can reference already existing links. Some of the relationships BIBFRAME holds include work-to-work relationships, work-to-instance relationships, instance-to-instance relationships, and work to authority relationships.

In 2013, LC issued a call encouraging libraries to test the BIBFRAME model. Inspired by a study testing the BIBFRAME model for audiovisual resources, the authors conducted an independent test focusing on e-books in the University of Illinois' online catalog.[6] Our hope was that we would be able to contribute to the revision of the BIBFRAME model for that specific format. It should be noted that at the time of this writing (late March 2016) there are now several proposed revisions to the BIBFRAME vocabulary, these draft documents are available as "BIBFRAME 2.0 Draft Specifications" on LC's BIBFRAME page.[7] Our project references the BIBFRAME specifications from 2014, and is one of fourteen projects registered at the LC BIBFRAME Implementation site as of March 2016.[8] The BIBFRAME implementation site includes projects from libraries in Cuba, England, Egypt, Germany, and the United States.

Innovation Grant Goals and Outcomes

The University of Illinois Library issues a biannual call for innovation proposals that will enable the library to explore new ways of working. Funding amounts vary, and have been supported up to $10,000. The funding source for the BIBFRAME grant provided graduate hourly student employees. The two graduate students who worked on this project were sourced from the Graduate School of Library and Information Science and the Department of Computer Science at the University of Illinois. Two professional tenured librarians led the investigation—first by way of manually derived

exploration of linked data transformation and enrichment, and after a model was developed for the e-book format within BIBFRAME, the transformation and enrichment was automated with original programming.

Objectives of the BIBFRAME innovation grant include the following:

- studying how to provide enhanced discovery of similar sets of content in the library system with the BIBFRAME model
- contributing a module of Bento-style search results in the BIBFRAME model[9]
- enriching the BIBFRAME model with linked data that connected to other open linked data projects
- writing a report on issues encountered and recommendations for e-book records in the BIBFRAME model.

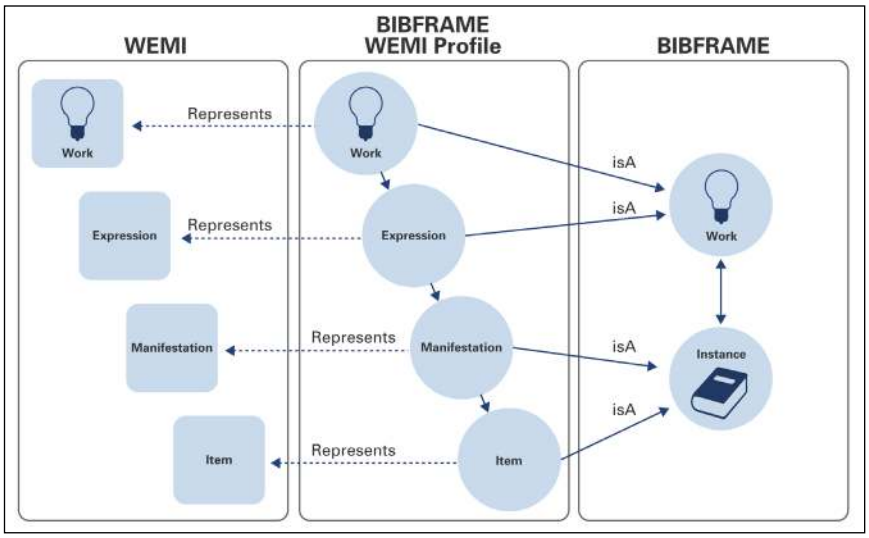

Figure 2. FRBR Work, Expression, Manifestation, and Instance mapped to BIBFRAME entities

By the conclusion of the innovation grant, the team transformed and enriched nearly 300,000 e-book records and has developed two prototype search interfaces. The two options for retrieval of linked data records include a Google Custom Search Engine that surfaces the structured data in the result list, and a Bento-style result layout for e-book search in addition to articles and other catalog data. The grant work is summarized on a project website.[10] The team has made the linked data enrichment code available through an online code repository.[11]

Literature Review

Enthusiasm for BIBFRAME has been high among several librarians whose work we review here, but since exemplars of large-scale implementations do not yet exist, the debate is still open as to whether BIBFRAME should be adopted. Among those reasons to pursue BIBFRAME projects is the concern that MARC may not be adequate to meet the demands of access and discovery on the World Wide Web and that a replacement is needed to leverage linked data like BIBFRAME. Kroeger provides an overview of literature leading to the BIBFRAME model.[12] She cites several sources including Tennant's 2002 paper "MARC Must Die."[13] In his paper, Tennant states that MARC has outlived its usefulness. MARC can no longer serve our users well. We reason however, that as the basis for a controlled identifier approach to sharing data, MARC has been instructive. Without adherence to standardization of controlled identifiers—of which MARC has been a leading exemplar—research such as the transformation and enrichment project described here would not be possible. The 2008 report "On the Record by the LC Working Group on the Future of Bibliographic Control," written by a group of well-known experts, argues that the library community needs to recognize that the World Wide Web is our technology platform and the appropriate platform for the delivery of our standards.[14] Since many e-book users primarily locate information resources through web searches, and not library discovery systems, we theorized that e-books are a natural target for BIBFRAME transformation and indexing on the web. Dean's paper indicates that we live in the age of Google, and our catalogs should reflect the information-seeking behavior of today's user, not the user of one hundred years ago.[15] Rollitt states in her paper that BIBFRAME might change libraries in a profound way.[16] It will link bibliographic data and will move bibliographic data to the web for access and management, which could generate new types of library services. Consider one library service available as a result of BIBFRAME transformation: locating e-books from your home library primarily from a search engine. This would be a welcome service of which users would seamlessly take advantage.

Pilot projects with BIBFRAME transformation are few, but among those early adopters and small prototypes, results have generally been favorable. Therani designed a project data model based on BIBFRAME, and transformed existing bibliographic data to BIBFRAME using relevant BIBFRAME vocabulary to implement linked data for a small collection at Harvard University library.[17] Therani's results indicated that BIBFRAME offers superior navigation control and access points for users to dynamically interact

with bibliographic data and concluded that users can find more information when bibliographic data are linked. The authors attempt in our Bento-style search result page of BIBFRAME data to assist users in finding sets of like items that are related to their initial search.

