BIBFLOW: A Roadmap for Library Linked Data Transition

Prepared 14 March, 2017

MacKenzie Smith
Carl G. Stahmer
Xiaoli Li
Gloria Gonzalez

University Library, University of California, Davis
Zepheira Inc.

Research support by the Institute for Museum and Library Services

Table of Contents

I. Introduction

BIBFLOW is an Institute for Museum and Library Services (IMLS) funded multi-year project of the University of California Davis Library and Zepheira Corporation. Traditional library data methods are out of sync with the data storage, transmission, and linking standards that drive the new information economy. As a result, new standards and technologies are sorely needed to help the library community leverage the benefits and efficiencies that the Web has afforded other industries. The findings in this report are the result of research focused on how libraries should adapt their practices, workflows, software systems, and partnerships to support their evolution to new standards and technologies. In conducting this research, the BIBFLOW team collaborated and communicated with partners across the library data ecosystem – key organizations like the Library of Congress, OCLC, library vendors, standards organizations like NISO, software tool vendors, commercial data providers, and other libraries that are trying to plan for change. We also experimented with various technologies as a means of testing Linked Data transition and operation workflows. The specific focus of this study was the Library of Congress' emerging BIBFRAME model, a framework developed specifically to help libraries leverage Linked Data capabilities.

This report is the result of two years of research across the spectrum of Linked Data implementation and operations. Its purpose is to provide a roadmap that individual libraries can use to plan their own transition to Linked Data operations. It makes specific recommendations regarding a phased transition approach designed to minimize costs and increase the efficiency and benefits of transition. An analysis of specific transition tools is provided, as well as an analysis of workflow transitions and estimated training and work effort requirements.

A key finding of the report is that libraries are better positioned than most believe to transition to Linked Data. The wider Linked Data ecosystem and the semantic web in general are built on the bedrock of shared, unique identifiers for both entities (people, places, etc.) and actions (authored, acquired, etc.). Libraries have a long history of shared data governance and standards; as such, library culture is well suited to transitioning to Linked Data, and library structured data (MARC) is well situated for data transformation. In light of the above, it is our conclusion that Linked Data represents an opportunity rather than a challenge, and this roadmap is intended to serve as a guide for libraries wishing to seize this opportunity.

II. Why Linked Data

In 1998 the World Wide Web Consortium (W3C) published Tim Berners-Lee's *Semantic Web Road Map*.[1] In this essay, Berners-Lee lays out an "architectural plan" that, to this day,

[1] See Berners-Lee, Tim. "Semantic Web Road Map." World Wide Web Consortium, 14 Oct. 1998. Web. 12 Oct. 2016. <https://www.w3.org/DesignIssues/Semantic.html.>. A full history of Linked Data, which extends at least back to Allan M. Collins, Ross Quillian and Elizabeth F. Loftus's *Semantic Network Model*, is beyond the scope of this report; however, it was Berners-Lee's vision of the Semantic Web that focused attention on the implementation of semantic network models as a foundation for information exchange over the World Wide Web.

provides the foundation of the Linked Data ecosystem. According to Berners-Lee, "The Web was designed as an information space, with the goal that it should be useful not only for human-human communication, but also that machines would be able to participate and help." This remains the fundamental ethos of the drive towards Linked Data—the idea that we can and should structure our data such that machines, without the aid of human readers, can follow threads of communication, building ever-deepening networks of knowledge.

A simple example will serve to clarify this concept. You're watching movie version of *The Lord of the Rings* and you become interested in what influences might have inspired Tolkien in the writing of the book; so you turn to Google and search for "The Lord of the Rings Influences." Here you find a Wikipedia page on Tolkien that tells you that he was a Catholic, and a student of Norse and Germanic mythology. You also see that Tolkien wrote many other works in addition to *The Lord of the Rings* and that all seem to be bear the influence of Tolkien's study of both contemporary and ancient religions. You see also that both Neil Gaiman and Ursula K. Le Guin (among others) were heavily influenced by Tolkien. You know Neil Gaiman as a contemporary fantasy author, but you aren't familiar with Ursula Le Guin, so you click on the link to her Wikipedia Page. Here you find out that she is a Science Fiction and Fantasy author whose works were, like Tolkien, heavily influenced by Norse Mythology and Anthropology. This prompts you to think about the relationship between Science Fiction and Fantasy as Genres, so you visit several websites devoted to the history of each; and, at each, you find that Tolkien occupies an important place in the lineage of both traditions. *Etc., Etc., Etc.*

The above is representative of how human readers traverse complex webs of information on a regular basis. At each stage in the traversal our reader could have followed multiple paths through the information web. The Wikipedia article on *The Fellowship of the Rings* alone contains 1,698 links to other sources of information. Our reader's decisions to traverse particular paths are rooted in formal semantics, the ability to use context to determine which paths are most likely related to the information retrieval task at hand. The choice to investigate the Fantasy Genre in the example above was rooted in the knowledge that the tree of Tolkien's influence included multiple relationships between Science Fiction and Fantasy authors.

Linked Data has one primary purpose: to allow machines to traverse the vast web of networked information with the same facility as human readers. Given a starting record or text, the computer, like our TV viewer above, should be able to identify webs of connectivity and traverse particular paths based on semantic decision making. The Non-Linked Data web does not allow for this kind of machine traversal. Linked Data does.

The image below is a rendering of a portion of an information graph created by a computer by traversing information about *The Fellowship of the Rings* and its relations across the various Linked Data resources already publicly available on the internet.

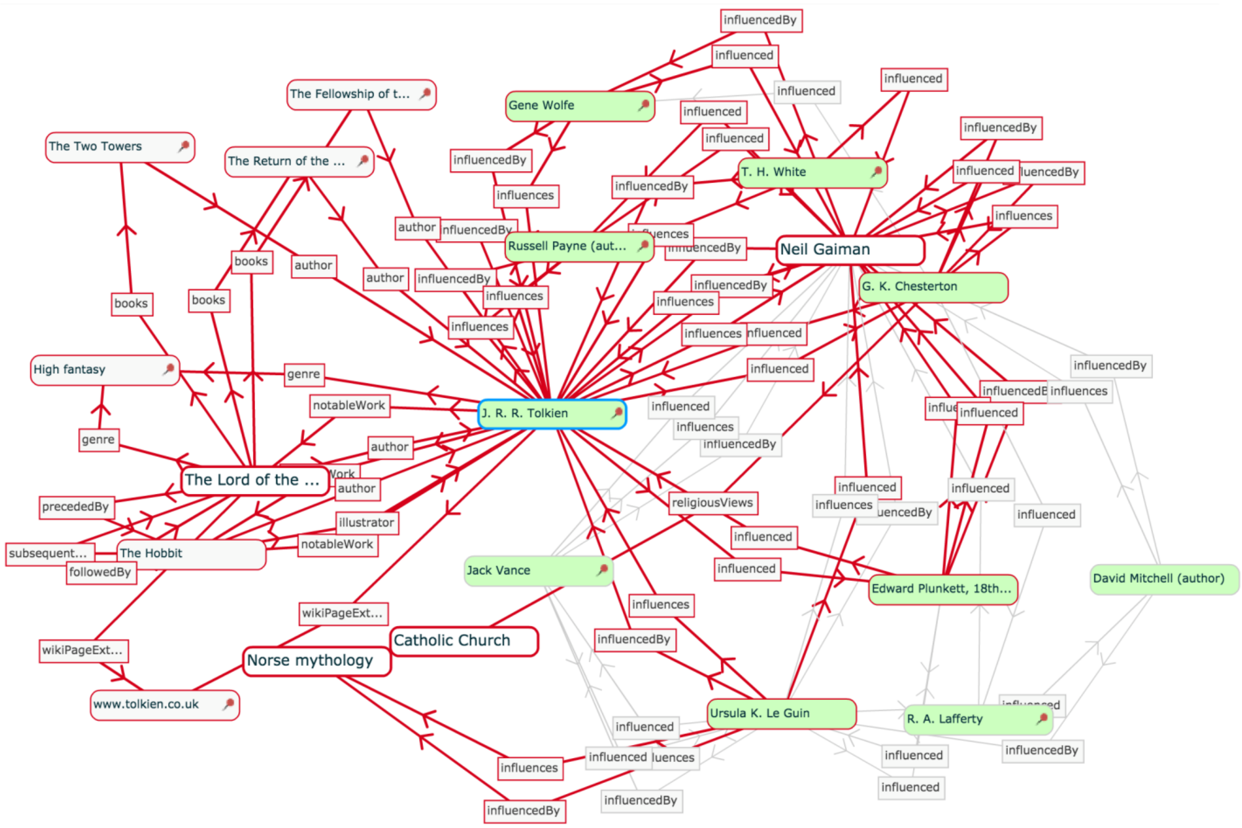


Figure 1*:* *The Lord of the Rings* seen as Linked Data

Here we see a vast network, or graph, of information surrounding an item of interest that the computer is able to generate using Linked Data. Graphs grow through iterations of traversal, starting with a core node, each of which reveals a new branch, or edge in the graph. These branches can be contained by limiting the number of traversal iterations, but they are theoretically infinite.

This is not the case for the **Ma**chine **R**eadable **C**ataloging (MARC) standards upon which current library catalogs are built. Certainly, MARC can, and has for some time, been used to link various knowledge repositories. When we search for J. R. R. Tolkien in a library catalog and receive a list of works written by and about the author, the computer has enacted a kind of linking around the name J. R. R. Tolkien. MARC's ability to facilitate this linking is, however, extremely limited for a variety of reasons.

MARC records are based on a complex data standard as currently defined and documented by the Library of Congress at https://www.loc.gov/marc. A key differentiator between MARC and Linked Data cataloging frameworks is that MARC is based on records whereas Linked Data is based on graphs. Unlike knowledge graphs, which are theoretically infinite, records have a fixed number of fields and subfields. MARC Authority records, for example, are composed of 183 fields. An individual cataloger cannot extend this structure vertically or laterally, which to say that one cannot add new fields to the system nor posit new relationships between fields. The standard's field/subfield structure insures that relational knowledge can only extend two iterations from the object defined, and it also limits the things that can be said at each iteration. The only way to extend the framework is through a complex, top-down driven process of discussion and adoption involving many institutions and governing bodies, followed by the reprogramming of all software systems that deal with the records.

Graph based knowledge systems are not subject to any of the above limitations. They simultaneously strengthen the ability to describe objects using reputable controlled vocabularies while at the same time providing an extensibility that allows users to add new knowledge nodes (fields) to their descriptive graphs. One can capture all of the fields currently represented in a MARC record using references to the same controlled vocabularies (when applicable) and add additional information as appropriate.

III. Transition Fundamentals

Transitioning to Linked Data is not a data transformation activity. Libraries have extensive experience transforming data from one format to another. While crosswalk processes can be cumbersome and time consuming, they are well understood and we are quite good at them. Transitioning to Linked Data, however, requires more than simply mapping fields across data models and performing necessary data reformatting to comply with the specifications of the new model. Transitioning to Linked Data requires adding new data to each record, data that can often be difficult to disambiguate by machine. Specifically, a successful transition to a Linked Data ecosystem requires adding numerous shared, publicly recognized unique identifiers (a Uniform Resource Identifiers, or URI) to each record at the time of transformation.

URIs form the backbone of the Linked Data ecosystem. The fundamental concept is to provide a unique, machine actionable identifier for all entities in a graph. Thus, for example, whereas a human might say:

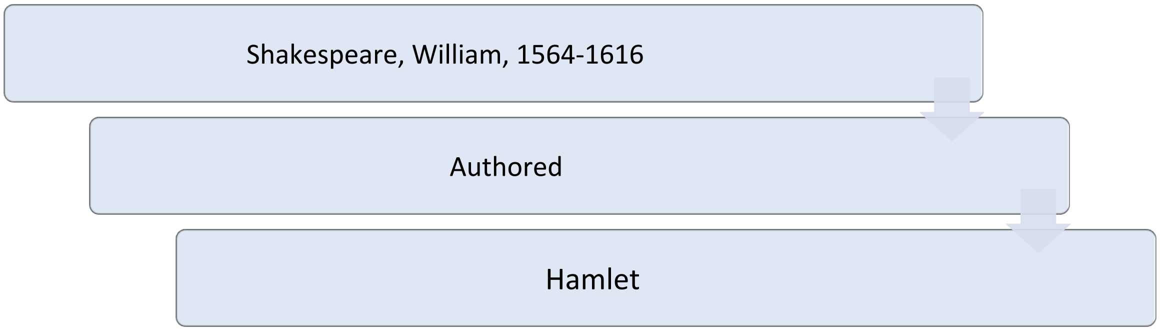


Figure 2: Human readable triple

A Linked Data representation of the same statement would look like:

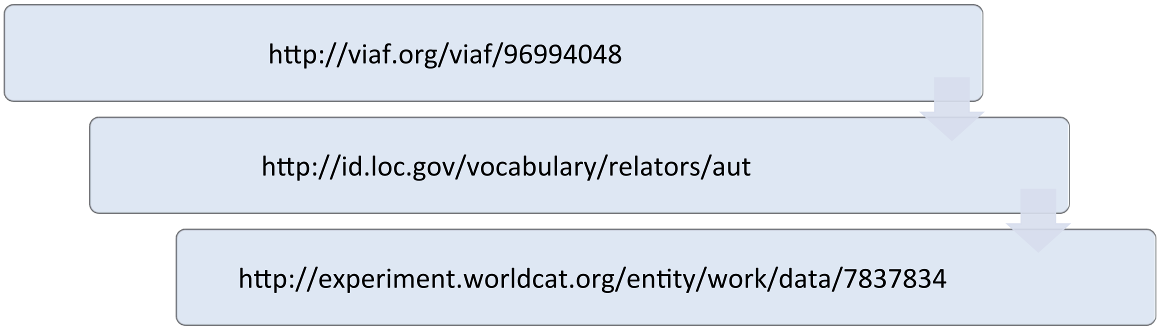

http://viaf.org/viaf/96994048
http://id.loc.gov/vocabulary/relators/aut
http://experiment.worldcat.org/entity/work/data/7837834

Figure 3: Machine readable triple

When we refer to URIs as "machine actionable" or "machine traversable," we mean to say that an identifier is uniformly recognized by independent computing systems, allowing them to use it to link things being said about the same entity by different people or telling it about a relationship that can be used to control function and output. For example, if *you* have a collection of records that says that "Shakespeare wrote *Hamlet"* and *I* have a collection of records that says "Shakespeare wrote *Romeo and Juliet*," adding URIs to our records allows a computer to infer that "Shakespeare wrote Hamlet and Romeo and Juliet." Similarly, if we used URIs to identify *Hamlet* and *Romeo and Juliet,* the computer could search across the network for things that others have said about each of these plays.

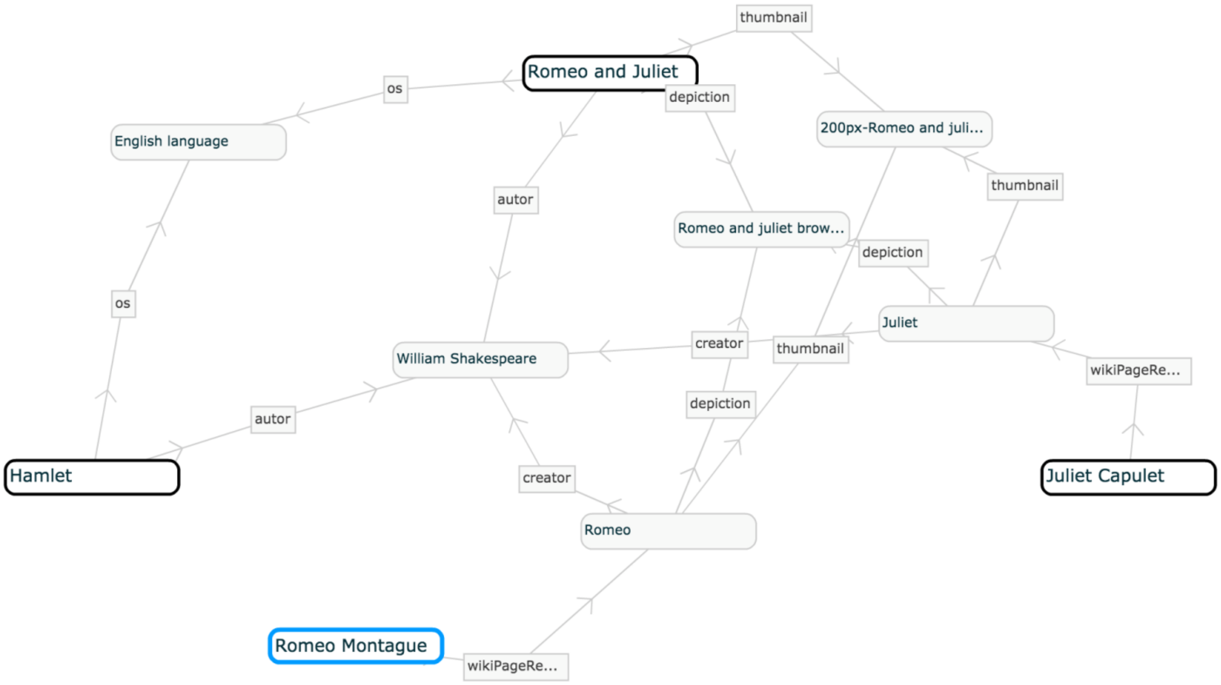


Figure 4: Dynamic Linked Data graph

The above figure shows a partial graph of relationships between *Hamlet* and *Romeo and Juliet* that was dynamically created, with no human intervention, by traversing URI based statements about the two plays that are currently available as Linked Data on the internet.

For a full discussion of the function and benefits of Linked Data see the "Why Linked Data" section of this report. For the present purposes, what concerns us is the role that URIs serve in the Linked Data universe. A Linked Data graph is only as good as its URIs. If two individuals use two different URIs for the same entity, William Shakespeare for example, then to the computer there are two William Shakespeares. As such, proper URI management is essential to the Linked Data effort.

Several organizations, such as Getty, the Library of Congress, OCLC, and VIAF, currently make available Linked Data gateways that provide URIs for entities and controlled vocabularies widely used by libraries and cultural heritage organizations.[2] Using these resources, organizations can lookup shared URIs for entities (people, organizations, subjects, etc.) Similarly, BIBFRAME defines a set of relationships for which public URIs have also been minted.

From a data perspective, the primary obstacle to transitioning to Linked Data is associating the literal representation of entities in MARC records (Shakespeare, William, 1564-1616) with machine actionable URIs (http://viaf.org/viaf/96994048).[3] This association must be backward implemented on all legacy records (a daunting task) and library systems must be updated to create the association when dealing with new records or editing existing ones (a potentially difficult task since most libraries rely on vendor software over which they have little control to perform this work.)

In addition to the technical problems presented by conversion of data, transitioning to Linked Data also brings with it a host of potential systems and workflow issues. Current library operations rest on workflows designed for and performed by staff with specialized and advanced training and knowledge. Changing the required output of these workflows could potentially have dramatic effects on the workflows that create it. *Section VII* of this document discusses these changes in depth.

Finally, transitioning our data and workflows will also necessarily impact library systems and information flow. The figure below is a diagram of the numerous systems in place at the UC Davis library that communicate either directly or by association with our library catalog:

[Continued on Next Page]

[2] See http://www.getty.edu/research/tools/vocabularies/lod/, http://id.loc.gov/, https://www.oclc.org/developer/develop/linked-data.en.html, and http://viaf.org respectively.

[3] There are workflow and cultural obstacles to Linked Data transition as well. These are addressed elsewhere in this roadmap.