The University of Washington evaluated BIBFRAME and the Resource Description Framework (RDF) as carriers for RDA cataloging.[18] They mapped RDA core elements to BIBFRAME, and concluded that both RDA/RDF and BIBFRAME can represent library metadata as linked data. While comparing RDA/RDF with BIBFRAME, they discovered that RDA/RDF is stronger in series, notes, technical details of a resource, and inverse properties, while BIBFRAME is stronger in administrative metadata, identifiers, subject headings, holdings information, support for both transcription (literals), and Uniform Resource Identifiers (URIs). Note, however, that RDA is a content standard for resource description and access. Catalogers have been creating MARC records based on RDA for the last several years. BIBFRAME is a structural framework. RDA and RDF are connected by FRBR to define the primary entities and relationships. FRBR has been extended to a name authority model (FRAD, Functional Requirements for Authority Data), and a subject authority model (FRSAD, Functional Requirements for Subject Authority Data). RDA supports FRBR, FRAD, and FRSAD.

Related Projects

Schema.org is an initiative launched in 2011 by Bing, Google, and Yahoo to create, maintain, and support a common set of schemas for structured data markup on web pages, and beyond (see https://schema.org). Ronallo in his seminal piece "HTML5 Microdata and Schema.org" explained the history of Schema.org and its different usages for search engines and libraries.[19] Schema.org provides a simple way for libraries, archives, and museums to expose liked data using microdata encoded in HTML5. For our BIBFRAME HTML display pages, we utilized Schema.org microdata. Clark's presentation at the American Library Association Annual Conference in 2014 about Schema.org markup demonstrates how Schema.org metadata can be used in library settings, noting that there are some descriptors like library holdings that lack one to one mapping.[20] Recently, however, new work developed by the World Wide Web Consortium's (WW3C) Schema Bib Extend Community Group addresses several of these needed mappings. Results of their work are available at the bib.schema.org webpage.[21] Before the availability of the bib.schema.org work we utilized the schema.org property *brand* to reference an e-book publisher, when we would have preferred the more library focused property *publishedBy*.

According Godby (OCLC) and Denenberg (LC), "the coverage of Schema.org is necessarily broad but shallow because library resources must compete with creative works offered by many other communities in the information landscape. Conversely, the coverage of BIBFRAME is deep because it contains the vocabulary required of the next-generation standard for describing library collections."[22] There are at least three high-level differences between LC's BIBFRAME and the Schema.org model adopted by OCLC. First, work and instances are defined in BIBFRAME, while work is defined in Schema.org, but not instance. Second, BIBFRAME defines an authority entity, but not Schema.org. Third, BIBFRAME defines the annotation entity, and Schema.org model does not.

The BIBFLOW project at the University of California Davis Library is an Institute of Museum and Library Services-funded initiative to examine workflows, systems, and processes necessary to move libraries into BIBFRAME. The grant includes partnership with Zepheira. The researchers hypothesize that,

> while these new standards and technologies are sorely needed to help the library community leverage the benefits and efficiencies that the Web has afforded other industries, we cannot adopt them in an environment constrained by complex workflows and interdependencies on a large ecosystem of data, software and service providers that are change resistant and motivated to continue with the current library standards (e.g. Anglo-American Cataloguing Rules (or AACR) and MARC. Research is required on how research libraries should adapt our practices, workflows, software systems and partnerships to support our evolution to new standards and technologies.[23]

Their work dovetails with the BIBFRAME project described in this research paper; we describe how transformed BIBFRAME data will be surfaced in a discovery view and also demonstrate how library systems can be modularly designed to mitigate some of the complexity inherent within the traditional Integrated Library System (ILS).

To summarize the three strands of disagreement regarding the potential usefulness of BIBFRAME implementation and the transition from MARC—one strand of thought leaders is looking to optimize discovery of resources that favor Schema.org metadata for MARC transformation. As we described above, Schema.org metadata without extensions lacks several library specific descriptors, however several researchers have found extensions to Schema.org to be sufficient.[24] There is a second somewhat cautionary thought that suggests that discarding MARC in favor of BIBFRAME is premature.[25] Most libraries will tread this path early on. While yet a third group of leaders are sympathetic

to projects like BIBFRAME and suggest that modeling the richness of MARC is an important component of transitioning library description into linked data.[26] Our approach was to use both BIBFRAME and Schema.org for enhanced discovery. We noted the extensions to Schema.org and find value in making use of microformats encoded in HTML. BIBFRAME was chosen as the library specific vocabulary for description encoded in RDF/XML, whereas Schema.org is utilized in our project when indexing HTML pages for a Custom Google Search Engine.

BIBFRAME Transformation and the Linked Data Enrichment Process

There are several ways the BIBFRAME model can be expressed using markup languages. In information modeling within the Library and Information Science community and digital librarianship specifically, it is common to express an information model in XML—the XML standard (more accurately a "meta-markup language") has proven to be a powerful tool for metadata transformation since many tools exist for traversing and transforming XML elements programmatically.[27] Due to XML's versatility, we chose to use RDF/XML encoding to model BIBFRAME resources. There are other ways to encode BIBFRAME, however, these other markup standards are highly specific to linked data in general and the Semantic Web in particular.[28] The modern use of XML for encoding MARC is exemplified in MARCXML, which is the starting point of the MARC records used in our experiment.[29]

RDF is a metadata model developed by the World Wide Web Consortium (W3C), which is implemented in Semantic Web resources and applications.[30] Many researchers have found RDF to be the de facto markup language for linked data, and many expected RDF to become the backbone of the Semantic Web. One challenge in working with RDF/XML is that while it is a standard markup for linked data applications, it is not easily readable and it serializes poorly. The reason for this poor serialization is that RDF/XML was meant as a data exchange format. The conceptual underpinning of RDF is quite basic: statements are made about resources using a subject, predicate, and object.[31] The implementation of this basic model in RDF/XML is the backdrop for our work.