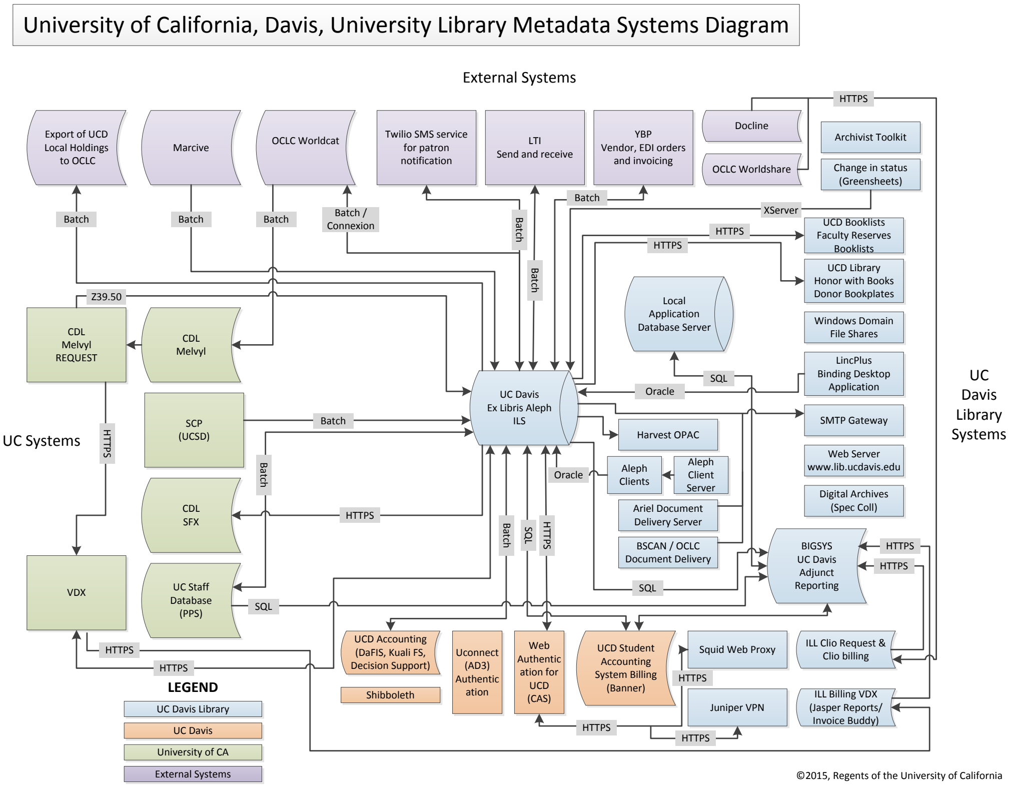

University of California, Davis, University Library Metadata Systems Diagram
External Systems
Export of UCD Local Holdings to OCLC
Marcive
OCLC Worldcat
Twilio SMS service for patron notification
LTI Send and receive
YBP Vendor, EDI orders and invoicing
Docline
OCLC Worldshare
Archivist Toolkit
Change in status (Greensheets)
HTTPS
XServer
Batch
Batch
Batch
Batch / Connexion
Batch
Batch
Batch
Z39.50
HTTPS
HTTPS
UCD Booklists Faculty Reserves Booklists
UCD Library Honor with Books Donor Bookplates
Windows Domain File Shares
LincPlus Binding Desktop Application
SMTP Gateway
Web Server www.lib.ucdavis.edu
Digital Archives (Spec Coll)
UC Davis Library Systems
Local Application Database Server
SQL
Oracle
CDL Melvyl REQUEST
CDL Melvyl
SCP (UCSD)
CDL SFX
UC Staff Database (PPS)
VDX
UC Systems
HTTPS
Batch
Batch
HTTPS
SQL
UC Davis Ex Libris Aleph ILS
Harvest OPAC
Oracle
Aleph Clients
Aleph Client Server
Ariel Document Delivery Server
BSCAN / OCLC Document Delivery
SQL
BIGSYS UC Davis Adjunct Reporting
HTTPS
HTTPS
Batch
SQL
HTTPS
UCD Accounting (DaFIS, Kuali FS, Decision Support)
Shibboleth
Uconnect (AD3) Authentication
Web Authentication for UCD (CAS)
UCD Student Accounting System Billing (Banner)
Squid Web Proxy
HTTPS
Juniper VPN
ILL Clio Request & Clio billing
ILL Billing VDX (Jasper Reports/ Invoice Buddy)
HTTPS
HTTPS
HTTPS
HTTPS
LEGEND
UC Davis Library
UC Davis
University of CA
External Systems
©2015, Regents of the University of California

Figure 5: Library systems diagram

As depicted in the above diagram, 40 different systems connect either directly or indirectly with our library catalog. Each of these connections represents a potential point of failure during a Linked Data Transition, further complicating any imagined or real transformation process.

IV. Roadmap Overview

The transition roadmap presented here is based on two years of experimenting with various approaches to making a transition to Linked Data. The plan is driven by the following seven primary principles:

1. Insure accuracy of resulting data
2. Insure proper function of data in the wider information systems ecosystem
3. Minimize impacts on daily operations during transition
4. Minimize impacts on library workflows except where changes will result in increased efficiency and improved quality of work
5. Minimize the need for additional staff training
6. Maximize benefits Linked Data offers with regard to data sharing and interoperability
7. Maximize benefits Linked Data offers in terms of extensibility of descriptive practices and methods (improve depth of records)

The proposed transition plan is a two-phased plan, each comprised of multiple steps. Importantly, *Phase One* can be undertaken as an end-game transition process and will situate libraries to function in a Linked Data library ecosystem. Libraries that complete *Phase One* will be able to exchange BIBFRAME and other Linked Data graphs with other libraries and cultural heritage institutions with minimal impact on staff and systems, but also without capitalizing on the full potential of Linked Data. Libraries that go on to complete *Phase Two* will add to this the ability to capitalize on the extensibility inherent in Linked Data graph description and also introduce efficiencies in cataloging workflows. Libraries should seek the level of engagement that aligns with their in-house technical expertise, efforts performing original cataloging, desire to create a deeper and more descriptive catalog, and budget.

The following figure presents a high-altitude view of the proposed conversion roadmap, including milestones of each phase:

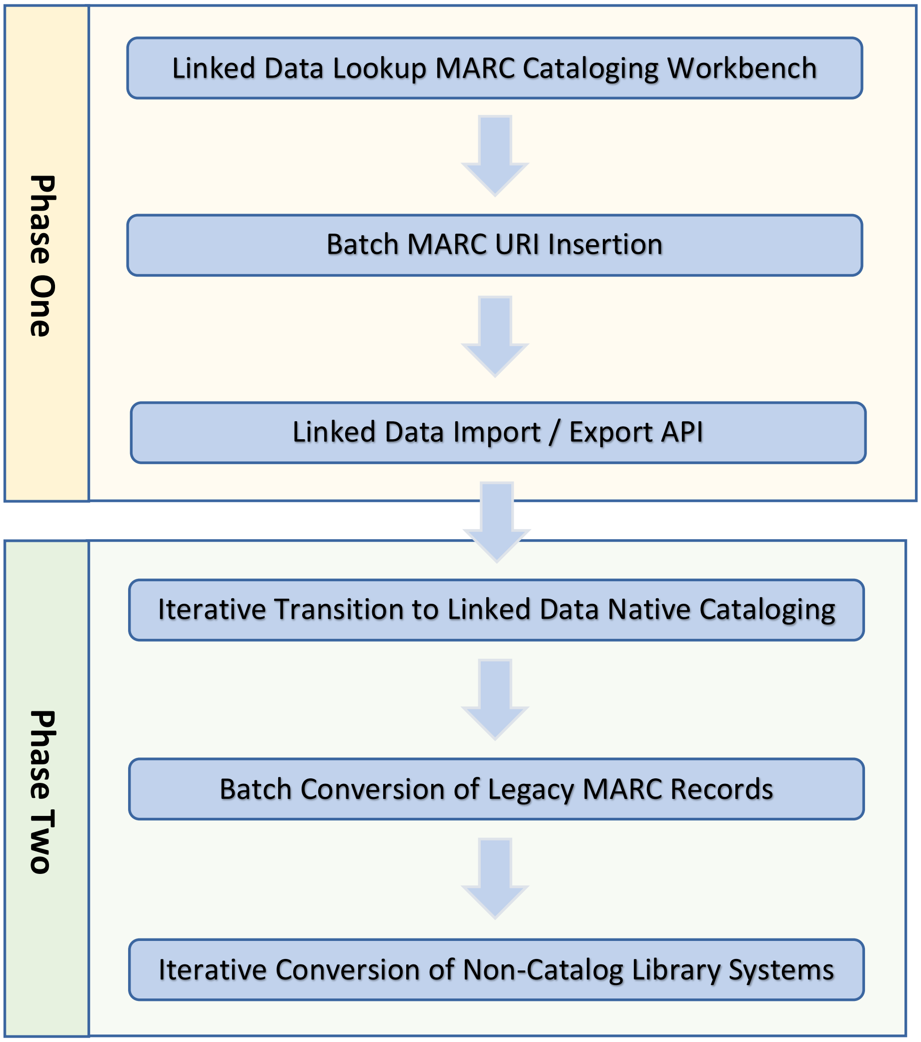

Phase One
Linked Data Lookup MARC Cataloging Workbench
Batch MARC URI Insertion
Linked Data Import / Export API
Phase Two
Iterative Transition to Linked Data Native Cataloging
Batch Conversion of Legacy MARC Records
Iterative Conversion of Non-Catalog Library Systems

Figure 6: Transition process overview

The Primary focus of *Phase One* is preparing existing MARC records for transformation to Linked Data graphs. The involves inserting appropriate URIs into MARC records so that records can be converted into functioning Linked Data graphs that include machine actionable URIs. At the conclusion of *Phase One,* the catalog's data store and cataloging user interface remain MARC based, but the presence of URIs in MARC records allows for the development of Application Programming Interfaces (API) to export and ingest Linked Data graphs. Libraries that lack the necessary resources, need, or are otherwise not interested in transitioning to complete internal Linked Data operations could stop at the completion of *Phase One* and function effectively in the wider Linked Data library ecosystem.

The Primary focus of *Phase Two* is converting the entire library information ecosystem to native Linked Data operations. During this phase of conversion, the catalog itself is converted to a Linked Data, graph-based architecture and cataloging interfaces and workflows are altered to maximize realization of the descriptive, search and discovery, and workflow benefits of Linked Data.

V. Phase One: Linked Data in a MARC Ecosystem

The library information ecosystem is comprised of a complex web of applications, scripts, and workflows that handle the range of library operations including acquisitions, cataloging, circulation, and analysis. At the UC Davis Library, for example, there are 15 non-cataloging systems that exchange data directly with the ILS on a regular basis, and an additional 25 systems that drive library operations which depend on the cataloging data in the ILS. Any conversion strategy must deal not only with the transformation of cataloging data, but also with the various points of exchange and interaction between all of these systems.

Each of the above systems is also intimately tied to human workflows. As noted elsewhere, library operations are performed by highly trained, specialized staff with well-established workflows. Altering the tools used by employees could have drastic impacts on quality, efficiency, and speed. Additionally, retraining of staff could be necessary, adding significant cost to the transition.

Adding to this complexity is the fact that the majority of the software systems deployed by libraries are licensed software applications provided by external vendors. This means that the vendors themselves must alter these applications to work with Linked Data graphs instead of MARC records, or complex connectors must be built for native Linked Data systems to exchange information with non-Linked Data systems.

The *Phase One* transition plan is designed to mitigate the costs and risks associated with the transition by establishing a minimally viable Linked Data infrastructure upon which a future, more complete transition can be accomplished. A *Phase One* conversion consists of the following primary steps:

[Continued on Next Page]

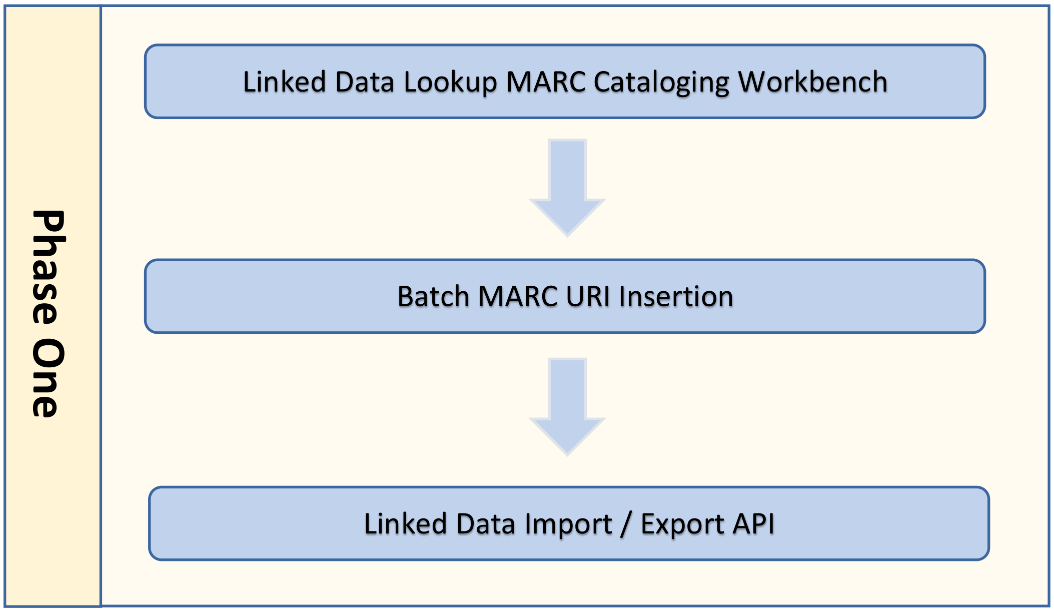

Phase One
Linked Data Lookup MARC Cataloging Workbench
Batch MARC URI Insertion
Linked Data Import / Export API

Figure 7: Transition phase one

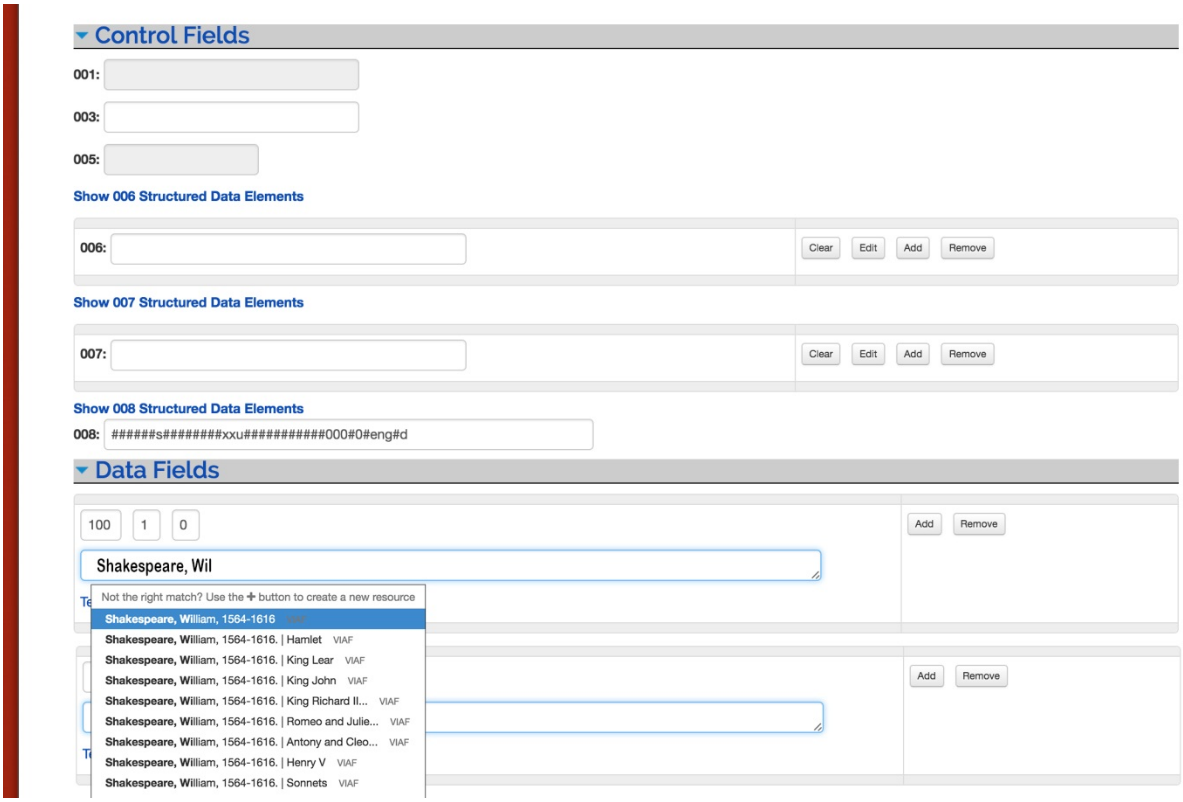
Step One: Linked Data Lookup MARC Cataloging Workbench

The first step in the *Phase One* process is to establish a system for capturing and inserting URIs into newly created and/or edited records. This initiates an important transition that insures that all current and future work efforts will support a Linked Data transition.

From a technical perspective, switching to a workflow that allows capture and insertion of URIs at the point of cataloging represents a minor modification to the ILS system. As part of the BIBFLOW study we were able to successfully modify the open source KUALI-OLE Describe Module to perform Linked Data gateway lookups on Library of Congress, OCLC, VIAF, and Getty Vocabularies and to insert captured URIs into MARC records with minimal effort. All gateway sources provide API documentation to facilitate query and retrieval. OCLC Research has also made available experimental Javascript code for performing lookups.[4]

[Continued on Next Page]

[4] See https://www.oclc.org/developer/news/2016/consuming-linked-data-using-javascript.en.html and https://github.com/oclc-developer-house/jquery-viaf-autocomplete.


Control Fields
Show 006 Structured Data Elements
Clear
Edit
Add
Remove
Show 007 Structured Data Elements
Show 008 Structured Data Elements
Data Fields
Add
Remove
Not the right match? Use the ✚ button to create a new resource

001:

003:

005:

006:

007:

008: ######s########xxu###########000#0#eng#d

100 1 0

Shakespeare, Wil

Shakespeare, William, 1564-1616 VIAF
Shakespeare, William, 1564-1616. | Hamlet VIAF
Shakespeare, William, 1564-1616. | King Lear VIAF
Shakespeare, William, 1564-1616. | King John VIAF
Shakespeare, William, 1564-1616. | King Richard II... VIAF
Shakespeare, William, 1564-1616. | Romeo and Julie... VIAF
Shakespeare, William, 1564-1616. | Antony and Cleo... VIAF
Shakespeare, William, 1564-1616. | Henry V VIAF
Shakespeare, William, 1564-1616. | Sonnets VIAF

Figure 8: Modified Kuali-Ole Describe interface showing VIAF Linked Data gateway lookup

Commercial ILS vendors are also working towards providing URI lookup and insertion as part of their standard offering. Ex Libris' Alma, for example, UC Davis's current ILS system, currently plans to offer this functionality. Additionally, all major ILS vendors and OCLC are currently running either public or internal pilot programs directed at providing Linked Data enabled versions of their products. Given the above, the human effort and associated costs of making this transition are minimal.

The workflow and systems impact of this transition on libraries currently using cloud-based ILS will be negligible. In this case, the technology overhead of the transition falls to the ILS vendor, and the staff training required to disambiguate from an authority file with no imbedded URIs to one with them is nil. This is similarly true for those using open source ILS. Running a local, open source ILS requires the in-house technical expertise to implement URI lookup and disambiguation developed internally or externally to the organization; however, our experimentation shows that this can be accomplished with minimal effort.

The libraries which will have the most difficulty in implementing this step in the first phase of the transition are those libraries currently running a non-cloud-based ILS. Moving to a URI enhanced ILS will require: 1) waiting until a URI enabled version of the ILS is available; and 2) implementing the new version.[5] Libraries that fall into this category could, thus, not begin a

[5] Note that another option here is to switch ILS, but this option is not covered in this report as it would most like be considered only in situations where an ILS change is already considered, and the impacts of switching ILS lie beyond the scope of this roadmap.

transition until such time as their vendor releases a URI enabled version of the browser; and, when the transition is made, upgrading to the new version would require a moderate level of internal technology effort.

Step Two: Batch MARC URI Insertion

Step One of *Phase One* conversion plan establishes a working environment in which all future-forward cataloging efforts will support Linked Data transition. *Step Two* of *Phase One* addresses the problem of legacy records. In October 2015 the Program for Cooperative Cataloging (PCC) charged a Task Group on URIs and MARC.[6] The specific charge of the Task Group was to investigate the feasibility of and make recommendations regarding the insertion of URIs in standard MARC records. Much of the Task Group's work focused on testing the potential impact of inserting URIs into MARC records, with an eye particularly to testing whether or not such an effort would negatively affect the functioning of current ILS systems. This testing necessitated the large-scale conversion of MARC records. To this end, librarians and staff at George Washington University, working under the guidance of the PCC Task Group's chairperson, Jackie Shieh, tested various methods of inserting URIs in the MARC records of their 1.7 million title catalog.[7]

The published results of George Washington University's experiments with URI insertion provide details regarding the exact process used as well as scripts for performing the insertion. As such, these specific details are not included in this report. Relevant to this report is the calculation of effort required to complete the transformation. The most successful method implemented by the George Washington University team involved automated conversion and validation followed by human validation, correction, and supplemental cataloging. According to Shieh and Reese, automated conversion of records resulted in few errors. Human catalogers were used to spot check machine output. One cataloger was devoted to this task for the duration of the project, resulting in a very high, verified rate of conversion accuracy.

A potential option for completing *Step Two* of *Phase One* of the conversion plan would be to share the conversion effort across libraries both through and with OCLC and other vendors. The present workflows of most libraries involve contributing and receiving records from OCLC and other vendors. There is opportunity for service models in which OCLC inserts URIs in bibliographic records and distributes the updated records to libraries as appropriate. Additionally, vendors could provide records for shelf-ready acquisitions that include records with URIs.[8] The costs of conversion as a service model are impossible to calculate without direct input from vendors; however, as such a service would dramatically reduce the work

[6] See https://www.loc.gov/aba/pcc/bibframe/TaskGroups/URI-TaskGroup.html

[7] Shieh, Jackie, and Terry Reese. "The Importance of Identifiers in the New Web Environment and Using the Uniform Resource Identifier (URI) in Subfield Zero ($0): A Small Step That Is Actually a Big Step." Journal of Library Metadata 15.3-4 (2015): 208-26. Web.