As we note in our introduction, the BIBFRAME model focuses on four main classes: work, instance, authority, and annotation. However, on closer inspection by other thought leaders concerning the model's construction, there are basically two entities: work and instance. According to Coyle, "The BIBFRAME Work Represents the content portion of the bibliographic description, and the instance describes the carrier."[32]

The URI plays a profound role within BIBFRAME. A URI is a string of characters to uniquely identify a resource. It is also the basis for interlinking and providing context to resources. As an example of how URIs are foundational to linked data, consider our example of a MARC record with repeating data with BIBFRAME data are not repeated in this way since there is not a record in the classic catalog sense, rather data are simply referenced with URIs within BIBFRAME resources. These references can then be utilized by multiple BIBFRAME resources, and thus provide the interlinking and contextual reference point that provides the "meaning," of resources within the context of the Semantic Web.

Our BIBFRAME transformation process was iterative and exploratory. The BIBFRAME RDF that we began enriching with URIs was created using the MARCXML to BIBFRAME transformation tools available on LC's GitHub software repository page.[33] Enrichment of URIs was required since after transformation the resulting BIBFRAME RDF included multiple placeholders for URIs. In effect the transformation process was complete, but enrichment was necessary to create a valuable BIBFRAME resource that referenced other linked data URIs. Our first research efforts were to manually develop a model of BIBFRAME with enriched URIs. In practice this meant examining the output of LC's transformation code and manually enriching several hundred resources with relevant URIs.

We curated the RDF down to four files for each of the core classes of work, instance, authority and annotation. In the second phase of our project, the results of manual modeling were automated so that the nearly 300,000 e-book records were transformed through programmatic methods. We considered modifying the LC codebase for MARCXML to BIBFRAME so that it would include enrichment while it transformed MARCXML, but because of the complexity of the codebase, we instead chose to automate enrichment after BIBFRAME RDF transformation was complete. The model shown in figure 3 was utilized to map MARC records to BIBFRAME for the project.

Authority Modeling

The Authority class of a BIBFRAME resource is defined as a "representation of a key concept or thing. Works and Instances, for example, have defined relationships to these concepts and things."[34] Project researchers first focused on BIBFRAME's authority section, replacing blank URI nodes, the example.org links in the RDF, with open linked data authority URIs for creators and subject headings. Each library transitioning to BIBFRAME makes an implementation decision whether to represent a BIBFRAME authority as a blank node or reusable resource. Some libraries may use local identifiers that then associate with equivalency tags

to open URIs. This two-step process gives the library local control over URIs should they decide to alter or add to existing URIs.

For names, the researchers chose to link to VIAF, which combines over thirty name authority files worldwide. Researchers eliminated LC Name Authority File (NAF) links as the main links in the RDF, and replaced the example.org URI with the VIAF URI. This was done because VIAF has authority records for most authors/creators listed in the e-books. Additionally, the LC NAF is part of VIAF.

An example of a personal name linked to VIAF is shown below.

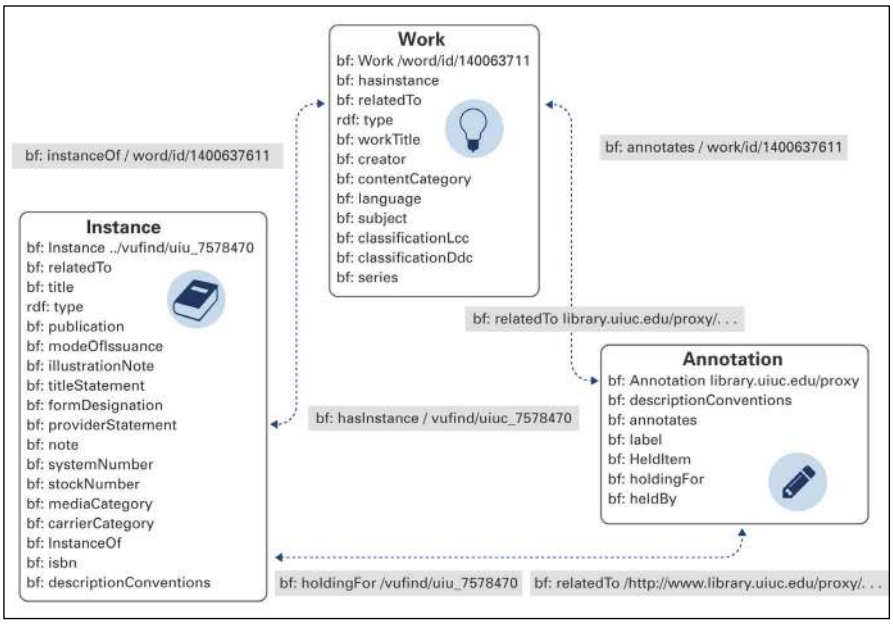

Figure 3. BIBFRAME ER Model utilized in project mapping

```
<bf:Person rdf:about= "http://viaf
.org/viaf/253339409">
<bf:label>Pivert, Olivier</
bf:label>
<bf:authorizedAccessPoint>Pivert,
Olivier</bf:authorizedAccessPoint>
<bf:hasAuthority>
<madsrdf:Authority>
<madsrdf:authoritativeLabel>Pivert, Olivier</
madsrdf:authoritativeLabel>
</madsrdf:Authority>
</bf:hasAuthority>
</bf:Person>
```

When the authors were unable to find names in VIAF, they linked them to WorldCat Identities, which has every name in WorldCat (over thirty million names), including named people, organizations, and fictitious characters. We also viewed WorldCat Identities as a reliable source for authority data.

Our first choice for subjects is to link to id.loc.gov. This database provides URIs for a large number of LC Subject Headings (LCSH) in our e-book bibliographic data among other authority files. An example linking a complex subject heading to id.loc.gov is provided below:

```
<madsrdf:isMemberOfMADSScheme
rdf:resource= "http://id.loc.gov/authorities/sub
jects"/>
</madsrdf:Authority>
</bf:hasAuthority>
</bf:Topic>
<bf:Topic rdf:about= "http://id.loc.gov/authorities/
subjects/sh85022943">
<bf:authorizedAccessPoint>Chemical plants—
Waste disposal</bf:authorizedAccessPoint>
<bf:label>Chemical plants—Waste disposal</
bf:label>
<bf:hasAuthority>
<madsrdf:Authority>
<rdf:type rdf:resource= "http://www.loc.gov/mads/
rdf/v1#ComplexSubject"/>
<madsrdf:authoritativeLabel>Chemical plants—
Waste disposal</madsrdf:authoritativeLabel>
```

While linking subject headings to id.loc.gov, the authors encountered challenges with subject headings not found in the database, or cases where only parts of complex subject headings are found. In the case that a subject heading could not be located in id.loc.gov, they then chose Faceted Application Subject Terminology (FAST), which is based on LCSH, but uses a simplified syntax.

An example linking to the FAST database:

```
<bf:Topic rdf:about= "http://experimental.world
cat.org/fast/1059826/">
<bf:authorizedAccessPoint>Petroleum refiner-
ies—Waste disposal</bf:authorizedAccessPoint>
<bf:label>Petroleum refineries—Waste disposal</
bf:label>
<bf:hasAuthority>
<madsrdf:Authority>
<rdf:type rdf:resource= "http://www.loc.gov/mads/
rdf/v1#ComplexSubject"/>
<madsrdf:authoritativeLabel>Petroleum refiner-
ies—Waste disposal</madsrdf:authoritativeLabel>
```

After those two searches were exhausted, the authors checked headings for medicine and health to see if URIs existed within Medical Subject Headings (MeSH), the National Library of Medicine's controlled vocabulary thesaurus. MeSH provides identifiers for main subject headings and their subdivisions. Both FAST and MeSH are reliable open linked data sources.

An example linking to MeSH:

```
</bf:Topic>
<bf:Topic rdf:about= "">
<bf:authorizedAccessPoint>Blood
Substitutes—adverse effects-Congresses</
bf:authorizedAccessPoint>
<bf:label>Blood Substitutes—adverse effects—
Congresses</bf:label>
<bf:hasAuthority>
<madsrdf:Authority>
<rdf:type rdf:resource="http://www.loc.gov/mads/
rdf/v1#ComplexSubject"/>
<madsrdf:authoritativeLabel>Blood
Substitutes—adverse effects—Congresses</
madsrdf:authoritativeLabel>
<madsrdf:isMemberOfMADSScheme
rdf:resource= "http://id.loc.gov/vocabulary/sub
jectSchemes/mesh"/>
</madsrdf:Authority>
</bf:hasAuthority>
```

Work Modeling

BIBFRAME's Work class is defined as a "Resource reflecting a conceptual essence of the cataloging resource."[35] To locate a proper WorkID for these e-books, the researchers considered several sources of "work identifier" information. OpenLibrary, the Internet Archive, and ebrary were each considered. The first two are open source resources that are similar to WorldCat. Ebrary, however, is a site that operates for profit. The WorldCat.org Work Identifier was chosen because it is part of a vast online database connecting libraries around the world. This service was still experimental at the time but was regarded by the authors to be a tentative best option.

An example link to a WorldCat Work Identifier:

```
<bf:Work rdf:about= "http://worldcat.org/entity/
work/id/1379076301">
```

Instance Modeling

BIBFRAME's Instance class is defined as a "resource reflecting an individual, material embodiment of the Work."[36] The authors chose the University of Illinois's VuFind link as an instance identifier. VuFind is our local online catalog.

```
<bf:hasInstance rdf:resource= "http://vufind.carli
.illinois.edu/vf-uiu/Record/uiu_7187480/
Description"/>
```

In our implementation, we linked our BIBFRAME work and instance by relationships expressed via the properties bf:hasInstance and bf:instanceOf.[37] A Work can have many Instances, and many Instances can point to one Work. Coyle has previously noted that in BIBFRAME, "instance is analogous to the FRBR manifestation. Item-level information is not treated as one of the primary bibliographic entities in BIBFRAME."[38] E-books are not tangible resources in the sense that there is an actual "item." Therefore, the folding of FRBR entity sets manifestation and item illustrated in figure 2 does not initially cause issues or necessitate additional workflows for e-book resource transformation for Work to Instance relationships in this round of data transformation. We note in the annotation model areas where item level data could be recorded as needed.

Annotation Modeling

BIBFRAME's Annotation class is defined as a "resource that asserts additional information about other BIBFRAME resource."[39] We investigated annotation modeling last because it is the model's most abstract part, though we found it useful for describing the item level information about a resource, as needed. As an example, within the "Annotation: about," we included a link to a site where we can access the e-book described in BIBFRAME data. The following link leads to the electronic access of the e-book.

```
<bf:relatedTo rdf:resource= "http://www
.library.uiuc.edu/proxy/go.php?url=http://
www.oxfordreference.com/view/10.1093/
acref/9780199738878.001.0001/acref
-9780199738878"/>
```

HTML model

The BIBFRAME RDF/XML was then hosted within a HTML page for the resource. Within that HTML, the project researchers included display elements for Access, Item Description, Subject Terms/Creators, and BIBFRAME RDF—where links to the individual pages of each RDF/XML section are linked (see figure 4). This enables our work to be reviewed and critiqued by others in the field and also allows others to observe our finalized model when creating their own BIBFRAME resources.

The researchers decided to include both the LCC number for the e-book and a short description of the item for which the record is created. The LCC number is taken from the RDF, as are the "notes" except for a few occasions when the notes are not available. While the authors believe that call numbers are important in linked data, yet for a few records, the RDF from e-books do not include a LCC number, which is problematic. Most of the records lacking a LCC number also lack a "Held Item" field in the RDF, and the authors searched WorldCat for a LCC number. If no number was found in WorldCat, the LCC number was not included in the HTML. Some of the records without "Held Item" portions are the proceedings from a meeting or conference.

Since the HTML records are web resources, several of the open linked data elements included in the BIBFRAME resources are also embedded in the HTML as Schema.org structured data. The project researchers used Google's Structured Data Testing Tool to properly enrich the HTML with linked data from the Schema.org vocabulary.[40] Including Schema.org markup in the HTML records allow a Google Custom Search engine to surface the linked data that are included in the BIBFRAME RDF. The Schema.org types utilized include Person, Book, Brand, URL, and Thing.