[8] At a recent IMLS funded gathering hosted at Cornell University devoted to discussion of authority control in the Linked Data ecosystem, vendor and publisher representatives indicated that they are interested in pursuing discussions with libraries along these lines.

effort required at each local institution, the resultant cost should represent a cost savings to participating libraries.

Step Three: Linked Data Import/Export API

The final step in *Phase One* of the Linked Data conversion plan involves providing gateways for the publishing and ingest of Linked data bibliographic and holding graphs. This involves exposing the catalog using either external or pass-through APIs that reformat MARC records as Linked Data graphs on export and reformat Linked Data graphs as MARC records on ingest. All major ILS systems include APIs for reading and writing data to the catalog. Individual libraries should check with their ILS vendor to determine if these APIs include Linked Data capabilities. Many ILS do not currently have Linked Data APIs; and currently no ILS includes production-ready BIBFRAME gateways.

Libraries that currently use ILS that have Linked Data APIs will be able to immediately operate in a Linked Data world. Those who do not will need to implement pass-through APIs that read BBIBFRAME and convert it to a format acceptable to their ILS API before passing the request on to the API and *vice versa*. The creation of the pass-through APIs will require technical expertise to implement and maintain. This could stand as a barrier to entry for smaller libraries that lack internal development expertise. However, it is possible that, once developed, these APIs could be shared, reducing time and expertise needs across the library community. Additionally, as BIBFRAME becomes more stable (which it is already doing), we can expect vendors to roll-out BIBFRAME native APIs to their ILS. In these cases, the time and cost impact would most likely be minimal.

Phase One Completion

Phase One completion represents a significant milestone in the transition to Linked Data operations. At the conclusion of this phase, libraries will be situated such that their entire record collection and ongoing record creation and maintenance will support Linked Data operations, and they will be able to deliver and ingest Linked Data records. Implementation timelines for *Phase One* are dependent on vendor implementation timelines for all but those libraries that currently implement open source ILS and have the expertise to add the needed functionality to the ILS. Some commercial ILS systems already contain the necessary URI lookup and insertion functionality. In all circumstances, the cost of implementing *Phase One* is minimal, as is the effect on cataloging workflows.

VI. Phase Two: Transition to a Native Linked Data Ecosystem

For the purposes of this report, a “Native Linked Data Ecosystem” is defined as one which exchanges data with other institutions as *serialized n-triples* (the most familiar form of which being RDF) and offers a Linked Data connected, oriented, and extensible cataloging workbench. While it is recommended, note that under this definition, it is not necessary that a system’s underlying data store be triples based. Contrary to popular belief, few ILS currently implement a truly MARC-based data store. User interfaces to the data are MARC oriented, but the data structures themselves are not. Well-designed software systems are comprised of three distinct components, or layers: 1) Data Layer, known as the Model; 2*)* User Interface Layers (human and machine), known as the View; and 3) Transaction Processing Layer, known as the Controller:

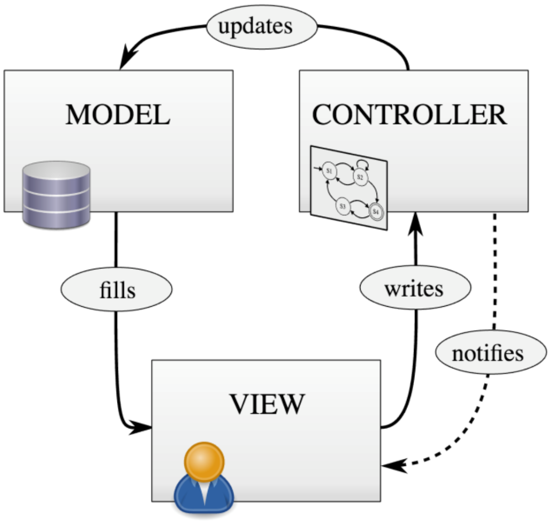


Figure 9: Model View Controller (MVC) architecture[9]

The View is the on-screen (GUI or command line) interface through which human and machine users interact with the rest of the application. This includes display screens, forms, APIs, etc. If you are reading this document electronically, the window in which you currently see this text is a component of the View. The Controller includes any components of the code that perform operations on data available to the application. In a PDF viewer, for example, this includes reading the raw data in the file and transforming it to a form that can be rendered by the View. Another example would be a program that calculates the mean of a series of numbers or converts a string to lower case. The actual computing process that performs these actions are part of the Controller*.* Last but not least, the *Model* is the data structure that an application uses to store data. A Model could be a collection of .CSV or XML files, a relational database, a graph database, or any other data storage schema.

Because the Views employed by current ILS systems are MARC oriented, the library community tends to think that ILS data Model is also MARC based. This is rarely the case. No widely implemented ILS (or sub components for modular environments) is MARC based at the Data Layer.[10] Most current systems store data in relational databases or other indexed document stores that bear only a passing resemblance to MARC itself. For example, the Kuali-OLE data store is comprised of 10,644 fields in 1,499 related tables—far greater than the

[9] By Grégoire Surrel, initial work by Deltacen [CC BY-SA 3.0 (http://creativecommons.org/licenses/by-sa/3.0)], via Wikimedia Commons

[10] The 1.0 release of Kuali-OLE was based on a MARC-XML document store, but performance considerations resulted in migration to a SQL based Data Layer for future releases.

number of fields and subfields in the MARC specification.[11] Similarly, MARC manipulation tools like MarcEdit rely on a SQL Data Layer to perform much of their work.

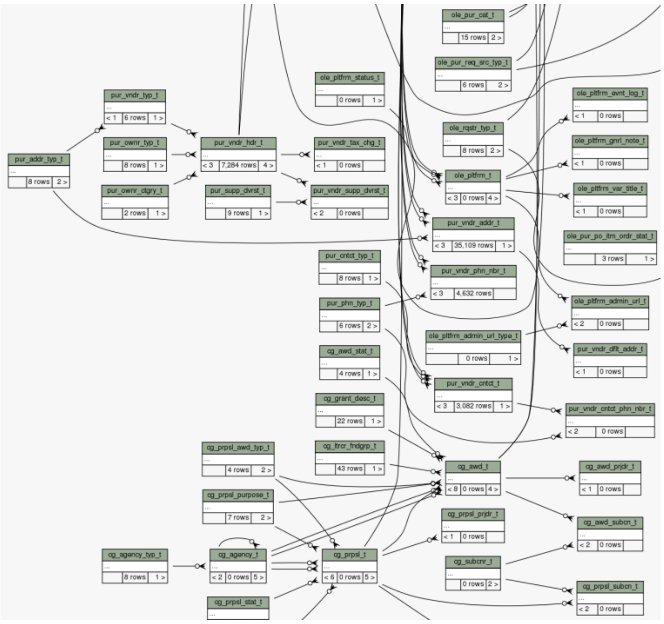


Figure 10: Portion of OLE SQL database structure[12]

Simply put, there is little direct relationship between the data Model and application View of most ILS. As such, it would be possible to implement a Linked Data graph Model without changing MARC-oriented Views at all. Similarly, it is possible to change Views to reflect a Linked Data, graph-based orientation to data creation and management while still using a relational database as the applications data Model. There are very good reasons why converting the application to a graph Model is preferable for operating in a Linked Data environment, but these reasons are largely technical in nature and beyond the scope of this report. What is important for the current purpose is recognizing that transitioning to a graph-based data Model is not a pre-requisite to operation in a fully Linked Data ecosystem.

Libraries must complete *Phase One* of the transition roadmap before commencing *Phase Two*, which consists of the following steps:

[Continued on Next Page]

[11] See http://ole-build-01.lib.duke.edu/ole-db/

[12] Courtesy of Jeff Flemming, Duke University Library: http://ole-build-01.lib.duke.edu/ole-db/relationships.html

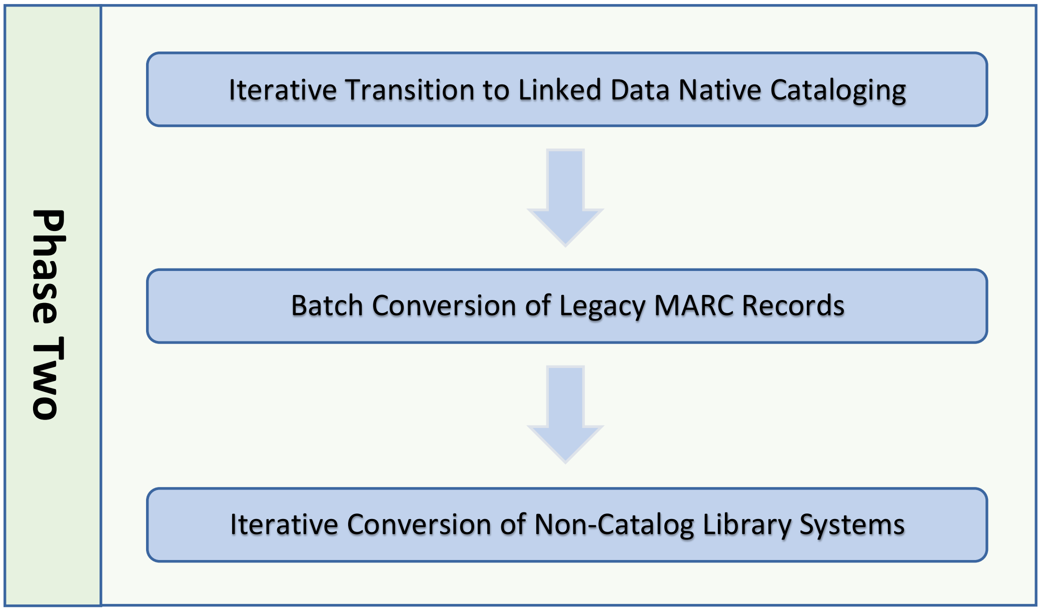
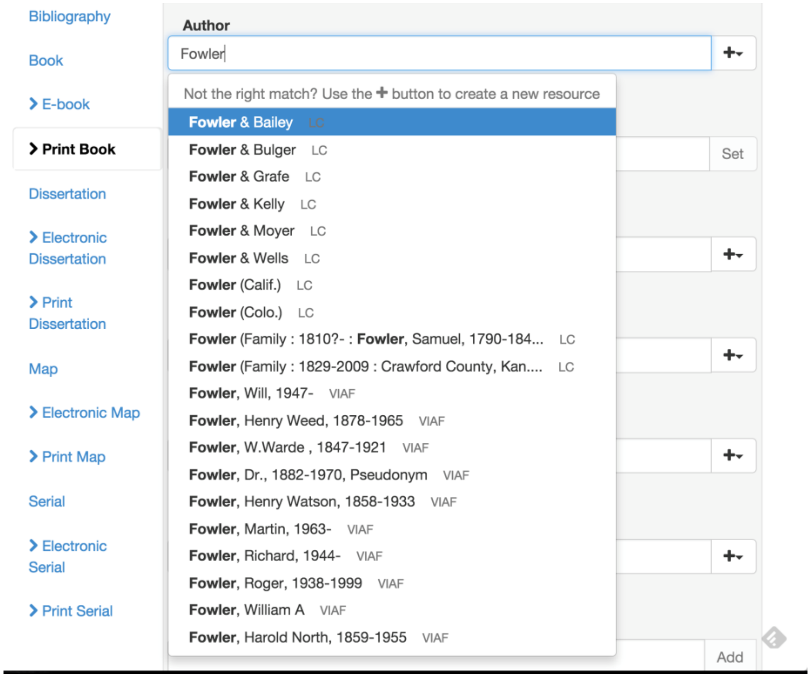

Phase Two
Iterative Transition to Linked Data Native Cataloging
Batch Conversion of Legacy MARC Records
Iterative Conversion of Non-Catalog Library Systems

Figure 11: Transition Phase Two

Step One: Staged Transition to Linked Data Native Cataloging

Step One of *Phase Two* of the transition roadmap is focused on migration of cataloging workflows to a Linked Data native cataloging workbench. By "Linked Data Native" we mean an interface that is designed specifically to interoperate with external Linked Data information resources as an integral part of cataloging workflows and that capitalizes on the extensibility offered by working with graph-based data models. As part of this study, we experimented with several such interfaces.

[Continued on Next Page]



Author

Fowler

Not the right match? Use the ✚ button to create a new resource

- **Fowler** & Bailey LC
- **Fowler** & Bulger LC
- **Fowler** & Grafe LC
- **Fowler** & Kelly LC
- **Fowler** & Moyer LC
- **Fowler** & Wells LC
- **Fowler** (Calif.) LC
- **Fowler** (Colo.) LC
- **Fowler** (Family : 1810?- : **Fowler**, Samuel, 1790-184... LC
- **Fowler** (Family : 1829-2009 : Crawford County, Kan.... LC
- **Fowler**, Will, 1947- VIAF
- **Fowler**, Henry Weed, 1878-1965 VIAF
- **Fowler**, W.Warde , 1847-1921 VIAF
- **Fowler**, Dr., 1882-1970, Pseudonym VIAF
- **Fowler**, Henry Watson, 1858-1933 VIAF
- **Fowler**, Martin, 1963- VIAF
- **Fowler**, Richard, 1944- VIAF
- **Fowler**, Roger, 1938-1999 VIAF
- **Fowler**, William A VIAF
- **Fowler**, Harold North, 1859-1955 VIAF

Figure 12: BIBRAME Scribe cataloging interface

Figure 12 shows a screenshot for a native Linked Data cataloging interface developed by Zepheira, Inc. and modified for testing by the UC Davis BIBFLOW team. Unlike MARC-based cataloging interfaces, the Scribe interface presents the cataloger with a Linked Data oriented view of the Universe. Rather than filling out form fields as appropriate as a means of defining the format, for example, of a particular object, Scribe asks the cataloger to first identify the kind of object being described. Once this has been done, it presents the user with a View based on one of many Linked Data Profiles, Linked Data models appropriate to the type of object being described. Each Profile contains a map of relevant Linked Data lookup services, and the View reflects this by providing type-ahead functionality on appropriate fields. Importantly, Profiles are highly configurable, allowing libraries to record extensible descriptions of objects. For example, one might combine traditional MARC-based content fields with Electronic Archival Description (EAD) descriptors in the same graph, something non-graph based systems cannot accommodate without extensive modification of the application Model itself.

[Continued on Next Page]

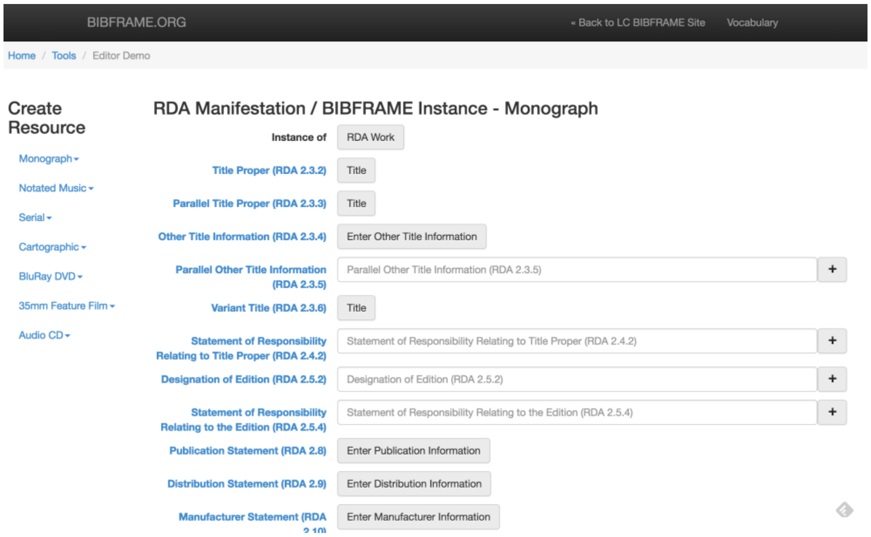

BIBFRAME.ORG
« Back to LC BIBFRAME Site
Vocabulary
Home / Tools / Editor Demo
Create Resource
Monograph
Notated Music
Serial
Cartographic
BluRay DVD
35mm Feature Film
Audio CD
RDA Manifestation / BIBFRAME Instance - Monograph
Instance of
RDA Work
Title Proper (RDA 2.3.2)
Title
Parallel Title Proper (RDA 2.3.3)
Title
Other Title Information (RDA 2.3.4)
Enter Other Title Information
Parallel Other Title Information (RDA 2.3.5)
Parallel Other Title Information (RDA 2.3.5)
+
Variant Title (RDA 2.3.6)
Title
Statement of Responsibility Relating to Title Proper (RDA 2.4.2)
Statement of Responsibility Relating to Title Proper (RDA 2.4.2)
+
Designation of Edition (RDA 2.5.2)
Designation of Edition (RDA 2.5.2)
+
Statement of Responsibility Relating to the Edition (RDA 2.5.4)
Statement of Responsibility Relating to the Edition (RDA 2.5.4)
+
Publication Statement (RDA 2.8)
Enter Publication Information
Distribution Statement (RDA 2.9)
Enter Distribution Information
Manufacturer Statement (RDA 2.10)
Enter Manufacturer Information

Figure 13: Library of Congress BIBFRAME Editor

The Library of Congress BIBFRAME Editor offers a different approach to a Linked Data native cataloging interface.[13] It focuses on creating Linked Data graphs while maintaining labels that reflect current cataloging rules (*ie.* RDA). It also builds on the BIBFRAME Work/Instance model.

Both of the above Linked Data cataloging workbenches are standalone products that output Linked Data graphs. Another approach to this transition could be the addition of native Linked Data workbenches to existing ILS. The addition of URI maintenance into current library workflows of several ILS (discussed in *Section Three*) marks a step in this direction. But adding an extensible interface capable of handling multiple profiles, communicating with a growing collection of Linked Data endpoints, and reflecting the Work/Instance BIBFRAME model will require significant effort on the part of the vendors who supply these ILS.

Linked Data adoption also opens the door to new, more automated modes of cataloging. As part of the BIBFLOW project, we experimented at UC Davis with systems that utilized available link data endpoints to construct catalog graphs on the fly.

[Continued on Next Page]

[13] See https://www.loc.gov/bibframe/implementation/index.html

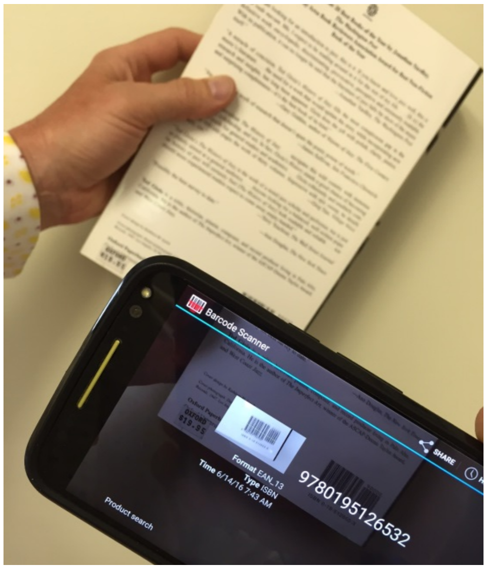


Figure 14: Barcode cataloging

Our barcode cataloging system allowed us to extract ISBN information by scanning a book barcode. The ISBN was used to make a series of queries to OCLC and Library of Congress Linked Data endpoints. When needed, a popup screen would ask catalogers to disambiguate information. At the completion of the process the appropriate graph was added to the triplestore. The system increased both efficiency and accuracy of bibliographic and holding data.

Regardless of the approach to native Linked Data cataloging pursued, there will be some constants. First, the transition will come at some cost. Libraries that host and maintain local ILS will be required to migrate the ILS to new, Linked Data native system. Libraries that use cloud-based ILS can similarly expect to pay for migration to new, cloud-based systems, as migrations of this magnitude legitimately constitute a release of a new system.

Regardless of the path to native Linked Data cataloging taken or the form of the data Model employed, new Linked Data workbenches must function in concert with existing MARC based systems. As noted in *Section III*, ILS and other library systems operate as part of a complex information ecosystem where data is exchanged regularly between systems. It is neither desirable nor likely that all of these systems will convert to a Linked Data exchange model at the same time. As such, libraries should expect to operate in a hybrid ecosystem for some time, where both Linked Data graph and MARC records exist in parallel.[14] Providing this parallelism requires coding efforts that are not incidental. As part of BIBFRAME's experimental effort, we were able to build bi-directional connectors between BIBFRAME Scribe's graph database and Kuali-OLE's relational database. These connectors functioned such that any time

[14] This issue is discussed in detail in *Section VII*, below, including system diagrams showing the nature of the relationship between the MARC Records and graph based models and their supporting systems.

a new Linked Data graph was created (whether by human cataloging or batch conversion) a "stub" MARC record was created in OLE, containing all necessary information to perform regular functions such as search, discovery, and circulation. Similarly, whenever a record was created or loaded into OLE, a parallel graph was saved to the graph database. Similar bidirectional functionality was added for edits as well.

While the above described parallel universe appears to create a great deal of unnecessary duplication and redundancy, its benefits outweigh this cost. Implementing a parallel system allows iterative conversion of both workflows and systems. Rather than having to convert all systems and workflows involved in the exchange of MARC records or MARC-based cataloging at one time, individual workflows and systems can be migrated to native Linked Data operation one at a time. At this systems level, this means that the transition can be made over time with a smaller, long-term or permanent staffing impact. This reduces the overall cost of the transition.