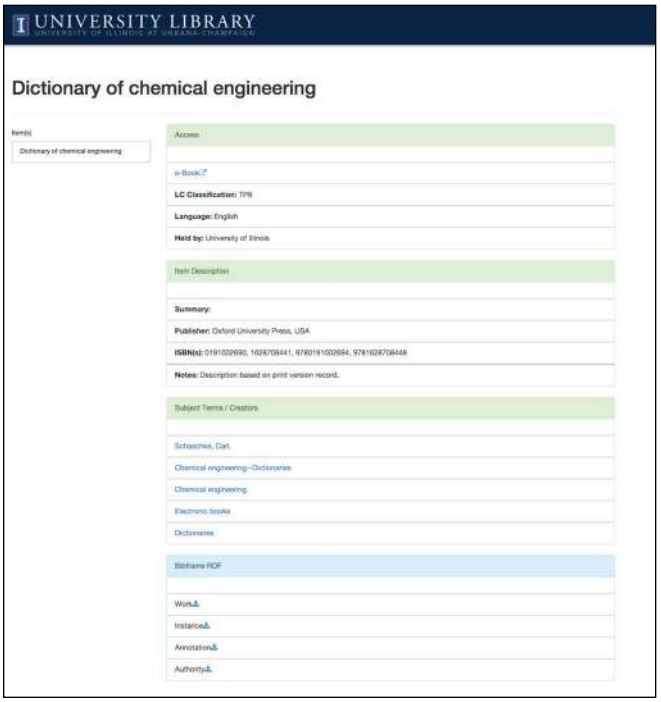

Figure 4. BIBFRAME HTML page

Process for Automated Transformation and Discovery

For each of the models described above, researchers developed a corresponding URI enrichment code written in Python. Python is a commonly used programming language for batch MARC data transformation and enrichment.[41] Several Python programs were developed to generate the enrichments for BIBFRAME elements programmatically using the master BIBFRME RDF/XML file.[42] It should be noted that the authors' BIBFRME RDF/XML file was generated from code available from LC. LC's code repository utilized a software language known as XQuery, which is a standard software tool employed for traversing and transforming XML.

Web-based Application Programming Interfaces (APIs), concise, specifically formatted data produced by programs to be consumed by other programs, were used to enrich the transformed RDF with linked open data. The Python programs take the transformed BIBFRAME RDF record from the marc2bibframe XQuery code and generate an Annotation, Instance, Work, and Authority RDF file with enriched linked data as an output. By enriching the records with linked data, we have a complete record that lacks blank nodes. Local nodes that pointed only to local resources are also avoided in the automation process. Target open data links are reviewed below.

Authority APIs against which the authors programmed included:

- VIAF Corporate Names:
  - https://viaf.org/viaf/search?query=local.corporateNames+all
- VIAF Personal Names:
  - https://viaf.org/viaf/search?query=local.personalNames+all+
- MeSH Linked Data:
  - http://id.nlm.nih.gov/mesh/servlet/query?query
- Library of Congress Linked Data Service
  - http://id.loc.gov/search/?q=
- FAST Heading
  - http://experimental.worldcat.org/fast/search?query=cql.any+all+

Annotation APIs:

- WorldCat XISBN Service (for Work id)
- http://xisbn.worldcat.org/webservices/xid/oclcnum/
- UIUC VuFind (Held item)
- http://vufind.carli.illinois.edu/vf-uiu/Record/uiu_

Instance APIs:

- WorldCat XISBN Service (for Work id)
- http://xisbn.worldcat.org/webservices/xid/oclcnum/

Work APIs:

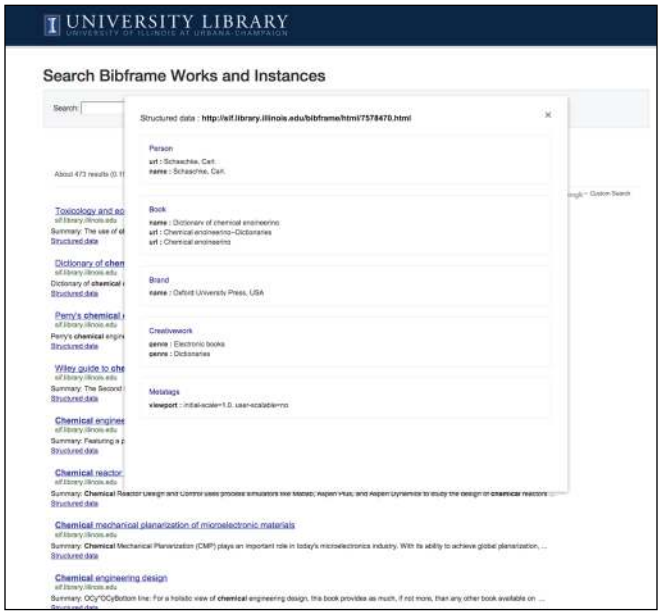

Figure 5. Structured data in Google Custom Search

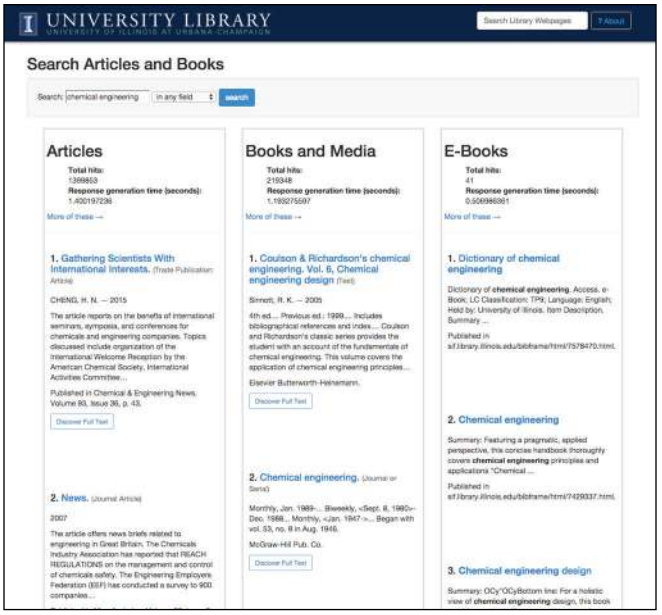

Figure 6. Bento-style discovery view with e-book search results

- WorldCat XISBN Service (for Work id)
- http://xisbn.worldcat.org/webservices/xid/oclcnum/
- VIAF Corporate Names:
- https://viaf.org/viaf/search?query=local.corporateNames+all
- VIAF Personal Names:
- https://viaf.org/viaf/search?query=local.personalNames+all+
- MeSH Linked Data:
- http://id.nlm.nih.gov/mesh/servlet/query?query
- Library of Congress Linked Data Service
- http://id.loc.gov/search/?q=
- Fast Heading
- http://experimental.worldcat.org/fast/search?query=cql.any+all+

After developing the automation code for the four BIBFRAME RDF/XML files and HTML page, the team transformed and enriched nearly 300,000 e-book records and has developed two prototype search interfaces.

We constructed an indexing program that would create sitemaps for 10,000 sets of records, which resulted in twenty-nine sitemaps that include URLs to 272,117 HTML BIBFRAME pages. The two options for retrieval of linked data records include a Google Custom Search Engine that surfaces the structured data in the result list (see figure 5), and Bento-style search (see figure 6) for e-book search simultaneously with articles and other catalog data. Google Custom Search provides results with structured data when retrieving BIBFRAME resources.[43]

Each HTML file (a BIBFRAME resource) incorporates BIBFRAME RDF/XML for a BIBFRAME Work, Instance, Authority, and Annotation. The BIBFRAME HTML also incorporates Schema.org structured data.

Discussion

There are several lessons learned from undertaking the BIBFRAME transformation and open linked data enrichment process.

The Transformation Process

Our strategy involved connecting to remote APIs to enrich records with linked data. Several times our Python scripts stopped retrieving data because of a "broken pipe" error. These errors are a result of one of the APIs not returning data. An API may stop returning data because it is programmed to stop responding, or cannot respond because of resource limits and will begin to drop responses during a high data load. We completed 272,117 HTML records for indexing, each of these pages has four RDF files linked for a total of 1,088,468 possible links. We identified 2,627 RDF links (a Work, Instance, Annotation or Authority RDF file) that are not transformed partly because of errors resulting from overloaded APIs. Since this is an experimental project, we are working to develop a process that runs a smaller number of records through the above referenced APIs. Currently, the project uses a folder input of 10,000 records,

but this could be reduced to as few as 100 and run with a queuing program. Smaller numbers of records being transformed may help to reduce the load on APIs, but may result in a more prolonged transformation process. Another option is to investigate alternatives to web-based APIs, and to use alternative data sources, such as static XML data stores where available.

As noted in our manual investigation, there are authority data that do not yet exist as linked data, and we may be left with blank links. Though OCLC makes available many Work IDs, the service was experimental at the time of this research, and did not yet include Work IDs for every resource required.

Searching and Indexing (Google Custom Search)

Earlier in the research process, we considered using Blacklight as an index for the transformed records.[44] It looked promising initially since it was developed for library data indexing and searching, and provides an API that we could have used to build a Bento-style search view. However, we later realized that Blacklight is optimized for indexing MARC records. We explored other indexing options for linked data and found that Google Custom Search provides indexing of structured data.

After testing the indexing of our HTML files within a Google Custom Search, we decided that this would be appropriate for the BIBFRAME search. Several digital library projects have also used search engine optimization for retrieval, including a recent project at Montana State University that used Schema.org markup to make better book viewers.[45]

Limitations

There are limits to what we could model in this project. Our current transformations model Work to Instance, and Instance to Work relationships. This is the output that is available from the marc2bibframe code. Since the BIBFRAME model can also incorporate several additional relationships, interlinking among all BIBFRAME relationships has not yet been fully realized in this project. According to the BIBFRAME documentation, "there are four types of relationships: Work to Work, Work to Instance, Instance to Work, Instance to Instance."[46]

It may be possible to leverage other APIs for this modeling. Specifically, OCLC makes available an xISBN web service that, when sent a string, will return a list of related ISBNs.[47] Such a tool can partially inform the finding of all manifestations. This may be helpful to complete instance-to-instance relationships. The xISBN web service is built from research at OCLC, notably, the FRBR Work-Set Algorithm.[48]

There are limitations of sustainability in any grant. To transform the University of Illinois' e-book MARC records to BIBFRAME resources, the researchers developed a prototype workflow, but there is currently no ongoing maintenance plan. To summarize, this is a discrete innovation funded grant. Project staff developed SQL queries to gather bibliographic identifiers for e-books that are then used to extract the MARC records as MARCXML. Next, we used XQuery from LC's *marc2bib* project to transform the BIBFRAME RDF and then enrich the BIBFRAME RDF with linked open data using Python. Finally, the data load included development of sitemaps for indexing Schema.org metadata by a Google Custom Search engine. Over time, additional e-books will be added to the catalog that are not captured by this process. The researchers will likely pursue an internal funding source to establish periodic updates to the corpus of e-books. Targeting newer bibliographic records will require altering our SQL queries to include titles that have been added since the previous cut-off date.

Conclusion

Because of our project, we have contributed an evaluation of the BIBFRAME model related to e-books. We have learned a great deal about the BIBFRAME model through converting the nearly 300,000 MARC records for e-books to BIBFRAME, developing an ER model for e-books, and creating two search interfaces for discovery of BIBFRAME linked data.

One challenging part of working with e-books using the BIBFRAME model is in choosing work identifiers. After much discussion, we decided on linking works to OCLC work identifiers. Another challenging part is to link people, families, corporate bodies, and works in bibliographic records to authority files. LC's linked data service is our top choice for this purpose. As a secondary source of authority linked data for people, families, corporate bodies, and works, we chose both the MeSH linked data service and the FAST linked data service to fill in these gaps. Unlike printed books, when a newer version of an e-book is imported to our catalog, the bibliographic record for the older version is deleted. This means we need to do more maintenance work for e-books. Serial resources may have similar issues since they are resources that may change over time because of possible title changes or interruptions and adjustments over time with regard to frequency of publication.