Running parallel, synchronized MARC/Graph data stores also increases efficiency and decreases the cost of migrating cataloging workflows from MARC to native Linked Data oriented workflows. With this model, migration can be accomplished by retraining and migrating small groups of staff at a time as opposed to attempting to train all cataloging staff and migrate the entire cataloging effort at one time. This reduces the impact on ongoing work efforts, all of which would be simultaneously affected during a mass transition, effectively shutting down work efforts during the transition. Additionally, managers and trainers will learn from each iteration, improving the efficiency of training and transition with each iteration. Further details of this iterative approach are provided in *Section VII*: Transitioning Workflows.

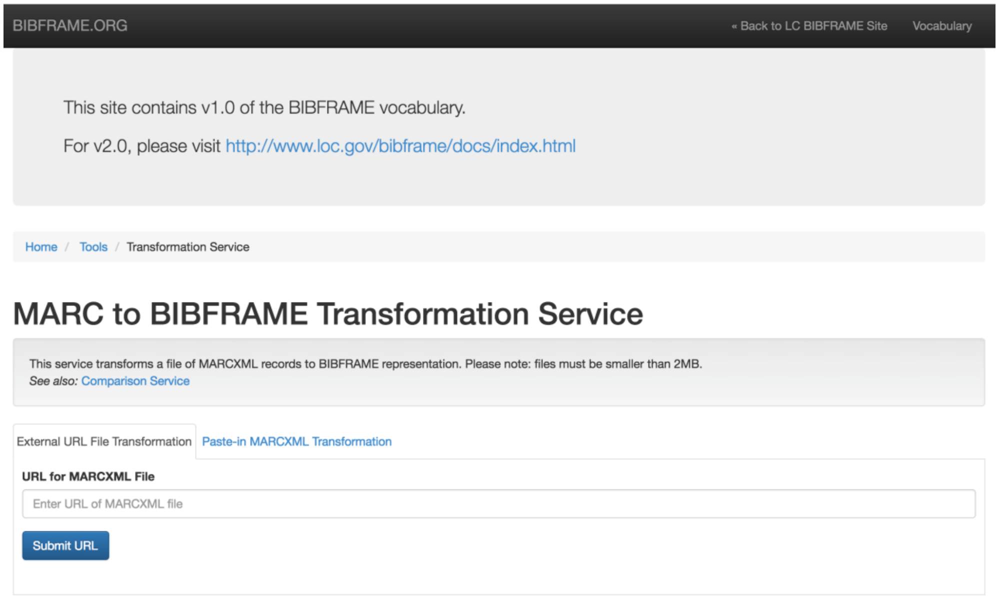
Step Two: Batch Conversion of Legacy MARC Records

Concurrently with, or after, migrating human workflows to native Linked Data operation, legacy MARC records must be converted to Linked Data graphs and stored in the new graph database. (As noted before, this database may not be strictly graph based, but the MARC records must be migrated to the new model regardless.) Automated transformation is made possible because needed URIs were added to MARC records during *Phase One* of this transition plan. This process will primarily involve technical staff, but libraries should expect to devote one cataloger familiar with both MARC and BIBRAME (or an alternate Linked Data model) to the effort in order to facilitate proper data mapping and to validate output.

Several viable tools are currently available for performing conversion of MARC records to Linked Data graphs.

[Continued on Next Page]

Library of Congress Transformation Service:


BIBFRAME.ORG
« Back to LC BIBFRAME Site
Vocabulary
Home / Tools / Transformation Service
External URL File Transformation
Paste-in MARCXML Transformation
Submit URL

This site contains v1.0 of the BIBFRAME vocabulary.

For v2.0, please visit http://www.loc.gov/bibframe/docs/index.html

MARC to BIBFRAME Transformation Service

This service transforms a file of MARCXML records to BIBFRAME representation. Please note: files must be smaller than 2MB.
See also: Comparison Service

URL for MARCXML File: Enter URL of MARCXML file

Figure 15: Current Library of Congress MARC to BIBFRAME Transformation Service

The current release of the Library of Congress MARC to BIBFRAME Transformation Service is a web-based service suitable for testing conversion from MARC to BIBFRAME 1.0. The Library of Congress is currently working on an open source, BIBFRAME 2.0 version of the software that can be installed locally and used to transform MARC to BIBRAME 2.0, the latest BIBFRAME standard. This software is soon to be released. The MARC to BIBFRAME Transformation Service has undergone extensive testing at the Library of Congress and will provide excellent MARC to BIBFRAME transformation. The software runs efficiently and produces a minimal required storage footprint. Additionally, the transformation engine is highly flexible, using an XSLT transformation service to traverse a MARC-XML DOM and output data in any text-based format. The Library of Congress provides XSLT for MARC-BIBFRAME conversion only, but with custom developed XSLT services the software could export transformations using any single or combination of ontologies and frameworks and in any Linked Data serialization. As such it represents a good choice for libraries interested in producing strict BIBFRAME with few alterations and for libraries with in-house XSLT expertise that are interested in converting to frameworks other than or in combination with BIBFRAME.

[Continued on Next Page]

MarcEdit:

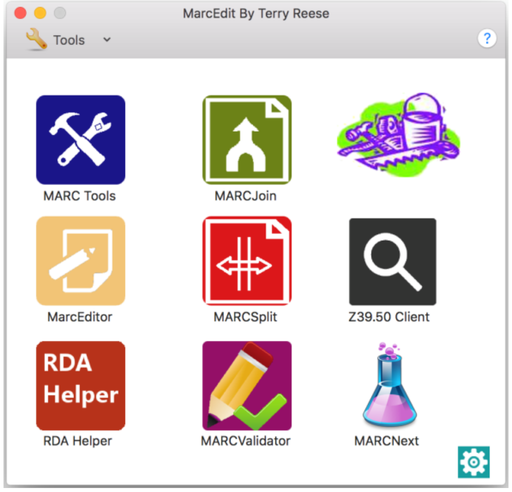


Figure 16: MarcEdit

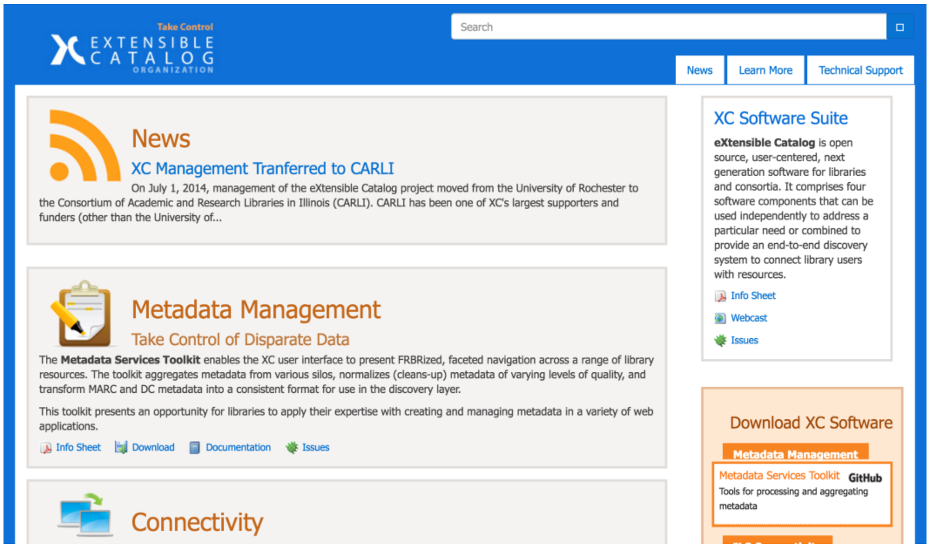
Most librarians are already familiar with Terry Reese's MarcEdit software. An import feature of MarcEdit is its MARCNext component, which provides a collection of tools for manipulating MARC with an eye towards Linked Data transformation. Two particular tools are of use in this regard: 1) a highly configurable transformation service; and 2) the ability to export MARC records as a SQL database.

MarcEdit's transformation engine is highly flexible, using an XSLT transformation service to traverse a MARC-XML DOM and output data in any text-based format. This could include RDF-XML, Turtle, or any other form of Linked Data representation. Using this system's libraries, one can easily run multiple transformations on the same collection of MARC records. This allows libraries to produce specific outputs for specific uses. For example, a library could run transformation as BIBFRAME for interlibrary use and another as Schema.org for search engine optimization. Additionally, Terry Reese also maintains a public forum where XSLT transformation scripts can be shared. This means that one library could use another library's BIBFRAME transformation out of the box, or modify it for a particular purpose and share with other libraries.

MarcEdit's ability to export MARC records as a collection of SQL scripts is also potentially quite useful. Exporting records to a SQL database opens the door for complex querying of data. Storing records in an accessible SQL database can simplify the transformation process for those libraries interested in writing their own, stand-alone transformation scripts or applications. All widely used scripting and programming environments have packages that provide easy access to a variety of SQL databases, simplifying the process of querying records as part of a transformation process.

MarcEdit provides a highly flexible platform for shared development of transformation script. As such, it is a good tool for libraries interested in performing multiple transformations and/or sharing in communal development of transformations. A potential drawback of the tool is that it is a Microsoft Windows only tool and can only be deployed on Windows based servers or desktops.[15] As such, it is only a suitable option for those libraries that operate in a Windows environment.

Extensible Catalog:



News

XC Management Tranferred to CARLI

On July 1, 2014, management of the eXtensible Catalog project moved from the University of Rochester to the Consortium of Academic and Research Libraries in Illinois (CARLI). CARLI has been one of XC's largest supporters and funders (other than the University of...

Metadata Management

Take Control of Disparate Data

The **Metadata Services Toolkit** enables the XC user interface to present FRBRized, faceted navigation across a range of library resources. The toolkit aggregates metadata from various silos, normalizes (cleans-up) metadata of varying levels of quality, and transform MARC and DC metadata into a consistent format for use in the discovery layer.

This toolkit presents an opportunity for libraries to apply their expertise with creating and managing metadata in a variety of web applications.

XC Software Suite

eXtensible Catalog is open source, user-centered, next generation software for libraries and consortia. It comprises four software components that can be used independently to address a particular need or combined to provide an end-to-end discovery system to connect library users with resources.

Download XC Software

Metadata Services Toolkit

Tools for processing and aggregating metadata

Figure 17: Extensible Catalog

The XC Software Suite is a suite of web applications focused on performing various transformation and connectivity functions. Like MarcEdit, The XC Metadata Services Toolkit (MST) provides a flexible engine for transforming MARC records into other formats. Whereas MarcEdit uses XSLT to perform transformations, the MST connects with ILS through the OAI-PMH protocol and then exposes records in a desired format based on customized Javascript transformations. Like MarcEdit, a community repository of transformation scripts is available, and can facilitate co-creation of scripts that allow libraries to expose record data in multiple forms.

The MST is a web-application that runs as a Java Servlet under server engines such as Apache Tomcat or Jeti. Administrative users use a web interface to manage transformation "Services" that map identified record sets to the Java transformation scripts. A valuable feature of the MST is that Transformations can be run one time only; or, the service can poll the ILS for changes and execute the transformation as need to keep the graph representation synchronized with the MARC data store. Transformed data sets are made available through an API. The MST can be run on any system that supports Java Servlets. This includes Linux, Mac, Unix, and Windows.

The MST is good option for libraries with in-house server administration technical expertise and the computing infrastructure necessary to run a Java Servlet container. An ILS that supports OAI-PMH is also required, or the ability to install and maintain a service that uses

[15] Note that there are versions available for other operating systems, but they do not offer the functionality of the Windows version.

APIs or exported MARC data to provide an OAI-PMH gateway. (The Extensible Catalog suite includes a MARC-XML to OAI-PMH gateway.) A particular disadvantage of Extensible Catalog's MST is it requires significant physical storage. In order to provide its synchronized transformation service, it maintains a local copy (SQL) of the entire catalog as pulled using OAI-PMH. As such, a single pipeline of transformation from the ILS to BIBFRAME results in three complete instantiations of the catalog: The original in the ILS, a copy in the MST SQL database, and the exposed BIBFRAME version.

Custom Application:

For libraries with robust technical services departments who are familiar with the various APIs of their various ILS, building a custom conversion tool could be an option. Our initial testing indicates that it will typically take from one to three months of full-time programming to code and test a fully functioning, stand-alone, custom conversion tool. Building a custom tool offers few advantages. It can, however be useful in cases where the records being converted are stored in more than one system or when attempting to combine records of different formats that reference the same object. For example, a not uncommon situation is for libraries holding special collections to maintain both a MARC record and an EAD record for the same object. Linked Data offers the opportunity to combine these two records into a single graph. In such cases, a custom application designed to communicate with both the MARC and EAD systems would be more efficient than using existing tools to create separate graphs and then applying a post-creation system of combining the graphs.

Third Party Service:

Zepheira Inc. will work with your library to either assist with or completely handle a transformation process. To date, Zepheira has worked with the Library of Congress, a host of public libraries, and the American Antiquarian Society, to name a few, to convert their existing MARC records. It can be expected that other vendors will also move into this space as the number of libraries planning on transforming records increases. Third party conversion services could focus on conversion of individual libraries or, taking advantage of economies of scale, provide a common, shared point of conversion and distribution. Libraries currently participate in shared cataloging through OCLC. A similar vendor service (OCLC is a natural point of service) that performs batch conversion and distributes converted records to libraries is a natural extension of the services that are already employed at libraries.

Step Three: Iterative Conversion of Non-Catalog Library Systems

The final step in the *Phase Two* Linked Data transformation is the conversion of non-cataloging library systems to Linked Data operations through either the development of necessary connectors or the adoption of Linked Data native versions of these systems as they become available. As with transitioning workflows, there is an advantage to pursing an iterative approach to this last phase of transformation. Attempting to transition all systems simultaneously would be highly disruptive to overall operations. It increases the likelihood of introducing a cascading error scenario where failures propagate across nodes in the information pipeline. This increases the impact of the inherent difficulty of troubleshooting. Transitioning one system at a time simplifies this process, localizing error potential, facilitating troubleshooting, and reducing potential impacts to the entire information ecosystem. Additionally, there are labor benefits to transitioning small teams at a time as opposed to transitioning the entire team over a short period of time. The small team approach offers management efficiencies and also simplifies human resources on-boarding and off-boarding.

VII. Transitioning Workflows

Cataloging is the process of creating metadata for libraries collections, whether owned or accessed. Workflows associated with cataloging largely depend on the ecosystem in which cataloging activities take place. The BIBFLOW project examined the effects and opportunities created by transitioning cataloging to a native Linked Data ecosystem by examining the following workflows:

1. Copy cataloging of a non-rare book
2. Original non-rare book cataloging
3. Original cataloging of a print serial
4. Original cataloging of a print map
5. Personal and corporate name authority creation

The study method employed was to document the current workflows in place at the UC Davis library, followed by testing of various approaches to the same cataloging tasks using native Linked Data cataloging workbenches. In each case, an eye was directed toward efficiency, accuracy, and the training required for catalogers to work in the new ecosystem. The workflows tested were chosen because they are representative of the range of cataloging practice employed in the library.

Workflows for authority creation and management are covered in the *Section VIII* of this report below. The remaining tested workflows are discussed in this section. Generally speaking, it was found that catalogers had little difficulty transitioning to a Linked Data ecosystem. The amount of training required was equivalent to that of transitioning from one MARC-based interface to another. With the exception of serials cataloging, discussed below, either a comprehensive knowledge of the technical details of Linked Data nor of the BIBFRAME model were required for catalogers to work successfully in the new environment. Additionally, cataloging in the Linked Data ecosystem offered various efficiencies in some workflows.

While completing *Step One* and *Step Two* of the transition plan outlined in this report, the Linked Data ecosystem consists of the following six components:

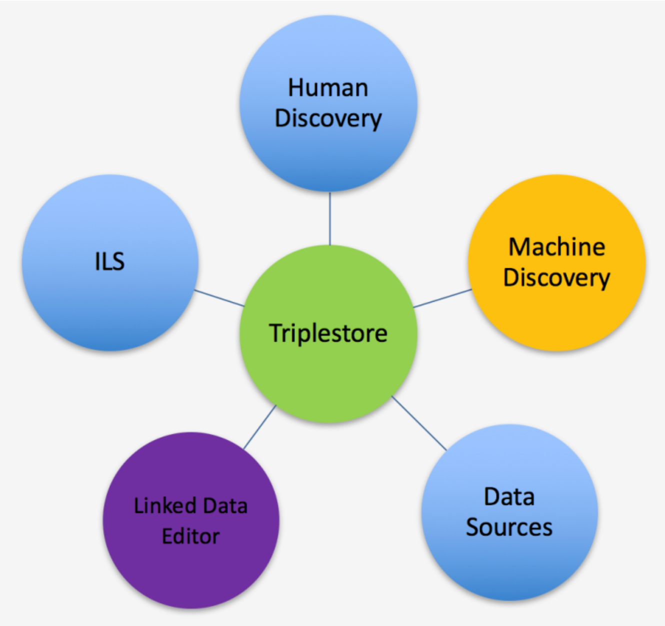


Figure 18: Six components of Linked Data ecosystem

At the center of this ecosystem is the *Triplestore*: the database management system for data in BIBFRAME format (RDF triples).

Human Discovery is comprised of application(s) that facilitate the transactions between patrons and the library's triplestore. It should also support the retrieval of additional information from external resources pointed to by the URIs recorded in the local triplestore.

The Integrated Library System (ILS) is an inventory control tool used to manage library's internal operations only, such as ordering and payment, collection management, and circulation. In this model, it also serves as a stand-in for all external systems that communicate with the library's catalog data. At the conclusion of *Phase One* of the transition plan, it will comprise a collection of applications that perform various functions such as acquisition, circulation, bibliometrics, etc. These systems may evolve to work directly with the triplestore, or they will continue to communicate with the triplestore through an API.

The Linked Data Editor is a tool that supports cataloging activities (metadata creation and management). At a minimum, an editor should have: 1) a user-friendly interface that does not require the cataloger to have a deep knowledge of the BIBFRAME data model or vocabularies; and 2) lookup services that can be configured to search, retrieve, and display Linked Data from external resources automatically.

Data Sources are resource locations available over the internet with which a Linked Data Editor can communicate in order to exchange data. These include endpoints such as OCLC WorldCat for bibliographic data and Library of Congress's Linked Data services for subject and name headings. To increase the likelihood of finding authoritative URIs and to make library data more interoperable, the community should also explore the use of non-library data and identifiers, such as ORCID, publisher's data, Wikidata, LinkedBrainz, etc.

Machine Discovery is a SPARQL endpoint that enables an external machine to query the library triplestore.

Figure 19 below illustrates the interactions among the six conceptual categories (OCLC and Authorities are used to represent “Data Sources”):

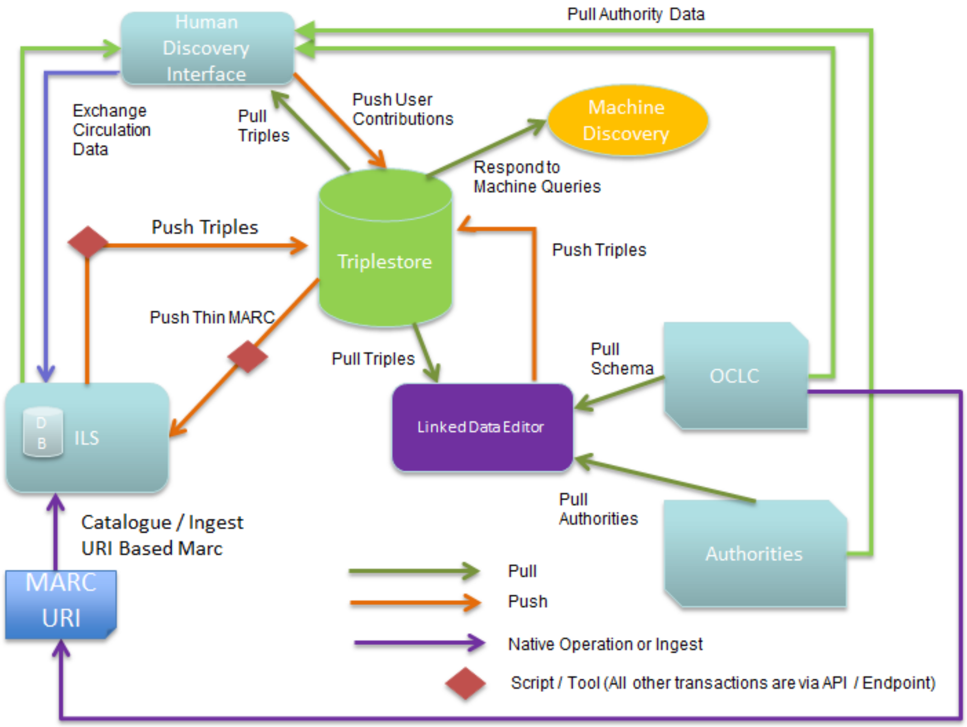


Figure 19: Interaction between the components of a Linked Data ecosystem

As can be seen, the information flows involved in a Linked Data ecosystem are more complex than in a MARC ecosystem. In the current MARC ecosystem, the Integrated Library System (ILS) acts as centralized information exchange point wherein external data is ingested and served through a single point of access. The Linked Data ecosystem dis-integrates the ILS. The triplestore serves as a partial, centralized data store, but graphs stored locally in the triplestore are supplemented on-the-fly by information provided by other Linked Data services and can be interacted with by a flexible suite of applications. The net result is a more complex data ecosystem, but one in which the workflows surrounding the data remain unchanged or are actually simplified.

Below we discuss the impacts of Linked Data adoption on three main types of cataloging workflows – copy, original, and serials cataloging. In each case we present proposed Linked Data native workflows and discuss how they relate to traditional MARC-based cataloging workflows. Readers will note that the two workflows presented are quite similar to their MARC ecosystem counterparts; however, each still presents its own issues and challenges. Some of the identified challenges may require further research and experimentation to address. Some may require the library community to rethink its cataloging rules and practices.

Copy Cataloging:

Linked Data copy catalogers will perform essentially the same tasks in a BIBFRAME ecosystem as they have traditionally in a MARC ecosystem: searching databases, finding existing bibliographic data, making local edits, checking access points, and saving data into a local system. During a Phase One implementation as defined in *Section V* above, the only required additional step is to synchronize thin MARC records with the existing ILS. The diagrams below illustrate the steps (workflow) used to perform copy cataloging. For demonstration purposes, OCLC WorldCat is used as an example of an external Linked Data data source (OCLC publishes its bibliographic data in Schema.org) and the BIBFLOW Scribe interface (as discussed in *Section VI* above) is assumed as a Linked Data cataloging workbench:

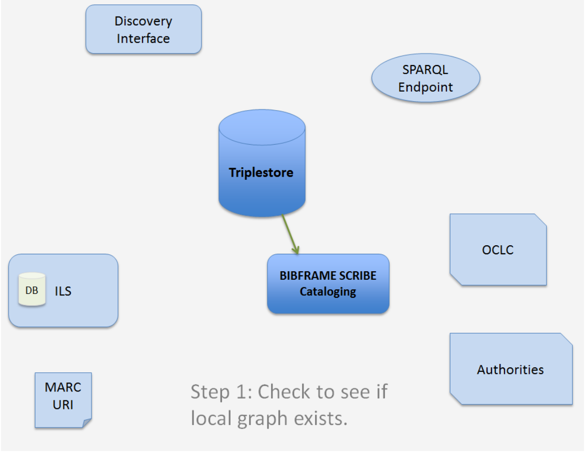


Figure 20: Step One of Linked Data copy cataloging workflow

In Step one, the copy cataloger uses the interface to see if a local bibliographic graph already exists for the item being cataloged. If a local graph does exist, a new local Holding is added to the local triplestore. If not, the cataloger moves to Step Two:

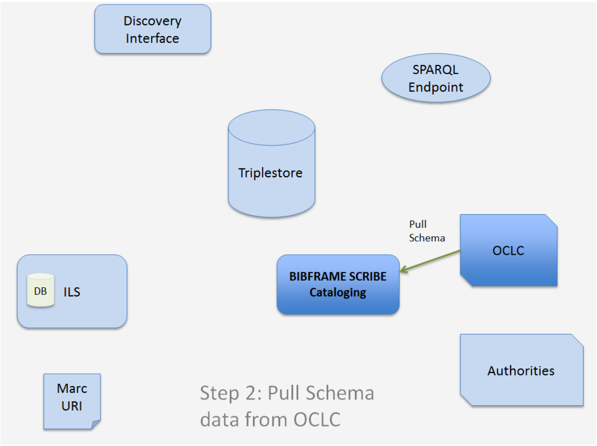


Figure 21: Step Two of Linked Data copy cataloging workflow

Step Two involves retrieving data about the item being cataloged from OCLC. This can be performed in one of two ways. *Figure 14 in Section VI* above depicts a system tested as part of the BIBFLOW project that allows users to scan the barcode of an item and automatically retrieve OCLC graph data based upon the extracted ISBN. Similarly, the BIBFRAME Scribe tool allows a cataloger to manually input an ISBN to perform the same search, or to perform a Title and or Author search. In both cases, the cataloger may be required to disambiguate results, as a single ISBN or search return can reflect multiple Work graphs. This same disambiguation is similarly required in a MARC ecosystem, and does not reflect an additional effort. Once an appropriate OCLC Work record has been identified, the Linked Data cataloging interface retrieves the graph for that resource from OCLC. This graph includes all information currently stored in exchanged MARC records. When a graph is pulled, its data is used to auto-fill all fields in the cataloging workbench for review by the cataloger.

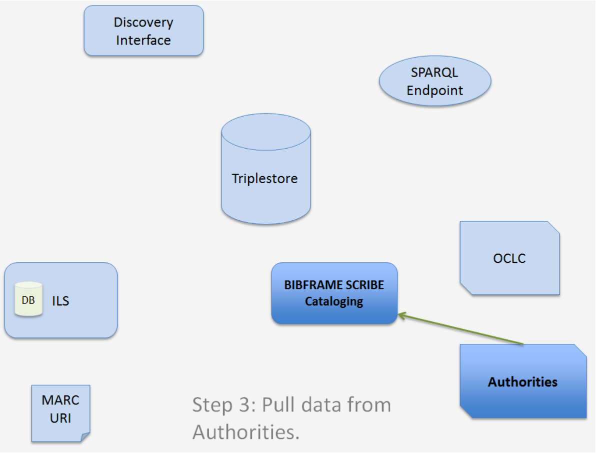


Figure 22: Step Three of Linked Data copy cataloging workflow

Step Three involves using similar lookup functionality to automatically discover URIs for authority entries. Using services such as VIAF, Library of Congress Authorities, and Getty Authorities, catalogers can search for authorities using human readable forms and automatically pull Linked Data representations of the authority, including URIs.

[Continued on Next Page]

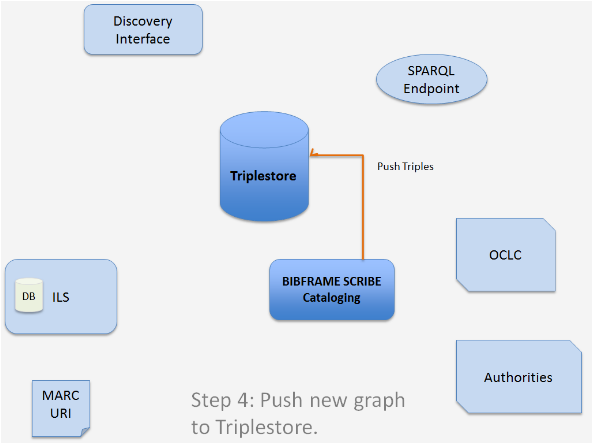


Figure 23: Step Four of Linked Data copy cataloging workflow

Once the cataloger is satisfied with the graph data pulled from OCLC and any made modifications, the final step in the human cataloging workflow is to push the new graph to the triplestore. In the case of items for which there is currently a local bibliographic graph, this involves adding an appropriate bibliographic record to the database as well as required Instance and Holding data. In a completely native Linked Data ecosystem, one in which all systems that surround the library's cataloging data have been converted to communicate directly with the triplestore, Step Four is the final step in the copy cataloging process. In cases where the cataloger is working in a hybrid ecosystem (prior to the completion of *Phase Two* as defined in *Section VI* above), a final, machine-automated step will be required:

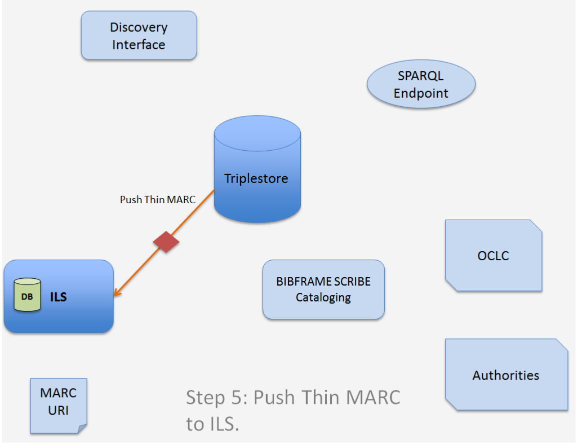


Figure 24: Step Five of Linked Data copy cataloging workflow

In cases where the library is currently not operating in a completely Linked Data ecosystem, when a cataloger pushes a new graph to the triplestore (or modifies an existing one), these changes must be propagated to any systems still relying on MARC data. This transaction is handled by a machine process and requires no human interaction.

As illustrated above, transition to a Linked Data ecosystem has no negative impact on the human workflows involved in copy cataloging and will improve efficiency in many cases due to the ability to auto-lookup and create graphs for items. Specific benefits of Linked Data copy cataloging include:

1. Catalogers do not need to search OCLC database separately because the lookup services embedded in the Linked Data cataloging workbench can retrieve both bibliographic and authority data, with associated URIs, and automatically put retrieved data into appropriate fields (auto-populate)
2. Catalogers do not need to have in-depth knowledge of BIBFRAME data model or BIBFRAME vocabularies because the data mapping between Data Source (e.g. OCLC - Schema.org) and BIBFRAME has been done behind the scenes
3. Catalogers do not need to input URIs manually because the machine will record and save them into the triplestore automatically; they just need to identify and select the correct entry associated with a URI
4. Automated methods such as barcode scanning can be used to perform record creation in a fraction of the time currently required

One potential issue stands as a barrier to proper BIBFRAME implementation using the proposed model. Schema.org (the Linked Data framework used by OCLC) does not differentiate title proper from the remainder of the title, but they are differentiated in the BIBRAME specification. For our implementation, we opted to include the complete Schema.org title in the BIBFRAME Title Proper element. This approach was taken because a full text search (or index) of a combined title element would return a successful search for any portion of the title. Given the nature of current full-text search capabilities, more discussion about whether multiple title elements are still useful and, if so, how to reconcile OCLC and LOC data will be necessary.

Transitioning to Linked Data cataloging using the proposed model raises the following questions for community consideration:

1. As per the discussion immediately above and given the nature of current full-text search capabilities, more discussion about whether multiple title elements are still useful and, if so, how to reconcile OCLC and LOC data will be necessary
2. How much data is needed in local triplestore? If most of the things can be identified by their associated URIs, and library discovery systems that sit on top of the local triplestore can pull information from external resources, how much data does the library still want or need in its local system?
3. If changes are made to source data, is it necessary to send the revised information back to the sources? If yes, what will we need to make this happen as an automatic process?

Original Cataloging

An original cataloging situation occurs when a cataloger is unable to locate, either locally or through an external authority, existing bibliographic data for the item being described. The process outlined above for copy cataloging an item included several options for searching both locally and through an external source (OCLC) for existing bibliographic graphs related to the item with which the cataloger is working. External lookup sources could include OCLC, publisher Linked Data endpoints, and even non-traditional data sources such as booksellers and Wikipedia. In the course of a cataloger's workflow, it is possible that no or partial data only can be found for an object. In this case, the cataloger must switch to an original cataloging workflow.

Once a cataloger has switched to an original cataloging workflow, very little will change from current original cataloging methods. The task of describing the details of the item being described will remain the same; however, cataloging in a Linked Data environment offers some distinct efficiency in the original cataloging workflow.

As discussed in *Section V* and *Section VI* of this report, Linked Data enabled cataloging workbenches have the ability to provide automatic lookup of entities at a variety of Linked Data endpoints such as OCLC, the Library of Congress, and Getty. This auto lookup feature facilitates original cataloging such that users can locate, disambiguate, and enter relevant data in a variety of fields that will be used to complete the bibliographic graph for an item. Current MARC-based cataloging systems employ similar functionality based on authority file lookup. When proper authority references are found, transitioning to Linked Data cataloging is a zero-sum-gain scenario. However, Linked Data cataloging offers workflow efficiencies in situations where no appropriate authority references can be found.

Currently, a cataloger confronted with the need for a nonexistent authority is faced with one of the following two workflows:

Option 1

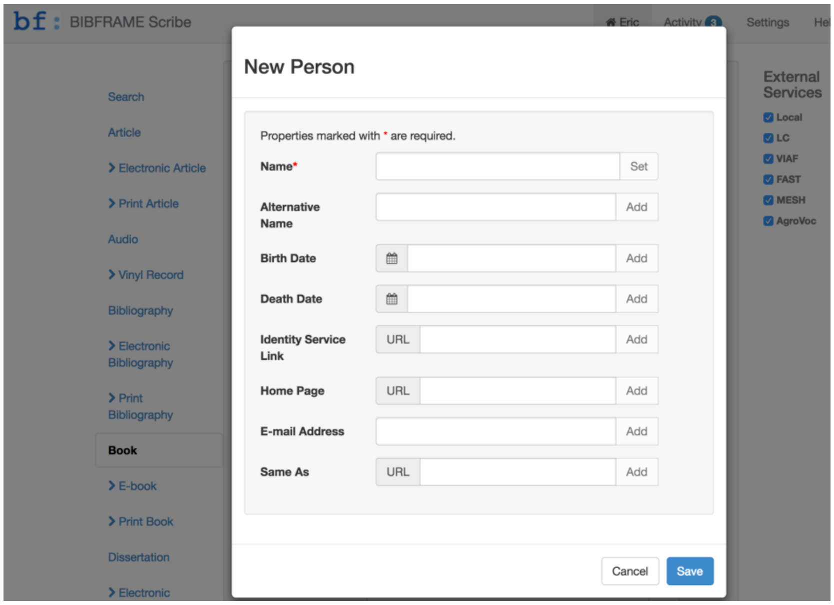
1. Identify need for new authority
2. Create new authority record
3. Submit new authority record to NACO
4. Return to original cataloging and continue cataloging item

Option 2

1. Identify need for new authority
2. Submit request for new authority
3. Wait for response to request
4. Return to original cataloging and continue cataloging item

Both of the above workflows involve the cataloger moving from the current cataloging work to another workflow (and often another computing system and interface) to create or request creation of a new authority before returning (either immediately or after an undefined period of time) to the cataloging workflow.

Linked Data workbenches, such as the BIBFRAME Scribe workbench tested as part of this project, eliminate the need to step away, as it were, from the current cataloging effort to deal with authority issues. When a cataloger is unable to locate a suitable authority, the workbench prompts the cataloger to create a new authority using whatever information is currently available to the cataloger:



New Person

Properties marked with * are required.

Name*

Alternative Name

Birth Date

Death Date

Identity Service Link URL

Home Page URL

E-mail Address

Same As URL

Figure 25: New authority in BIBFRAME Scribe

When a user creates a new authority entry, a graph for this authority is created in the local triplestore with a new, local URI. The cataloger is then returned to their ongoing cataloging effort.

When a cataloger creates a new authority using the above system, the authority is subsequently available within the local domain for all future cataloging efforts. This insures that all local cataloging efforts run efficiently, but does not, de facto, solve the larger problematics of authority control. As discussed in *Section II* above, Linked Data's ability to facilitate information traversal rests on the availability of URIs over the network and also on the assumption that each entity is uniquely represented. As such, a local instance of a URI cannot function as an authority unless it is distributed across the network and is done so in a way that can be properly linked to or differentiated from other URIs in the Linked Data universe.

Section VIII below provides a more in-depth discussion of processes for managing the production of local URIs for new authorities. Relevant to the present discussion is the fact that systems can be put in place to allow for on-the-fly authority graph creation, thereby streamlining the workflows of catalogers involved in original cataloging. These efficiencies include:

1. Catalogers do not need to have in-depth knowledge of BIBFRAME's data model or BIBFRAME vocabularies to perform cataloging because the terms used by Linked Data workbenches are the same ones currently used by catalogers

2. Catalogers do not have to leave the Editor in order to complete the cataloging work when confronted with authority issues
3. Catalogers have an option to create authority data on-the-fly and to mint local URIs which can be connected to other related URIs through a reconciliation service as discussed in *Section VIII*

In order to implement the above described workflow, the following systems need to be in place:

1. Robust lookup services which can interpret source data and present it in a readable format to catalogers
2. Systems for performing local authority reconciliation as described below in *Section VIII*

Transitioning to Linked Data cataloging using the proposed model raises the following questions for community consideration:

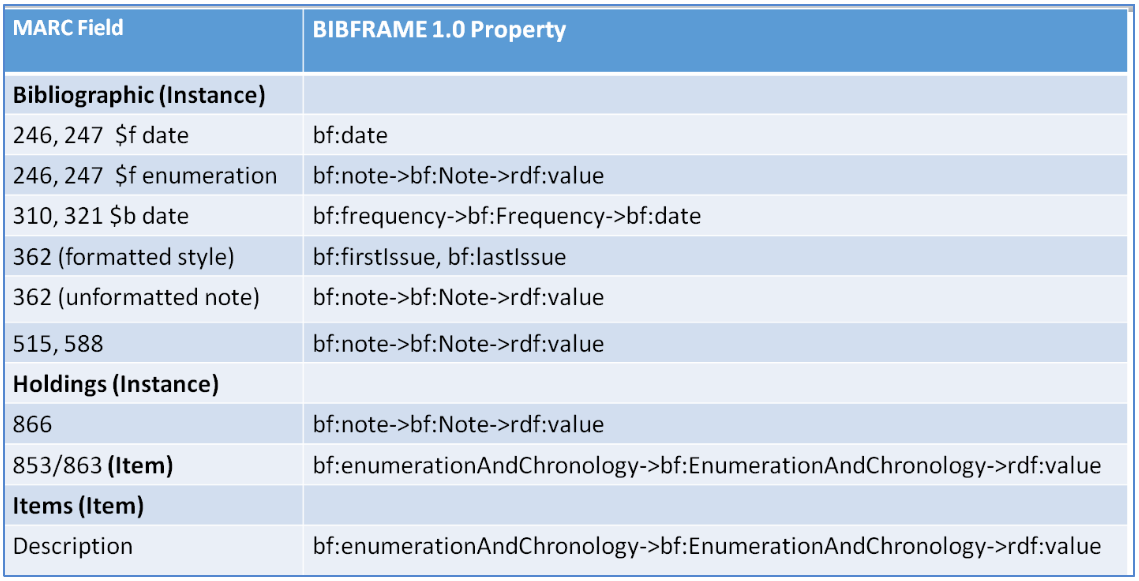
1. URIs are crucial in order to disambiguate or retrieve information in the Linked Data environment. As a result, the more sources a library can use, the less work needed locally. But how to find right balance? To what extent should we consider using non-traditional information sources such as commercial book sellers and Wikipedia?
2. Cataloging descriptive rules have played an important role in the card or MARC cataloging environment. In a world where most of the entities we describe can be identified by a unique ID (URI), how much descriptive data do catalogers still need to create if that information can be retrieve from other data sources, such as publishers or vendors?
3. Library of Congress subject strings played an essential role in the era when the discovery technology was string based and not always automated. With faceted navigation and other features a 21st century library discovery tool can offer, library users can narrow down their search results more easily. Given this new environment, how much value is added by having tightly controlled, nested subject strings presented to library users?
4. Instead of creating new name authority data, would it make sense for the library community to start using other authoritative URI enabled name identifiers, such as ORCID (researchers) and ISNI (individuals and organizations) IDs and focus on building context around these identifiers?

Serials Cataloging

Cataloging workflows described above can be used for cataloging serials. However, because of the changing nature of serial publications and the need to accommodate complex holdings information, cataloging serials in BIBFRAME has its own unique issues. During the life time of a serial publication, the serial title, issuing body, publication information, frequency, numbering, etc., may change. As a result, it is essential that catalogers are provided a means to associate dates or date ranges with assertions (triple statements). In this report, we want to highlight the following two areas where the current data BIBFRAME model will fail to maximize the potential of Linked Data:

1. The current state of BIBFRAME does not seem to be able to address adequately the issue of change over time to serials metadata. For example, there is not a way to express a start and end date for changes to titles and publication information. It may make sense for the serials cataloging community to explore other vocabularies that are more suitable for modeling serials, such as PRESSoo, for use in conjunction with BIBFRAME. [16]

2. Enumeration and chronology information is ubiquitous and important for describing serials. It is used with serial titles and appears in notes, item, and holdings records in the MARC environment. *Figure 26* shows the mappings of enumeration and chronology data in MARC records to corresponding BIBFRAME properties.



MARC Field	BIBFRAME 1.0 Property
Bibliographic (Instance)	
246, 247 $f date	bf:date
246, 247 $f enumeration	bf:note->bf:Note->rdf:value
310, 321 $b date	bf:frequency->bf:Frequency->bf:date
362 (formatted style)	bf:firstIssue, bf:lastIssue
362 (unformatted note)	bf:note->bf:Note->rdf:value
515, 588	bf:note->bf:Note->rdf:value
Holdings (Instance)	
866	bf:note->bf:Note->rdf:value
853/863 **(Item)**	bf:enumerationAndChronology->bf:EnumerationAndChronology->rdf:value
Items (Item)	
Description	bf:enumerationAndChronology->bf:EnumerationAndChronology->rdf:value

Figure 26: Enumeration and chronology information mapping

As illustrated above, there are two problems with how Enumeration and chronology information are expressed in BIBFRAME: 1) several different properties are often used to encapsulate a single datum point, resulting in an overly complex representation; and 2) none of those data are machine-actionable because they are

[16] Patrick Le Boeuf and François-Xavier Pelegrin. " FRBR and serials: the PRESSoo mode" http://library.ifla.org/838/1/086-leboeuf-en.pdf. See also http://www.slideserve.com/erasto/a-brief-presentation-of-press-oo.