We believe our work in enriching data is particularly instructive for future projects in the University of Illinois Library, and applies to library data work across institutions. With the Python code developed for this grant, we can help to programmatically address other components of the catalog for enrichment. We envision that we will still need to do

local transformations even if OCLC eventually transforms all of their existing bibliographic records into linked data in the future. Institutions will need to transform the data themselves to be part of the OCLC community.

One of the key issues for our users to find library resources is to provide consistency in the form of access points used to identify people, families, corporate bodies, and works. The next phase this project will be to work with 7 million MARC records in our online catalog to address those limitations with BIBFRAME relationships between Work to Work, and Instance to Instance, which were not part of the initial innovation project.

The cataloging world is in transition. BIBFRAME is a profound step for the library community. It uses linked data to make discoverable library bibliographic and authority data on the web. Libraries considering piloting BIBFRAME transformations will be taking a leap forward in helping their users discover library resources across the web—and beyond the classic catalog paradigm.

References and Notes

1. Library of Congress, "Bibliographic Framework as a Web of Data: Linked Data Model and Supporting Services," accessed March 11, 2016, www.loc.gov/bibframe/pdf/marcld-report-11-21-2012.pdf.
2. Peter Pin-Shan Chen, "The Entity-Relationship Model: Toward a Unified View of Data," *ACM Transactions on Database Systems* 1, no. 1 (1976): 2.
3. "Bibliographic Framework Initiative," Library of Congress, accessed March 28, 2016, www.loc.gov/bibframe.
4. Ibid.
5. Library of Congress, "Bibliographic Framework as a Web of Data: Linked Data Model and Supporting Services," accessed March 11, 2016, www.loc.gov/bibframe/pdf/marcld-report-11-21-2012.pdf.
6. Kara Van Malssen, "BIBFRAME AV Modeling Study: Defining a Flexible Model for Description of Audiovisual Resources," accessed March 11, 2016, www.loc.gov/bibframe/docs/pdf/bibframe-avmodelingstudy-may15-2014.pdf.
7. "BIBFRAME Model & Vocabulary," Library of Congress, accessed March 25, 2016, www.loc.gov/bibframe/docs/index.html. The original research that is the subject of this article does not review the BIBFRAME 2.0 changes, however at a recent Coalition for Networked Information (CNI) Meeting, Robert Sanderson gave a talk "The Future of Linked Data in Libraries: Assessing BIBFRAME Against Best Practices" which addressed how BIBFRAME 2.0 changes align with linked data best practices. Sanderson's talk is available on Vimeo at https://vimeo.com/152611387.
8. "BIBFRAME Implementation Register," Library of Congress, accessed March 11, 2016, www.loc.gov/bibframe/implementation/register.html.
9. Jonathan Rochkind, "A Comparison of Article Search APIs via Blinded Experiment and Developer Review," *Code4Lib Journal* no. 19 (2013), accessed March 22, 2016, http://journal.code4lib.org/articles/7738; Cory Lown, Tito Sierra, and Josh Boyer, "How Users Search the Library from a Single Search Box," *College & Research Libraries* 74, no. 3 (May 2013): 227–41. We use the term Bento-style search software as previously defined by library technologist Tito Sierra, and explored in technical detail as it relates to article search APIs in Rochkind, as "custom designed software that provided catalog search and article search in separate sections of the result page."
10. "BIBFRAME at University of Illinois," University Library, University of Illinois at Urbana-Champaign, accessed March 11, 2016, http://sif.library.illinois.edu/bibframe.
11. "minrvaproject-admin / bibframeuiuc --Bitbucket," Bitbucket, accessed March 28, 2016, https://bitbucket.org/minrvaproject-admin/bibframeuiuc/. There are several online resources that offer free public repositories to host open source code, we chose Bitbucket for sharing our original Python programming work described in this paper, whereas the Library of Congress is using GitHub for their code repository.
12. Angela Kroeger, "The Road to BIBFRAME: The Evolution of the Idea of Bibliographic Transition into a Post-MARC Future," *Cataloging & Classification Quarterly* 51, no. 8 (2013): 873–90.
13. Roy Tennant, "MARC Must Die," *Library Journal* 127, no. 17 (2002): 26–27.
14. Working Group on the Future of Bibliographic Control, "On the Record: Report of the Library of Congress Working Group on the Future of Bibliographic Control," accessed March 11, 2016, www.loc.gov/bibliographic-future/news/lcwg-ontherecord-jan08-final.pdf.
15. Jason W. Dean, "Charles A. Cutter and Edward Tufte: Coming to a Library near You, via BIBFRAME," *In the Library with the Lead Pipe*, 2013, www.inthelibrarywiththeleadpipe.org/2013/charles-a-cutter-and-edward-tufte-coming-to-a-library-near-you-via-bibframe/.
16. Karen Rollitt, "MARC21 to Bibframe: Outcomes, Possibilities and New Directions," *New Zealand Library & Information Management Journal* 55, no. 1 (2014): 16.
17. Karim Tharani, "Linked Data in Libraries: A Case Study of Harvesting and Sharing Bibliographic Metadata with BIBFRAME," accessed March 11, 2016, http://ejournals.bc.edu/ojs/index.php/ital/article/view/5664 .
18. Joseph Kiegel, "BIBFRAME Projects at the University of Washington" (paper presented at the annual meeting for American Library Association, San Francisco, California, June 25 to June 30, 2015), http://slideplayer.com/slide/5383520/.
19. Jason Ronallo, "HTML5 Microdata and Schema.org," *Code4Lib Journal*, no. 16 (2012) accessed March 11, 2016, http://journal.code4lib.org/articles/6400.
20. Jason A. Clark, "Beyond Description: Using Schema.org to Describe Networks and Actions in Book, People, and

Discovery Settings" (paper presented at the annual meeting for American Library Association, Las Vegas, Nevada, June 26 to July 1, 2014), www.lib.montana.edu/~jason/talks/ala2014-session-schema.pdf.