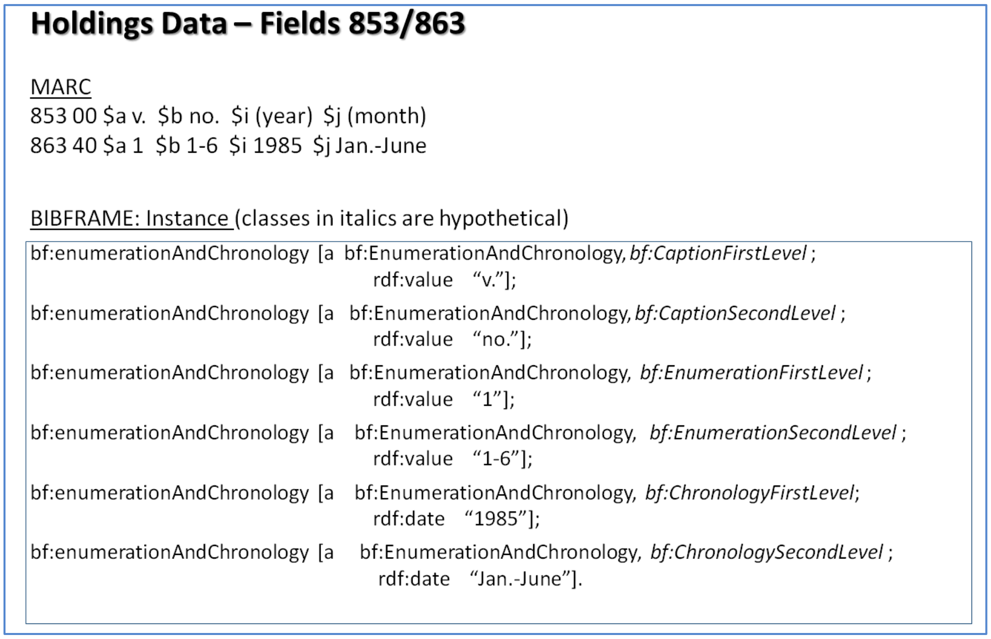
literals (strings of text). The serials cataloging community should consider the following questions:

a. Should enumeration/chronology data appearing at BIBFRAME *Instance* level be coded in a uniform way?

b. Should enumeration/chronology data appearing at both BIBFRAME *Instance* and *Item* be coded in the same way?

c. Does Linked Data offer the possibility of simplifying the ways in which we encode enumeration/chronology data while still achieving same end-user functionality for which they are intended? For example, dropping enumeration when chronology alone is sufficient.

d. Would it be more useful to parse enumeration and chronology data currently recorded in MARC 853/863 fields into similar pieces like this:



Holdings Data – Fields 853/863

MARC
853 00 $a v. $b no. $i (year) $j (month)
863 40 $a 1 $b 1-6 $i 1985 $j Jan.-June

BIBFRAME: Instance (classes in italics are hypothetical)

bf:enumerationAndChronology [a bf:EnumerationAndChronology, *bf:CaptionFirstLevel* ;
rdf:value "v."];
bf:enumerationAndChronology [a bf:EnumerationAndChronology, *bf:CaptionSecondLevel* ;
rdf:value "no."];
bf:enumerationAndChronology [a bf:EnumerationAndChronology, *bf:EnumerationFirstLevel* ;
rdf:value "1"];
bf:enumerationAndChronology [a bf:EnumerationAndChronology, *bf:EnumerationSecondLevel* ;
rdf:value "1-6"];
bf:enumerationAndChronology [a bf:EnumerationAndChronology, *bf:ChronologyFirstLevel*;
rdf:date "1985"];
bf:enumerationAndChronology [a bf:EnumerationAndChronology, *bf:ChronologySecondLevel* ;
rdf:date "Jan.-June"].

Figure 27: Possible model for holdings data

e. Should we explore other ontology/vocabularies such as *ONIX for Serials Coverage Statement (Version 1.0)* or *Enumeration and Chronology of Periodicals Ontology?*

f. Would incorporating other models or vocabularies enable the reusability of data? For example, harvesting existing enumeration and chronology data from content providers.

Several groups have been, and remain actively involved in discussions surrounding modeling serials using BIBFRAME and other vocabularies. These include groups from the LD4P, LD4L, Library of Congress BIBFRAME working group, and the PCC BIBFRAME CONSER working group. Future reports from these groups may shed more light on modeling serials. Given the efforts currently devoted to this area of Linked Data implementation, it is reasonable to expect that best practices will be achieved before libraries are situated to begin the transition. It is also worth noting that work on this front could continue with different libraries adopting different serials models. While this scenario is not preferred, a multi-model ecosystem could be made functional through reconciliation graphs that use multiple *sameAs* designations to linked disparate graphs. The following section provides an in depth discussion of reconciliation models.

Moving from MARC to Linked Data affords us the opportunity to take a fresh look at the way we describe serials. The answers to the challenges mentioned may be found by rethinking existing practices. Regardless of the path forward in serials cataloging, this is an area where we can expect the necessity for staff re-training.

VIII. Authority Control

Authority control is the area of Linked Data transition that has caused the most concern. According to *Maxwell's Guide to Authority Work,* "Authority work is so called because it deals with the formulation and recording of authorized heading forms in catalogue records," such that, "names and other headings that are access points to records are given one and only one conventional form."[17] Prior to the internet, when humans and non-networked computers were the only consumers of information, heading forms were string based, which is to say that the written, human readable form of a heading was the functioning authority. Humans and computing systems could only match records if the values of individual fields were identical as strings. Thus, for example, two records, each of which recorded an Author field with the value "Mark Twain" would be seen as connected through the Author field. But a collection records with Author field values "Mark Twain", "Twain, Mark", "Samuel Langhorne Clemens," and "Samuel L. Clemens" would not connect despite that fact that all of these name forms refer to the same, physical author.[18] This is a familiar concept to catalogers.

From one perspective, Linked Data authorities function much the same as MARC's human readable authorities. As with strings, when URIs are the same they stand as authority for the same named entity and for different entities when they are different. Thus, for example:

[17] Maxwell, Robert L. Maxwell's Guide to Authority Work. Chicago: American Library Association, 2002, p. 1.

[18] MARC does, of course, have mechanisms for dealing with variants through cross reference. The point here is to recognize that in a string based world even slight variations in spelling differentiate forms.

http://id.loc.gov/authorities/names/n79021164

matches

http://id.loc.gov/authorities/names/n79021164

but not

http://id.loc.gov/authorities/names/n79021165

As with authorities meant for human consumption, a variation of just one character (in the above case "4" to "5" in the last character of the string) results in treatment as a distinct authority.

When cataloging in MARC, the authorized, human readable version of a heading will always appear in the record access point regardless of how the name, subject, etc. may appear on the actual item, and cross referenced literal values may or may not be provided elsewhere in the record. In Linked Data cataloging, the same URI must be used to create a linkable node in the graph, but any given graph can contain any version of the human readable label (name, subject, etc.) without affecting the field's linking function.

Given the above, it is not necessarily the case that moving to Linked Data dramatically affects how we work with authorities. We could, in fact, use the same centralized authority control systems that we use today and the workflows that surround them. Linked Data, however, opens the possibility for radically new forms of authority.

Figure 28 below depicts the current, centralized model of authority control.

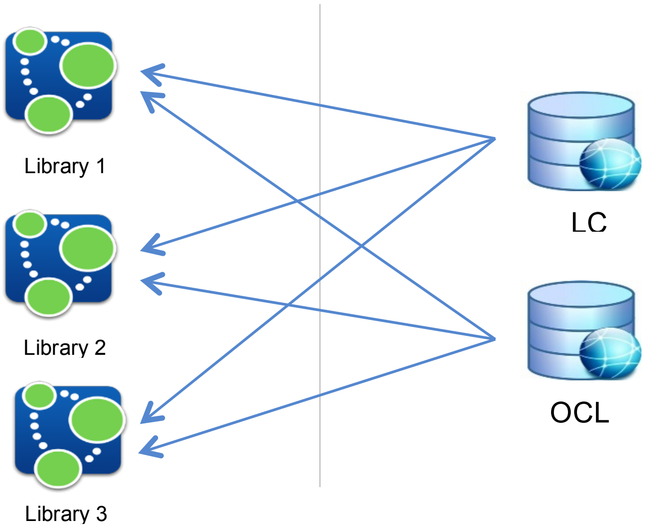


Figure 28: Centralized authority control

By contrast, *Figure 29* blow depicts a completely decentralized model for authority control:

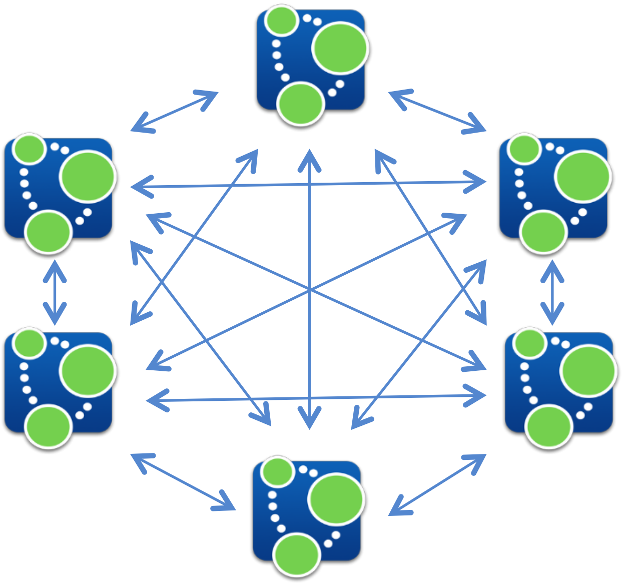


Figure 29: Decentralized authority control

It is the centralized authority control with which we are currently familiar. Authority headings are managed by one or more centralized authority. Individual libraries both request and submit headings from the appropriate managing authority. The decentralized model, by contrast, removes the authority managing organizations from the equation. Instead of going through central points of authority to manage authorities, libraries rely on each other.

In a completely decentralized authority model, rather than turning to a Library of Congress authority file, individual libraries would query each other's Linked Data points in search of authority URIs. For example, if cataloging a work credited on the title page as authored by "Mary W. Shelley," a cataloger would submit a query to other libraries for any triple in their graph store with the label "Mary W. Shelley," or "Mary Shelley", or even just "Shelley."   If a matching triple(s) were found, the cataloger would then pull the extended graph in the holding institution's data-store in order to disambiguate. Provided the cataloger determined from traversing this graph that it represented the same "Mary W. Shelley," the cataloger would use the found URI in the local graph. In cases where no graph can be located by querying other libraries for triples with the Label "Mary W. Shelley" the cataloger would mint a URI locally and insert it into the local graph for the work being cataloged, making the new URI findable and usable by other libraries through the Linked Data gateway to the cataloger's library.

The above system allows URIs to propagate organically through the extended library information network in a matter that is both efficient and provides a growing graph of context for disambiguation. Once a URI is in circulation, each library that uses the URI extends the graph of information available for other libraries to use in disambiguation. This extend graph would very quickly surpass the current level of context that surrounds existing authority methods.

There are, however, some potential difficulties with the completely decentralized model. Most obvious is the problem of finding an appropriate URI with a non-matching label.

The current, authorized heading for “Mary W. Shelley” is “Shelley, Mary Wollstonecraft, 1797-1851.” A query for the label “Mary W. Shelley” would not find a referenced URI for “Mary W. Shelley,” even though the two are actually the same person and should be represented with the same URI. The solution to this problem is a reconciliation process commonly known as *sameAs*.[19] The *sameAs* entity provides a mechanism for indicating that two URIs refer to the same entity. Thus, for example, if one graph assigns the following URI to Mary W. Shelley:

http://library1.com/entity/person/72312031

And another graph assigns the following, different URI to Shelley, Mary Wollstonecraft, 1797-1851:

http://library2.com/agent/person/q09eqe9mws

The following *sameAs* statement indicates that both URIs represent the same person, with the two name variants “Mary W. Shelley” and “Shelley, Mary Wollstonecraft, 1797-1851”:

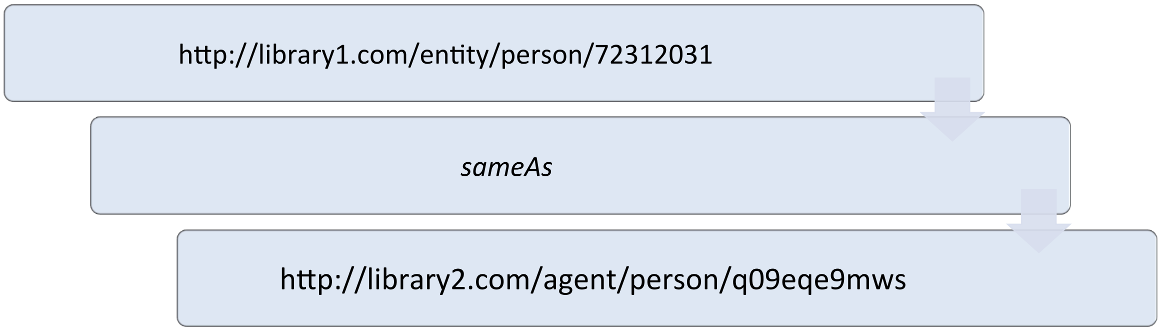

http://library1.com/entity/person/72312031
sameAs
http://library2.com/agent/person/q09eqe9mws

Figure 30: *sameAs* entity linking

Once a *sameAs* statement has been made and published, it becomes available for others to take advantage of. A traversal for the “Mary W. Shelley” URI would find the *sameAs* statement and know that it also need to query for the “Shelley, Mary Wollstonecraft, 1797-1851” in order to produce a complete graph of the referenced person—provided the querying institution has access to the graph that contains the *sameAs* statement.

There are two primary obstacles to a completely decentralized authority model of URI creation and *sameAs* reconciliation. The first is the problem of determining the scope of query traversal. Were the entire library community to transition to Linked Data, the number of graph endpoints would be staggering. This number would continue to grow as commercial vendors and services enter the ecosystem. As such, traversing the entire knowledge graph represented by the Linked Data web is not computationally practical. Making such a traversal would require computing resources on the order of that currently provided by major search engines—a level

[19] The *sameAs* moniker is derived the *owl:sameAs* entity, an entity of the structural specification of the Web Ontology Language (OWL), a W3C specification for defining ontologies.

of technology support not now nor likely ever to be in the grasp of even the most major resource libraries.

History provides a lesson in the above regard. In the early days of the internet it was common for people and institutions to perform their own crawls of the entire internet and store a local cache for searching.[20] However, within a year of the advent of the World Wide Web, such traversals became impractical based on both time of crawl and space required to store crawl caches, and the search engine as service was born. Farming out crawling and caching functions to a handful of centralized systems solved the computing barriers of local crawling and caching.

As the number of cultural heritage institutions and supporting commercial interests increases, libraries will quickly face the same technological barrier that confronted information consumers of the early internet. As such, some form of centralized authority operations will be a technological necessity for the future Linked Data library ecosystem. There are, however, multiple forms that such an operation could take.

Several organizations that currently maintain widely used authority lists have already made their MARC-based authorities available as Linked Data. This includes organizations such as the Library of Congress, OCLC, and Getty.[21] As more libraries move into the Linked Data ecosystem, we can reasonably expect that others will do the same. None of those organizations currently making the authorities available as Linked Data have changed the process through which they manage their authorities to reflect a Linked Data environment. The Library of Congress, for example, still employs the same NACO system of authority management. Their Linked Data gateways are simply a Linked Data representation of the Library of Congress authority files.

Similar to the Library of Congress, OCLC has made its WorldCat, FAST, and VIAF data available as Linked Data. As with the Library of Congress, the bulk of these services represent a re-presentation of traditional MARC-based data, with no significant modification of resource management practice. OCLC has, however, recently been engaged in a variety of pilot projects aimed at capitalizing on the potential of Linked Data to facilitate authority management.

Several of the OCLC Linked Data pilots have focused on solving the *sameAs* problem discussed earlier. The first iteration of the pilot service provided what can best be described as an authority registry, as system for centralized *sameAs* aggregation of authority URIs created by various institutions, including local libraries—a process that has come to be known as URI reconciliation. *Figure 31* presents an overview of the basic methodology:

[20] For several years in the early 1990s, one of the authors of this report maintained his own, searchable local cache of the entire internet on a Sparc server in his office.

[21] See http://id.loc.gov/, https://www.oclc.org/developer/develop/linked-data.en.html, and http://www.getty.edu/research/tools/vocabularies/lod/ respectively. Also see the “Vendor Engagement” section for a more complete view of Linked Data offerings.

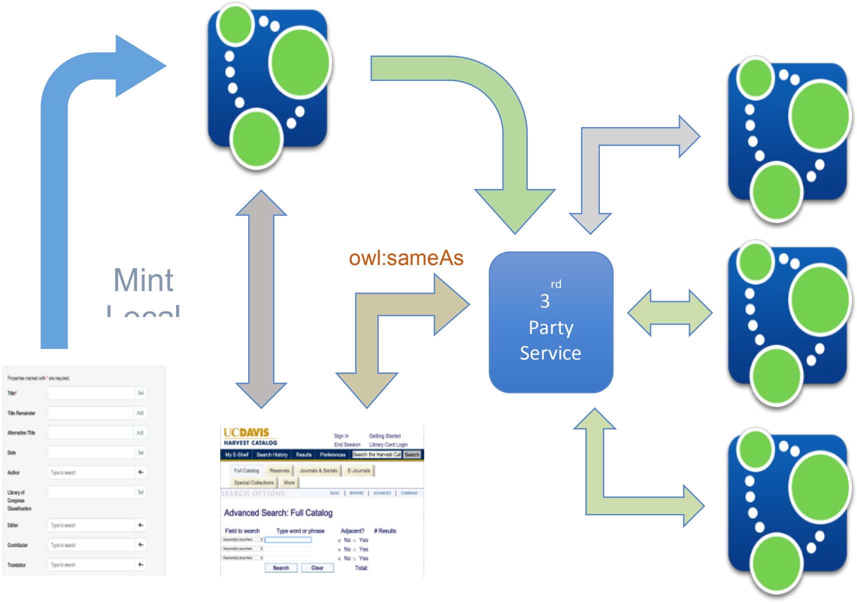


Figure 31: Centralized authority reconciliation model

The above model allows individual libraries to submit locally created URIs to a 3rd party service for reconciliation. Needed local URIs would be created and submitted to the reconciliation service where it would be aggregated through a *sameAs* relationship with other URIs that refer to the same entity. During search and discovery (whether by end-users or internally as part of the cataloging workflow) the aggregated set of URIs provided by the reconciliation service are used to build the graph to be presented to the user.

The type of service described above (for which OCLC is currently planning a pilot) dramatically streamlines the process of building associative graphs. For example, consider a situation where three URIs have been minted for the same entity. In order to build an associative graph, an institution would have to query the Linked Data ecosystem first for the known label of the entity for which they are searching. This would return one URI. They would then have to re-query the universe for all instances of the returned URI looking for *sameAs* statements. And for each returned *sameAs* statement, they would again have to query the entire ecosystem for other *sameAs* relationships that contain references to URIs not already known. Building the complete list of sameAs relationships for an entity with three sameAs URIs in circulation would require a minimum of three and a maximum of five traversals of the ecosystem. A centralized reconciliation system similar to the one depicted in *Figure 31* would reduce this to a constant two traversals—a significant improvement in efficiency that would result in significant savings in both speed of query and cost.

The above discussion is meant only as an introduction to the problem of authority control. Its intent is to provide a foundation for understanding what is involved in considering Linked Data authority; and, more importantly to demonstrate that there are viable solutions to this perceived barrier to adoption to Linked Data. As demonstrated by the current Linked Data offerings of major authority providers, it is possible to provide reliable Linked Data authority without changing anything about the way authority management is currently conducted. As such, the perceived Linked Data authority problem is, in fact, a Linked Data opportunity—a chance to improve operational efficiency and the depth of contextual information that surrounds authority headings.

Cornell University is currently mid-cycle of an IMLS grant devoted specifically to understanding and modeling processes for Linked Data authority focused on seizing the opportunity of Linked Data transformation to improve both the quality of authority data and the efficiency of the workflows that create and manage it. This effort is already producing valuable results, and promises to conclude with a collection of community developed principals and models for Linked Data authority control. Those with an interest in this area of Linked Data implementation should follow the work of this project.