21. Richard Wallis, "bib.schema.org | Schema Bib Extend Community Group," W3C Community & Business Groups Blog, June 24, 2015, accessed March 28, 2016, www.w3.org/community/schemabibex/2015/06/24/bib-schema-org/.
22. Carol Jean Godby and Ray Denenberg, *Common Ground: Exploring Compatibilities Between the Linked Data Models of the Library of Congress and OCLC* (Dublin, OH: Library of Congress and OCLC Research, 2015), www.oclc.org/research/publications/2015/oclcresearch-loc-linked-data-2015.html.
23. "BIBFLOW about page," accessed March 28, 2016, https://www.lib.ucdavis.edu/bibflow/about/.
24. Timothy W. Cole et al., "Library Marc Records Into Linked Open Data: Challenges and Opportunities," *Journal of Library Metadata* 13, no. 23 (2013): 163–96.
25. Diane Hillman, Gordan Dunsire, and Jon Phipps, "Maps and Gaps: Strategies for Vocabulary Design and Development," Proceedings of the International Conference on Dublin Core and Metadata Applications 2013, accessed March 23, 2016, http://dcpapers.dublincore.org/pubs/article/view/3673.
26. Gordon Dunsire et al., "Linked Data Vocabulary Management: Infrastructure Support, Data Integration, and Interoperability," Information Standards Quarterly 24, no. 2/3 (2012): 9. With regards to modeling the descriptive richness of MACR21, RDA, & ISBD, several thinkers opined that these are "important aspects of current experimentation in linked bibliographic data." The authors were referencing early trailblazing linked data work that reused several vocabularies in transforming rich library bibliographic description to the Semantic Web. Their examples were drawn from the British Library, Cambridge University, and The German National Library circa 2012.
27. J. David Eisenberg, "Comparing XSLT and XQuery," accessed March 25, 2016, www.xml.com/pub/a/2005/03/09/xquery-v-xslt.html.
28. Other semantic markup in which BIBFRME could be expressed include N3/Turtle (https://en.wikipedia.org/wiki/Turtle_(syntax)) and n-triples (https://en.wikipedia.org/wiki/N-Triples). The result of our work was not to create the classical RDF triple store common in Semantic Web modeling, but rather our key project deliverable was to model a process of transformation, indexing, and discovery. Schema.org microdata in HTML provided a nimbler path to implementation.
29. "MARC 21 XML Schema," Library of Congress, accessed March 25, 2016, www.loc.gov/standards/marcxml.
30. "RDF 1.1 Primer," W3C Working Group, accessed March 25, 2016, www.w3.org/TR/rdf11-primer.
31. Ibid.
32. Karen Coyle, *FRBR, Before and After: A Look at Our Bibliographic Models* (Chicago: ALA Editions, 2016), 145.
33. "lcnetdev/marc2bibframe," GitHub, accessed March 28, 2016, github.com/lcnetdev/marc2bibframe.
34. "Vocabulary," Bibliographic Framework Initiative Project, accessed March 28, 2016, bibframe.org/vocab/Authority.html.
35. "Overview of the BIBFRAME Model," Library of Congress, accessed March 28, 2016, www.loc.gov/bibframe/docs/model.html.
36. Ibid.
37. "BIBFRAME Relationships," Library of Congress, accessed March 11, 2016, www.loc.gov/bibframe/docs/bibframe-relationships.html#connect.
38. Karen Coyle, *FRBR, Before and After*, 145.
39. "Overview of the BIBFRAME Model," accessed March 11, 2016, https://www.loc.gov/bibframe/docs/model.html.
40. "Structured Data Testing Tool," Google Developer, accessed March 11, 2016. https://developers.google.com/structured-data/testing-tool/.
41. "About Python," Python Software Foundation, accessed March 22, 2016, https://www.python.org/about/.
42. All data inputs of this project can be accessed from the Illinois Data Bank at https://doi.org/10.13012/B2IDB-1065549_V1. To replicate automated components of BIBFRAME enrichment, follow the instructions available at the project code repository here: https://bitbucket.org/minrvaproject-admin/bibframeui-uc, the main project parts are in the transform folder which holds the RDF enrichment code for a BIBFRAME Work, Instance, Annotation, and Authority (https://bitbucket.org/minrvaproject-admin/bibframeuiuc/src/7863b737d7dc51f77d3ebcb5215228b6d8082825/transform/?at=master) folder setup of the Python code requires an input and an output path. The one dependency that is required is that of an input RDF that can be generated using the Library of Congress XQuery transformation code *marc2bibframe* (https://github.com/lcnetdev/marc2bibframe) and a starting MARC XML file to begin the process. In sum the process takes in an initial MARC XML file, transforms it to BIBFRAME RDF/XML, and then four separate python files corresponding to the BIBFRAME model (Work, Instance, Annotation, and Authority) are run over the BIBFRAME RDF/XML output.
43. "Search Articles and Books," University Library, University of Illinois at Urbana-Champaign, accessed March 28, 2016, sif.library.illinois.edu/megasearch ; "Search Bibframe Works and Instances," University Library, University of Illinois at Urbana-Champaign, accessed March 11, 2016, http://sif.library.illinois.edu/bibframe/search.php.
44. For more information regarding the Blacklight software project see http://projectblacklight.org.
45. Jason A. Clark and Scott W.H. Young, "Building a Better Book in the Browser (Using Semantic Web Technologies and HTML5)," *Code4lib journal* 29 (2015), http://journal.code4lib.org/articles/10668.

46. "BIBFRAME Relationships," Library of Congress, accessed March 11, 2016, www.loc.gov/bibframe/docs/bibframe-relationships.html#intro.
47. "WorldCat Web service: xISBN [OCLC - WorldCat Affiliate tools]: Home," OCLC, accessed March 29, 2016, http://xisbn.worldcat.org/xisbnadmin/index.htm
48. Thomas B. Hickey and Jenny Toves, "FRBR Work-Set Algorithm," accessed March 11, 2016, www.oclc.org/content/dam/research/activities/frbralgorithm/2005-04.pdf.