IX. Vendor Engagement

This section characterizes the state of Linked Data readiness and awareness within the community of library services and product providers. Innovations in digital media applications on the Web from companies like Google and Amazon are clear wake-up calls to libraries and their service providers which, in response, need to expand their strategy to work in new and different ways. There are primarily two possible reactions to this major technological change: try to delay or deny the development, or seize the opportunity and use it to redefine the relationships between libraries and their communities of users.

Librarians and their service providers must work together to ensure that libraries are well positioned to take advantage of evolving technology and offer their rich resources to users in their communities and across the globe. In order to do so, library systems must become compatible with a range of external and internal systems including acquisitions, cataloging, circulation, discovery layers, and content management systems. An overview of the findings reported in this section can be found in *Appendix A*.

Methodology

In pursuit of the objective to provide an assessment of Linked Data strategies in the library industry, a multi-method approach was employed. The information synthesized by this report was gathered through direct conversations with service providers and combined with material made publicly available to document the business and product development strategies of companies that provide library services. First, Zepheira worked with academic and public librarians to identify key library service providers. It then reached out to these companies in order to assess the following areas:

1. Have these service providers experienced demand for Linked Data integration or any Linked Data services yet?
2. Have the companies established any collaborative partnerships with customers or other companies for Linked Data developments, grant-funded or otherwise?
3. Have they published any reports, white papers or other public documents relevant to Linked Data initiatives?

For those companies who are not yet incorporating Linked Data into services, Zepheira then began educational discussions to explain the increasing interest in Linked Data their customers are experiencing in the library community. Some companies did not respond to requests for information, or were not willing to share information at the time of this report because their business plans are confidential for competitive reasons. In such cases, research was performed to gather public documentation on Linked Data products, services, and strategies. There may be service providers working on Linked Data products that are not addressed by this report. Due to the fast-paced and constantly changing nature of Linked Data adoption, this report is not intended to be comprehensive and does not provide recommendations to libraries for purchasing specific services.

Summary of Linked Data Assessment

With the exception of a few forward thinking companies including Atlas Systems, EBSCO, Ex Libris, Innovative Interfaces, Inc. (III), OCLC, Overdrive, ProQuest, SirsiDynix, and Zepheira, library vendors in general are either unaware or minimally aware of Linked Data developments and benefits.[22] Libraries, archives, and museums are starting to working together with their service providers to solve these challenges to move forward towards a future with visible resources on the Web that can be used by a variety of Semantic Web applications. The following summary of Linked Data assessment is divided up by service provider and arranged in alphabetical order. Details follow for each service provider on their plans for incorporating Linked Data to the extent they were willing to share publicly.

Atlas Systems

Atlas Systems is the provider of Aeon, circulation and workflow automation software for archives and special collections, and ILLIAD, resource sharing management software for automating interlibrary loans. Atlas Systems believes Linked Data and BIBFRAME will play a key role in how the Web understands libraries and reflects what libraries have to offer. Atlas Systems became a founding Libhub Initiative sponsor in spring 2015.[23] The Libhub Initiative is an effort founded by Zepheira to bring people together to explore and experiment with Linked Data technologies in service of increasing library relevance through the Web.[24] Atlas' support, along with support from many other service providers, funded a forum and experimental space for librarians, libraries, and industry leaders.

In the fall of 2015, Atlas Systems partnered with Zepheira to start exploring Linked Data

[22] See https://www.atlas-sys.com, https://www.ebscohost.com, https://.www.iii.com ,or https://www.sirsidynix.com, https://www.ebscohost.com/novelist , http://www.oclc.org , https://www.overdrive.com , http://www.proquest.com , http://library.link, and https://zepheira.com/ respectively.

[23] Atlas Systems One of First Sponsors of the Libhub Initiative." Atlas Systems: Library Excellence Through Efficiency. March 18, 2015. Accessed August 1, 2015. http://www.atlas-sys.com/atlas-systems-one-of-first-sponsors-of-the-libhub-initiative/.

[24] "Libhub: University Leading, Learning, Linking." Accessed October 15, 2015. http://www.libhub.org/. For a full list of Libhub Initiative Partners and Sponsors, see: http://www.libhub.org/sponsors-partners.

for Archives and Special Collections. At the end of 2015, Atlas Systems became a Registered Service Provider for ArchivesSpace, an open source content management and publishing platform for archives and special collections. At ALA Mid-Winter, Atlas Systems presented findings of their Linked Data research to the Association for Library Collections & Technical Services MARC Formats Transition Group in Boston. Currently, Atlas Systems is continuing to explore methods for integrating Linked Data with ArchivesSpace, Aeon, and ILLIAD.

EBSCO and Novelist

In February 2015, EBSCO announced that they will be funding development of Koha, an open source Integrated Library System created by librarians for librarians. Koha Linked Data updates will include MARC to RDF cross-walking to enhance capabilities of linking to online data repositories.[25] However, in April 2016 EBSCO announced that they would no longer be supporting Koha or Kuali OLE development, and will instead fund the development of an open-source Library Service Platform.[26] The outlined functional expectations for this new open-source Library Service Platform include support for Linked Data Services. An initial version of the software will be available in early 2018.

EBSCO and their subsidiary company, NoveList, became Libhub Initiative sponsors in October 2015. EBSCO explained that they are "showing support for Zepheira and moving forward to support BIBFRAME and Linked Data which are seen as essential to opening up library collections to the World Wide Web."[27] NoveList launched the Linked Library Service in April 2016 at the Public Library Association in Denver. The service, created for public libraries, publishes Linked Data to the Web via the Library.Link Network. The Library.Link Network provides global infrastructure for publishing library Linked Data. NoveList is currently researching enrichment products and services for Linked Data.[28]

Innovative Interfaces, Inc. (III)

In 2014, Innovative Interfaces, Inc. (III) demonstrated strong interest in Linked Data innovation and support for the library industry's BIBFRAME transition by becoming a founding sponsor of the Libhub Initiative. In March 2016, after reviewing the results of Libhub Initiative experimentation done by III customers, the company partnered with Zepheira to release a new

[25] "Koha Receives Massive Support from EBSCO for Enhancements to Its Web-Based, Open-Source ILS." EBSCO. February 11, 2015. Accessed August 1, 2015. https://www.ebsco.com/news-center/press-releases/koha-receives-massive-support-from-ebsco-for-enhancements.

[26] Breeding, Marshall. "EBSCO Supports New Open Source Project Software for academic libraries will be developed collaboratively." April 22, 2016. Accessed April 22, 2016. http://americanlibrariesmagazine.org/2016/04/22/ebsco-kuali-open-source-project.

[27] McEvoy, Kathleen. "EBSCO Information Services and NoveList Show Commitment to BIBFRAME and Linked Data through Sponsorship Agreement with Zepheira." PR Web. October 5, 2015. Accessed October 10, 2015. http://www.prweb.com/releases/Zepheira/NoveList/prweb13004978.htm.

[28] "Library Visibility on the Web: Linked Library Service." Accessed April 20, 2016. https://www.ebscohost.com/novelist/our-products/linked- library-service.

service, Innovative Linked Data. The goal of the Innovative Linked Data service is to extract bibliographic data from Polaris, Sierra, Millennium and Virtua library systems, transform the information, and publish the descriptions as Linked Data on the Web via the Library.Link Network. The Innovative Linked Data pages can be found on the open Web, including discovery via search engines. The pages direct users to the library's interface where they can complete their interaction with the library. Leif Pedersen, Executive Vice President at Innovative, explains "the Innovative Linked Data service publishes regular updates of library data to the Web, and this constant exposure to search engines will help drive our library partners' visibility among search results. Innovative Linked Data plays a critical role in the relevance and sustainable discovery of libraries, and catalog content and geographic locations are just the first step in our commitment to strengthen and expand the library user experience."[29]

III realized the importance of Linked Data before incorporating the technology into their tools and services.[30] In early April 2015, III announced that "Linked Data is going to fundamentally change some of the assumptions which we have operated upon."[31] III continues to partner with Zepheira to streamline the transformation of their customers' MARC records into Linked Data. At the 2016 Innovative User Group Meeting in San Francisco and the 2016 Public Library Association Annual Meeting in Denver, III launched Innovative Linked Data and made subscriptions to the service available.

Online Computer Library Center, Inc. (OCLC)

OCLC is broadly known for their support of Linked Data and actively speaks about integration of Linked Data into their strategy. Publication of the Virtual International Authority File (VIAF) and Faceted Application of Subject Terminology (FAST) as Linked Data were early demonstrations of OCLC's strategic initiatives to provide authoritative library data in open formats native to the Web.[32] Connexion, OCLC's tool for creating, acquiring, and managing

[29] "Innovative Advocates for Library Visibility on Semantic Web with Launch of Innovative Linked Data." March 16, 2016. Accessed April 13, 2016. https://www.iii.com/news-events/pr/innovative-advocates-library-visibility-semantic-web-launch-innovative-linked-data.

[30] "Learning about Linked Data." August 12, 2015. Accessed January 5, 2016. https://www.iii.com/community/inn-side-view/learning-about-linked-data.

[31] "What Is Innovative Thinking about Linked Data and Its Impact on Libraries?" October 31, 2015. Accessed August 1, 2015. https://www.iii.com/ community/inn-side-view/what-innovative-thinking-about-linked-data-and-its-impact-libraries/. For more information about the Virtual International Authority File, see: https://viaf.org/. For more information about Faceted Application of Subject Terminology, see: http://experimental.worldcat.org/fast/. For more information on how to explore WorldCat Linked Data, see: https://www.oclc.org/developer/develop/linked-data/linked-data-exploration.en.html.

[32] For more information about the Virtual International Authority File, see: https://viaf.org/. For more information about Faceted Application of Subject Terminology, see: http://experimental.worldcat.org/fast/. For more information on how to explore WorldCat Linked Data, see: https://www.oclc.org/developer/develop/linked-data/linked-data-exploration.en.html.

bibliographic and authority records does not include Linked Data services. However, OCLC provides access to over 197 million bibliographic work descriptions in Linked Data format via WorldCat Works. These Linked Data entities are incorporated into WorldCat and made available to software applications via API. To support a more human-friendly understanding of these data, work entities are also available via the WorldCat Linked Data Explorer Interface. Libraries can use OCLC's work entities to consistently identify works in a way the Web understands. Through creating actionable URIs to identify works, OCLC is providing the infrastructure that will be needed to identify works in future Linked Data based systems and services.

In January 2015, OCLC published a white paper with the Library of Congress entitled "Common Ground: Exploring Compatibilities Between the Linked Data Models of the Library of Congress and OCLC." A major outcome of the paper is the recommendation that OCLC develop and test technical solutions that capture information expressed in BIBFRAME that cannot be expressed using the schema.org model. The report also recommends the development of services that can export and import BIBFRAME into OCLC systems without data loss.[33]

In February 2015, OCLC featured BIBFLOW as part of the Collective Insight Series titled, "Linked Data [R]evolution: Applying Linked Data Concepts." The goal of this session was to explain OCLC's work with Linked Data and provide presentations from people experimenting with Linked Data in libraries, including "Linked Data in the Library Workflow Ecosystem" presented by Carl Stahmer, Director of Digital Scholarship at the UC Davis University Library.[34]

A primary goal for OCLC's work with Linked Data is to understand the library workflows that will drive the use of tools that use Linked Data. To support this strategy, OCLC is working with the Library of Congress, the BIBFRAME community, and the schema.org community.[35] OCLC Research is also experimenting with the beta version of a discovery layer for Linked Data, called Entity JS, to demonstrate other uses for WorldCat Entities. In September 2015, OCLC announced a person entity lookup pilot project. The pilot aims to help library professionals reduce redundant data about people by linking related sets of identifiers and authorities. The libraries participating in the pilot include University of California, Davis, Cornell University, Harvard University, the Library of Congress, the National Library of Medicine, the National Library of Poland, and Stanford University. Together OCLC and these libraries will improve the relationships between authorities and the librarian's ability to identify the vast number of people who create and are described by library collections.[36]

[33] Godby, Carol Jean, and Ray Denenberg. "Common Ground: Exploring Compatibilities Between the Linked Data Models of the Library of Congress and OCLC." January 2015. Accessed August 1, 2015.

[34] Price, Gary. "Videos and Slides: Five Presentations From OCLC's "Library Data [R]evolution: Applying Linked Data Concepts." Event Now Available Online." InfoDOCKET. February 26, 2015. Accessed August 1, 2015.

[35] OCLC. "Making Breakthroughs Together." December 2015. Accessed May 1, 2016. https://www.oclc.org/en-US/annual-report/2015/ breakthroughs.html.

[36] OCLC to launch Linked Data pilot with seven leading libraries." September 11, 2015. https://www.oclc.org/news/releases/ 2015/201526dublin.en.html.

OverDrive

To date, OverDrive has limited their public use of Linked Data to incorporating a limited amount of schema.org decoration into their interfaces in order to make high-level information available to Bing and Google. OverDrive continues to monitor Linked Data adoption in the library industry. The company is evaluating how Linked Data can be incorporated into their strategies for eBook, video, and audiobook access for public libraries. OverDrive is also engaged with Libhub Initiative partners and participants. Currently, OverDrive is working with customers to assess the potential utility of the Library.Link Network and possible integration with OverDrive content.

ProQuest and Ex Libris

In October 2015, ProQuest agreed to acquire the Ex Libris Group in order to "support ProQuest's mission to innovate across libraries across the world."[37] In December 2015, Ex Libris announced their vision and roadmap for incorporating Linked Data into two products: Alma, their resource management service and Primo, their discovery layer solution, will enhance workflows and allow new methods for exploring library resources. In addition, Ex Libris plans to make the Linked Data provided by each product available to third party tools.

In the outline of their plan for Linked Data services, Ex Libris explained, "While there is a shared understanding that the use of Linked Data will have many benefits in the form of new services for both library staff and end users, the precise nature of the possibilities is still a matter of discussion and debate. Ex Libris is working closely with libraries around the world to identify the scenarios and use cases that are expected to yield the greatest value to libraries and patrons, and is actively leading the way in planning and implementing linked-data services as part of the Alma resource management and Primo discovery and delivery solutions."[38] Ex Libris plans to incorporate BIBFRAME into their Linked Data services, which will include BIBFRAME import and export from Alma. The Alma Linked Data pilot has already produced demonstration functionalities to this end.

SirsiDynix

SirsiDynix was the first company to offer a Library.Link Network service for integrated library systems in partnership with Zepheira. In Fall 2015, SirsiDynix launched BLUECloud Visibility in order to transform their customers' MARC records into Linked Data and make library resources visible on the Web. To make library Linked Data freely available to search engines and other applications, the service allows any library using Symphony or Horizon along

[37] "ProQuest and Ex Libris Join to Accelerate Innovation for Libraries Worldwide." October 6, 2015. Accessed April 15, 2016. http:// www.proquest.com/about/news/2015/ProQuest-and-Ex-Libris-Join-to-Accelerate-Innovation-for-Libraries-Worldwide.html

[38] "Putting Linked Data at the Service of Libraries: the Ex Libris Vision and Roadmap." December 2015. Accessed April 14, 2016. http://www.exlibrisgroup.com/files/Publications/LinkedDataattheServiceofLibraries.pdf

with the BLUECloud web application to have their catalog data harvested, transformed to Linked Data, and published to the Library.Link Network.[39]

In an announcement about their partnership with Zepheira, Bill Davison, the CEO of SirsiDynix said "Our goal is to take the mystery and complexity out of Linked Data and deliver to our customers a product that is plug-and-play. We want libraries to easily transform their MARC data into robust, Web-searchable, geo-locatable Linked Data—ready for the world to find."[40] Many libraries are starting to set-up the infrastructure needed to publish a Local Graph of Linked Data. So far, 26 library organizations have published with SirsiDynix's BLUECloud Visibility. These organizations include large consortia like the Houston Area Library Automated Network, the Library Integrated Network Consortium, and the System Wide Automated Network Consortium as well as individual libraries like Randwick City Library and special libraries like the International Bureau of Fiscal Documentation in the Netherlands.

Zepheira

In Fall 2014, Zepheira founded the Libhub Initiative, an effort to bring libraries together with data and service providers to explore and experiment with Linked Data technologies in service of increasing library relevance through the Web. The Libhub Initiative sparked more than 500 conversations, meetings, interviews and experiments with library professionals as well as library data and service providers, all committed to greater library relevance through better library visibility on the Web. With many successful partnerships with libraries across North America and early support from Atlas Systems, III, SirsiDynix and EBSCO/NoveList, Zepheira felt there was strong confirmation from libraries and vendors alike and saw a clear need for global Linked Data infrastructure. Currently, Zepheira's top priority is offering its Linked Data infrastructure service, the Library.Link Network, to libraries, cultural history organizations, and their service providers who wish to improve the visibility of libraries and their collections on the Web. Partnering with library service providers to create Linked Data services lowers the barriers of entry for libraries that may not be able to participate in experimental projects.

Launched in 2016, the Library.Link Network is a global infrastructure for allowing libraries and other cultural heritage organizations to increase their visibility on the Web while maintaining the uniqueness of their own local identity. The Library.Link Network is the direct result of successful library-led collaborations for large-scale Linked Data experimentation completed under the umbrella of the Libhub Initiative.[41] The Library.Link Network brings together libraries and their providers on the Web to share their localized, comprehensive, connection-rich stories. While Zepheira established the Libhub Initiative as a community space

[39] "BLUEcloud Visibility." Accessed August 1, 2015. http://www.sirsidynix.com/products/bluecloud-visibility.

[40] "SirsiDynix and Zepheira Announce New Details of Strategic Partnership." June 25, 2015. Accessed August 1, 2015. http://www.sirsidynix.com/ press/sirsidynix-and-zepheira-announce-new-details-of-strategic-partnership.

[41] For more information on the Libhub Initiative, see: libhub.org. See also "Ending the Invisible Library" Enis, Matt. February 24, 2015. Accessed April 14, 2016. http://lj.libraryjournal.com/2015/02/technology/ending-the-invisible-library-linked-data.

for libraries to share best practices around implementing Linked Data, the Library.Link Network provides shared infrastructure that libraries can use to make their resources and events visible on the Web by publishing their resources in Linked Data format.[42]

The Library.Link Network infrastructure is used to reveal library resources including events, collections, bibliographic data and archival description in a Web-actionable format. Zepheira works with publishing partners and libraries to transform data from MARC and other formats into Linked Data to seed the Web with structured, openly published and interLinked Data.

Library.Link Network partners include Atlas Systems, Counting Opinions, Innovative, SirsiDynix, and most recently, EBSCO's NoveList. Contributions to and participation in Library.Link Network are possible at different levels. Some services are free to libraries, including the description of library locations and hours of operation with Linked Data. Other Library.Link Network services and partner services are fee-based, including Linked Data transformation for entire catalogs and publication of Linked Data to the Web via the Library.Link Network. Zepheira, SirsiDynix, Innovative Interfaces, and Novelist all offer library services that publish Linked Data to the Library.Link Network. All libraries are free to contribute their identifying information to the Library.Link Network in order to make the organization more visible on the Web. Participating in Library.Link Network gives libraries and archives the opportunity to contribute collection details into an open data store, known as a Local Graph. The Library.Link Network also connects the shared resources across Local Graphs to create trustworthy Linked Data on the open Web.

Over 1,110 public library locations have published Linked Data via the Library.Link Network, including Denver Public Library, Arapahoe Public Library, Dallas Public Library, Worthington Public Library, and Tulsa Public Library.[43] In total, 29,378,381 MARC records have been transformed resulting in 118,799,193 Linked Data resources and 326,009,018 links connecting the data. Academic libraries are also beginning to join Library.Link Network. Most recently, Boston University transformed their MARC catalogs into Linked Data and published via the Library.Link Network. Jack Ammerman, Associate University Librarian for Digital Initiatives and Open Access, explains "We are committed to making the resources of Boston University Libraries discoverable in the preferred discovery environments of our users. Publishing our

[42] Scardilli, Brandi. "Zepheira Helps Libraries Tell Their Stories on the Web." May 3, 2016. Accessed May 4, 2016. http:// newsbreaks.infotoday.com/NewsBreaks/Zepheira-Helps-Libraries-Tell-Their-Stories-on-the-Web-110777.asp.

[43] Fewell, Rachel. "DPL Announces Linked Data Launch." June 9, 2015. https://www.denverlibrary.org/blog/rachel-f/dpl-announces-linked-data-launch. See also "Library Working to Establish Greater Visibility on the Web" May 28, 2015. Accessed April 15, 2016. http://arapahoelibraries.org/visible-library 39 "Where We're Going We Don't Need Catalogs?" August 5, 2015. Accessed April 15, 2016. http://dallaslibrary2.org/blogs/bookedSolid/2015/08/where-were-going-we-dont-need-catalogs, "Worthington Libraries Announced Launch of Linked Data." August 12, 2015. Access April 15, 2016. http://www.worthingtonlibraries.org/about/ news/2015-8/worthington-libraries-announces-launch-linked-data, and "Free the Books." Accessed April 15, 2016. http://m.tulsalibrary.org/freethebooks.

records as Linked Data is an essential first step for us. We are convinced that publishing these bibliographic data in Linked Data formats not only increases their discoverability, but enables their re-use in ways we can't yet imagine."[44] The University of Manitoba was the first academic Canadian library to use the Library.Link Network.[45] Les Moor, Head of Technical Services at University of Manitoba Libraries, is working to improve how users find resources. Les says, "Linked Data allows our faculty, students and researchers to use popular search engines like Google to find our resources. As a result, we take a big step towards closing a giant discovery gap."[46]

X. Discovery

Discovery is the aspect of Linked Data implementation that is least studied, tested, and understood, as well as being the aspect most likely to have the biggest impact on library operations. It offers the potential for radical changes in the way users search and browse for library information. One of the most obvious and talked about aspects of Linked Data adoption is the extent to which it positions library information to be accessible through search engines and connected with a graph of information beyond the library.

In his February 2015 presentation "Making Library Collections Discoverable on the Web" as part of the OCLC Collective Insight Series titled, "Linked Data [R]evolution: Applying Linked Data Concepts," Ted Fons outlines the way Linked Data enabled library records can and should circulate in the wider information universe of the World Wide Web.[47] According to Fons, as libraries increasingly shift focus from physical collections management to information access management, the need to make information discoverable through user-preferred mechanisms increases. This requires structuring information such that it can be located through non-library interfaces such as major search engines.[48]

Search engine optimization is a valuable benefit to Linked Data adoption; but it is not the only, or even most important discovery advance that it serves. Linked Data makes possible new visual discovery interfaces that speak to one of the most voiced laments about the

[44] Daly, Janet. "Boston University libraries choose Zepheira to convert catalog to Linked Data opening it to the Web. February 23, 2016. Accessed April 15, 2016. http://zepheira.com/news/boston-university-libraries-choose-zepheira-to-convert-catalog-to-linked-data-opening-it-to-the-web/.

[45] For a full list of Library.Link Network users, see: http://library.link/statistics.html.

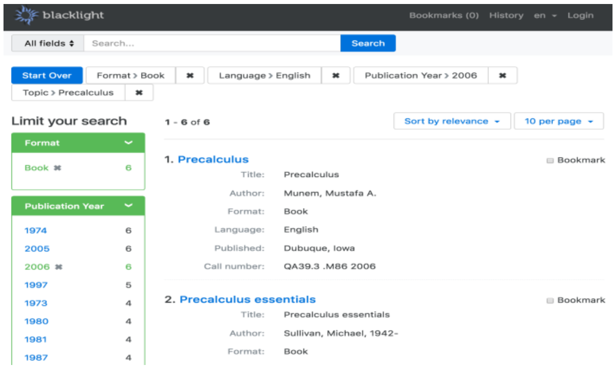
[46] Daly, Janet. "University of Manitoba Chooses Zepheira to Convert Catalogue Collections to Linked Data." February 16, 2016. April 15, 2016. http://zepheira.com/news/university-of-manitoba-chooses-zepheira-to-convert-catalogue-collections-to-linked-data/.

[47] See the complete video at https://www.youtube.com/watch?v=kcJ2uNvZY5s&list=PLWXaAShGazu5MybuhCd7Kf_4jh0moRXj1&index=2.

[48] The desire expose library information to commercial search engines was behind OCLC's decision to adopt Schema.org, the preferred Linked Data ontology of the search engine community rather than BIBFRAME. It, however, worth noting that entity based URI traversal will work regardless of the encoding Linked Data framework.

transition from stacks to screens: the serendipity of browsing. Consider the rudimentary network visualization of *the Lord of the Rings* presented in *Figure 1* at the beginning of this report. Here we see a focus on a text of interest spread to include an expansive multitude of context, much of which will inevitably be unknown to the user. This type of interface allows users to follow threads of relationship in a manner that harkens back to the days browsing the stacks, moving from one node to the next, with the option of focusing in on a new node and following the subsequent traces growing from it.

The above is just one example of a potential new discovery mechanisms made possible through Linked Data adoption. Current Linked Data discovery efforts either present experimental, pilot demonstrations of this potential (such as the thin network graph presented in *Figure 1*) or provide traditional search and discovery interfaces. One of the most adopted platforms for exposing Linked Data graphs for search and discovery is Blacklight.



1 - 6 of 6

Format	
Book	6

Publication Year	
1974	6
2005	6
2006	6
1997	5
1973	4
1980	4
1981	4
1987	4

1. Precalculus
Title: Precalculus
Author: Munem, Mustafa A.
Format: Book
Language: English
Published: Dubuque, Iowa
Call number: QA39.3 .M86 2006

2. Precalculus essentials
Title: Precalculus essentials
Author: Sullivan, Michael, 1942-
Format: Book

Figure 32: Blacklight demonstration screenshot

Figure 32 above is a screenshot of the online demonstration of Blacklight.[49] Blacklight is, "A multi-institutional open-source collaboration building a better discovery platform framework."[50] It provides a traditional but sophisticated search and discovery interface to both MARC and Linked Data data-stores. Built on an Apache SOLR/Lucene index, it provides fuzzy search, with full text search capability and faceted browsing.

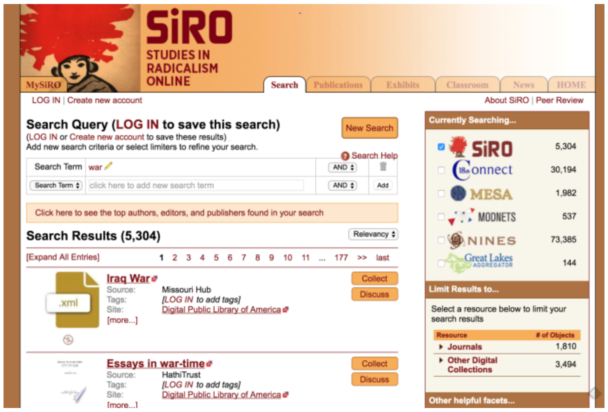
Another Linked Data discovery platform is Collex, a native Linked Data platform maintained by the Advanced Research Council and Institute for Digital Humanities, Media, and Culture at the IDHMC.[51]

[Continued on Net Page]

[49] See http://demo.projectblacklight.org.

[50] See http://projectblacklight.org.

[51] See http://idhmcmain.tamu.edu/arcgrant/ and http://idhmcmain.tamu.edu/ respectively.



Search Query (LOG IN to save this search)
(LOG IN or Create new account to save these results)
Add new search criteria or select limiters to refine your search.

Search Term: war

Search Term: click here to add new search term

Click here to see the top authors, editors, and publishers found in your search

Search Results (5,304)

[Expand All Entries] 1 2 3 4 5 6 7 8 9 10 11 ... 177 >> last

Iraq War
Source: Missouri Hub
Tags: *[LOG IN to add tags]*
Site: Digital Public Library of America
[more...]

Essays in war-time
Source: HathiTrust
Tags: *[LOG IN to add tags]*
Site: Digital Public Library of America
[more...]

Currently Searching...

☑ SiRO	5,304
□ 18th Connect	30,194
□ MESA	1,982
□ MODNETS	537
□ NINES	73,385
□ Great Lakes Aggregator	144

Limit Results to...

Select a resource below to limit your search results

Resource	# of Objects
Journals	1,810
Other Digital Collections	3,494

Figure 33: Collex Linked Data browser

Figure 33 presents a screenshot or the Collex platform as implemented at the Michigan State University Library's *Studies in Radicalism Online*. Much like Blacklight, Collex provides a fairly traditional library interface to its Linked Data data-store. But it also includes several features designed to capitalize on the extended web of Linked Data information. Note the "Currently Searching…" at the top of right side menu column of the screenshot in *Figure 33*. Built into the Collex platform is the ability to direct the platform to either query an aggregated triplestore or query a configured list of SOLR endpoints at other institutions, thereby producing an aggregated search and browse environment. The aggregating through linking functionality of the platform represents a significant step towards the type of network expanded search and discovery made possible by Linked Data.

As noted at the beginning of this section, we are only beginning to explore the potential of Linked Data discovery, and experimentation along these lines is likely to continue for the next several years. Two important initiatives devoted to this area of research are the Mellon funded Linked Data for Libraries (LD4L) and Linked Data for Production (LD4P) initiatives.[52] These ongoing initiatives bring together Columbia, Cornell, Harvard, Library of Congress, Princeton, and Stanford University in a combined effort to examine the potential for Linked Data production and discovery, building on products such as Hydra, Blacklight, Fedora, Vivo, and Vitro.[53] LD4L and LD4P has initiated wide engagement with the library community, and are actively developing new search and discovery methodologies and platforms based on their own research and engagement with other libraries.

It is difficult to predict exactly what new search and discovery approaches and capabilities will be developed out of initiatives like LD4L and LD4P. However, we already have enough examples of novel interfaces to begin to see some of the possibilities. Importantly, we currently lack sufficient library Linked Data data-stores to properly test and develop scalable Linked Data search and discovery platforms. We can, however, reasonably expect the functionality of existing systems, which already provide capabilities on par with current library

[52] See https://www.ld4l.org/.

[53] See https://projecthydra.org/, http://projectblacklight.org/, http://fedora-commons.org/, http://vivoweb.org/, and http://vitro.mannlib.cornell.edu/ respectively.

search and discovery, will expand over time as Linked Data uptake in the library community expands.

XI. Survey of Current Library Linked Data Implementation

Below is a representative Survey of national and research libraries currently engaged in Linked Data implement or Experimentation. It is by no means a comprehensive list. Rather, it is meant to serve as an indicator of the various states of adoption and readiness.

Library of Congress:

Library of Congress has initiated several Linked Data projects since 2009. It created its Linked Data Service by publishing its authority data and other vocabularies in Linked Data format (id.log.gov) and developed BIBFRAME model. In late 2015, LC launched its BIBFRAME pilot project which was designed to test "efficacy of BIBFRAME." More than 40 LC catalogers participated in Phase One of the pilot (Oct. 2015-March 2016). The report and assessment of the pilot project can be found at: https://www.loc.gov/bibframe/docs/pdf/bibframe-pilot-phase1-analysis.pdf. LC is planning to launch Phase Two in early 2017 which will experiment with BIBFRAME vocabulary version 2.0.

Library of Congress has created three sets of BIBFRAME tools which libraries can use for their own experiments:

1. BIBFRAME Editor: Library of Congress BIBFRAME Pilot Training for Catalogers Module 3 Unit 2 (http://www.loc.gov/catworkshop/bibframe/Module3Unit2External.pdf) provides detailed instructions on how to use the Editor.
2. BIBFRAME Profile Editor: The document, *BIBFRAME Profiles: Introduction and Specification* (http://www.loc.gov/bibframe/docs/bibframe-profiles.html), defines what BIBFRAME Profiles are and describes how they are constructed.
3. MARC to BIBFRAME transformation tools: The software that underpins these two services can be downloaded from GitHub site: https://github.com/lcnetdev/marc2bibframe.

National Library of Medicine:

The National Library of Medicine was one of the BIBFRAME Early Experimenters and a registered BIBFRAME Early Implementer. In late 2014, NLM proposed a modular approach to BIBFRAME (BF) experimentation through development of a core ontology, i.e., a widely shareable BF vocabulary, that could be extended with other RDF ontologies for greater granularity. To test this approach, NLM collaborated with Zepheira, George Washington University, and University of California, Davis (UCD) in the early development of the BIBFRAME Lite (BF Lite) ontology suite (http://bibfra.me/).

In addition to its early work on BIBFRAME and BIBFRAME Lite, in 2014 NLM published beta versions of two of its well known datasets as Linked Data: PubChemRDF, containing information on the biological activities of small molecules (https://pubchem.ncbi.nlm.nih.gov/rdf/) and MeSH RDF, NLM's thesaurus of Medical Subject Headings (https://id.nlm.nih.gov/mesh/). Both RDF products are searchable from their own SPARQL query interfaces or querying can be directly integrated into programs and services using their SPARQL endpoints.

George Washington University Libraries:

The Linked Data experiment George Washington University Libraries (GW) has undertaken is to insert URIs into MARC records by utilizing MARCNext Linked Identifiers of the MarcEdit open source Tookit. To date GW entered over four million URIs into its existing legacy MARC records. Internal cataloging workflow was adjusted to accommodate and ensure the newly added MARC records will contain appropriate URIs. The embedding of URIs via MARCNext is automatically set to Library of Congress ID service (id.log.gov) as the default setting. Additional vocabularies, e.g. VIAF, other national libraries, and OCLC are also possible. The URIs are embedded in $0 of the following MARC fields: 1xx, 240, 6xx, 7xx, and 830. GW's URI project is an excellent example demonstrating how libraries can apply Linked Data concepts within their existing MARC-centric systems as a transition to a data model that is more Linked Data friendly.

Linked Data for Libraries and for Production Projects:

Linked Data for Libraries (LD4L) is a series of two integrated projects (LD4L and LD4L-Labs) supported by grants from the Andrew W. Mellon Foundation (http://www.ld4l.org/ld4l-original/). They involve several major research libraries, including Cornell, Harvard, Stanford, and the University of Iowa. Those projects were designed to examine and test the discovery of Linked Data (LD4L) and to create tools and services that support the creation of Linked Data (LD4L-Labs). The ultimate goal of the LD4L projects is to create Linked Data solutions and infrastructures that can be implemented in a production environment at research libraries within the next three to five years.

Linked Data for Production (LD4P) is a related project also funded by the Mellon Foundation. This project involves Columbia, Cornell, Harvard, the Library of Congress, Princeton, and Stanford. The goal of LD4P is to begin the transition to the native creation of Linked Data in a library's current production environment using existing tools. Issues LD4P discovers with current tools will be fed back to LD4L-Labs to aid them in their future tool development.

The British Library:

According to Neil Wilson, Head of Collection Metadata at the British Library:

> The British Library believes Linked Open Data to be a logical evolutionary step for the established library principle of freedom of access to information. As such it offers trusted and authoritative knowledge organizations such as libraries a new important role in the emerging information landscape. The vision of a global pool of semantically rich, reusable metadata is also highly attractive to such organizations by enabling the concentration of scarce resources on adding unique value. Similarly, the potential of Linked Data for cost effective exposure of library datasets to search engines, application developers and new forms of resource discovery has significant appeal.

Given the above commitment, it is not surprising that the British Library was an early adopter of Linked Data technologies. In 2011 they released metadata for a subset of the British National Bibliography (BNB) under a Creative Commons CC0 1.0 Universal Public Domain license. Metadata from the bibliography is made available via a SPARQL endpoint (bnb.data.bl.uk) in addition to downloadable data dumps serialized as RDF/XML and N-Triples.

Data in the BNB dataset is described using the following vocabularies:

- Bibliographic Ontology
- Bio
- British Library Terms
- Dublin Core
- Event Ontology
- FOAF
- ISBD
- MADS/RDF
- Org
- OWL
- RDA
- SKOS
- WGS84 Geo Positioning

Explicit URI based linking in the metadata is established with the following data sources:

- ISNI
- VIAF
- LCSH
- Lexvo
- GeoNames
- MARC Country and Language codes
- Dewey.info
- RDF Book Mashup

The Library’s Linked Data model gives preference to the use of pre-existing ontologies in order to ensure the widest options for interoperability with new user groups in addition to the library world. Since its creation, the Linked Open BNB has continued to be enhanced through the addition of new features such as the ISNI and regular monthly updates. The British Library is also looking at new options to expand its Linked Data services including the use of content negotiation for selected areas of its web site to offer Turtle (TTL) versions of pages in addition to standard HTML.

Although The British Library was a member of the BIBFRAME early experimenters group and remains committed to offering linked open data, it currently has no plans to adopt BIBFRAME as the framework for doing so. This does not mean the British Library is uninterested in BIBFRAME; but simply reflects that the organisation is waiting for BIBFRAME to achieve the necessary maturity, stability and critical mass of users in order to justify the commitment of limited development resources.

University of California, Davis:

The research contained in this report is a reflection of the current state of the UC Davis Library's engagement with Linked Data. The UC Davis Library plans to continue its efforts on this front by moving to implement the transition plan suggested in this report as part of its regular operations. To this end, a Linked Data transition has been formed, and we expect to begin an implementation process during the Summer of 2017. As we engage in this transition, we will continue to update this roadmap to document the transition process and communicate specific, relevant findings. Changes will appear in versions made available at bibflow.library.ucdavis.edu.

Appendix A: Vendor Engagement Matrix

Service Provider	Experienced demand?	Established collaborative partnerships?	Published public documents?	Linked Data Services	Service Availability Date	Area(s) of Focus
Atlas Systems	yes	yes	no	Service currently in development, scheduled for release in 2016	Summer 2016	Archives and Special Collections, Rare Book Libraries, interlibrary loan, resource sharing
EBSCO/ Novelist	yes	yes	yes	Linked Library Service	April 2016	Public Libraries, reader's advisory, enrichment, open-source Library Service Platform software
Innovative Interfaces	yes	yes	yes	Innovative Linked Data	April 2016	Public and Academic Libraries, users of Polaris, Sierra, Millennium and Virtua
OCLC	yes	yes	yes	VIAF, FAST, WorldCat Entities	2011	Authoritative link creation and management, WorldCat services
OverDrive	yes	yes, informally	no	Limited schema.org decoration for harvest in OverDrive Web interfaces	2012	Public Libraries, ease of access to eBooks, videos and audiobooks
ProQuest/ Ex Libris	yes	yes	yes	Alma and Primo Linked Data features currently in development	2016, Release Date TBD	Academic Libraries, Alma and Primo customers
SirsiDynix	yes	yes	yes	BLUECloud Visibility	2015	Academic, Public, and Special Libraries, BLUECloud Library Services Platform customers, users of Symphony and Horizon
Zepheira	yes	yes	yes	Library.Link Network	2014	Global Linked Data infrastructure, service provider partnerships and collaboration

Appendix B: Glossary of Terms

Actionable [Machine]: An object is said to be machine actionable when it is in a form that allows a computer to interact with it in some automated manner.

Application Programming Interface (API): An application programming interface (API) is a set of communication protocols that provide a clearly defined method of communication between various software components, programs, or network services.

BIBFRAME: Short for Bibliographic Framework—a data model created for bibliographic description. The design of BIBFRAME began in 2011 through a partnership between the Library of Congress and Zepheira. BIBFRAME's goals include the replacement of MARC encoding standards with methods that integrate Linked Data principles in order to make bibliographic data more useful both within the library professional community and to the world at large.

Crosswalk: The process of migration data from one serialized form to another.

Disambiguate: A process directed at distinguishing between distinct entities.

Graph: A graph is a data arrangement that consists of nodes (objects) connected to each other via edges (relationships). A family tree is a common example of a graph where the persons represent nodes (John, Jane, etc.) and relationships represent edges (child, parent, etc.).

International Resource Identifier (IRI): An IRI is a version of a URI that is encoded in a form that can render international characters.

Linked Data: According to the W3C the term Linked Data refers to a set of best practices for publishing structured data on the Web that includes the use Uniform Resource Identifiers (URIs) as names for things, the use of HTTP URIs so that people can look up those names, insuring that when someone looks up a URI, provide useful information, and including links to other URIs so that users can discover more things. Additionally, Linked Data describes a semantic data structure based on collections of n-triples preferably (but not necessarily) serialized as RDF. See https://www.w3.org/wiki/LinkedData.

LD4L: Acronym for Linked Data for Libraries, a Mellon funded initiative focused on examining a variety of issues surrounding Linked Data implementation in libraries. See https://www.ld4l.org/.

LD4P: Acronym for Linked Data for Production. An extension of the LD4L project. See https://www.ld4l.org/ld4p/.

n-triple: An n-triple is the fundamental structure of Linked Data graphs, wherein relationships between objects are described through "subject::predicate::object" statements. For example, "John::hasMother::Sarah" is an n-triple.

Resource Description Framework (RDF): A standard model by the World Wide Web Consortium (W3C) for expressing Linked Data on the Web. See "Resource Description Framework (RDF) Model and Syntax Specification". 22 Feb 1999. Accessed August 1, 2015. http://www.w3.org/TR/1999/ REC-rdf-syntax-19990222/.

Serialization: Serialization is the process of representing data in a particular form. In the Linked Data universe, this refers to the one of many forms that can be used to represent n-triples. Examples of such formats include RFD, Turtle, JASON, etc. A non-technical way to understand serialization is to think of it as the way a triple is formatted.

Schema.org: A Linked Data standard ontology implemented by most major search engines. See http://schema.org/.

Thin [MARC, Record, or Graph]: A thin information is a sparse collection of data that describes only the minimum necessary depth of information for a particular context as opposed to the full range of information that may be known about the object.

Traversable: When a person or computer follows the chain of relationships represented by a data graph, moving from one related node to the next, she/he/it is said to traverse the graph.

Uniform Resource Identifier (URI): In information technology, a Uniform Resource Identifier (URI) is a string of characters used to identify a resource. Such identification enables interaction with representations of the resource over a network using specific protocols. In practical terms, to human readers URIs look like the URLs used to navigate the World Wide Web. URIs, however, by convention, are intended to be permanent identifiers for a resource, regardless of it might live (or move to) on the network. In other words, an item’s URL could change, if, for example, a web-based resource moved to another hosting environment, but its URI would not and any person or machine that traverses the URI would be directed to the current URL for the resource.

Workflow: The steps involved in completing a defined task.