

Summary notes and comments on BIBFLOW Roadmap

Simeon Warner - 2017-07-21

“BIBFLOW: A Roadmap for Library Linked Data Transition”, MacKenzie Smith, Carl G. Stahmer, Xiaoli Li, Gloria Gonzalez, 2017-03-14. https://bibflow.library.ucdavis.edu/wp-content/uploads/2017/03/bibflow_roadmap_revised_3_14_2017.pdf

Summary notes and comments by section:

I. Introduction

Argues that libraries are well positioned for the transition to linked data because of both collaborative culture and data structures (including authority and identity use).

II. Why Linked Data

Pretty standard pro-linked-data introduction. Argues (correctly) that linked data aims to make a machine traversable web. To my mind, over-emphasizes that another key driver is independent extension of the graph (data model) avoiding top-down process. I don't think this is a key near term goal (beyond a parallel with local use of 9xx fields served the purpose of local extensions in MARC) and there are significant dangers and hurdles. (There are later comments about mixing MARC and EAD data which does seem like an aspect of this that is more relevant.)

III. Transitioning Fundamentals

Argues that success is mapping to linked data plus URI assignment and reconciliation. Argues that the “*primary obstacle*” is reconciliation (association of literals with URIs).

IV. Roadmap Overview

Notable that in the 7 principles there is no mention of improving the function or utility of the catalog for users:

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)

propose two phases, each with three components:

1. Phase one

- Linked Data Lookup MARC Cataloging – sounds like \$0,\$1
- Batch MARC URI Insertion – more \$0,\$1
- Linked Data Import/Export API

2. Phase two

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

An implication is that completions of phase one is predicated on round trip conversion to support input/export as linked-data.

V. Phase One: Linked Data in a MARC Exosystem

Argues that the ILS is highly connected with other systems, notes dangers of high employee retraining and efficiency costs.

In BIBFLOW, KUALI-OLE was modified to better support linked data lookup (how similar is this to Questioning Authority work? Note images on p11) and to support URIs.

Comments on second step – batch URI insertion – experiments at GWU, and notes alignment with [PCC work to define \\$0, \\$1](#)

Discussion of support for APIs is very fuzzy, mainly pushing it onto ILS vendors. Does not point out that with good converters could even externally shim a MARC import/export to talk BIBFRAME.

VI. Phase Two: Transition to Native Linked Data Ecosystem

Odd comment in introduction that in a linked data ecosystem data is exchanged “*serialized as n-triples (the most familiar form of which being RDF)*” – I’m not sure whether that is intended to say something specific, or is just a little mangled. Also asserts that MVC is *the* right model for software systems which seems a little bold. Notes that both Kuali-OLE and MarcEdit rely on SQL – Kuali first used MARCXML data but it was too slow.

Again suggests that the focus of transition to linked data cataloging is to “capitalize[s] on the extensibility offered by working with graph-based data models”? On p17 there is another lookup screenshot. Describes Zepheira’s [Scribe editor](#) which is configured with different profiles for different object types, notes mixing EAD and MARC data.

Mentions practice of making linked data records tied to stub in legacy ILS (OLE in their case), suggests that parallel universe is a good way to proceed because it allows phased migration.

In discussion of second step – batch conversion – mentions LC MARC->BF1 tool and development of MARC->BF2 tool. Notes that MarcEdit can run XSLT transformation and could thus run LC converter (though notes Windows only). Mentions XC and MST edit tools, and Zepheira and OCLC as vendors also, but very vague.

Claims then that last step is to transition other data, but warns not to try everything at once.

VII. Transitioning Workflows

Claims that MARC->LD transition is about as difficult as transition from one MARC environment to another – which implies that it is really just about learning a new UI. Did experiments on 5 workflows, reports on three categories of work. Figure 19 (p27) has nice picture of system interconnections.

Copy cataloging workflow - Argues essentially same process as now from the catalogers point of view: “*searching databases, finding existing bibliographic data, making local edits, checking access points, and saving data into a local system*”. Nice description of the data transfers according to components of figure 19. Note that in this model OCLC data comes as schema.org and thus there is an assumed converter from schema.org to BF. Suggests that use of linked-data authorities might actually offer an efficiency gain over current practices.

Original cataloging workflow - Argues that ability to create new local authority on-the-fly will lead to efficiency increases in cases where no existing authority is found. Notes need to authority reconciliation to include local data as well as established authorities, suggests extension to include identities from ORCID, VIAF, Wikipedia, etc.

Serials cataloging - Notes complexity of holdings information with changing titles, publishers, frequency, numbering, etc., and deficiencies in BIBFRAME for dealing with this.

VIII. Authority Control

Opening “*Authority control is the area of Linked Data transition that has caused the most concern*” seems odd as I think it is one of the most compelling parts of the transition. In later discussion it seems that the main worry is about moves to a fully decentralized model, though later still it is argued that (following experience from evolution of web search) there will end up being centralized authority services based on distributed data, and these might

actually be operated by familiar players (or organizations like them). Mentions Cornell IMLS Authorities grant for way forward with community principles.

IX. Vendor Engagement

Notes that most vendors are unaware or minimally aware of linked data. Pages 47–51 detail status for a few vendors doing linked data (mainly Zepheira partners).

X. Discovery

Notable that this section is the first time possible impact on discovery is mentioned, opens with reasonable preface:

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.

Discussion of discovery systems quite limited. Oddly mentions Blacklight first, and though correctly notes it is “traditional”, says it can be deployed over indexes build from linked data. Also mentions [Collex](#) search engine which provides similar features by building Solr index from linked data, an example instance is [SIRO](#) which doesn’t show much linked-data-ness in the browse interface. Finally mentions LD4L Labs and LD4P work as technologies being built upon.

XI. Survey of Current Library Linked Data Implementation

Somewhat selective list of libraries engaged in linked data experiments, no mention of DNB or National Library of Finland for example.

- *Library of Congress* - [BIBFRAME](#), BIBFRAME editor, [profile](#)) editor, [MARC->BIBFRAME1](#) converter
- *National Library of Medicine* - experiments, published [PubChemRDF](#) and [MeSH](#) as RDF, with SPARQL endpoints.
- *GWU Libraries* - has added 4M URIs into its MARC records
- *LD4L Labs and LD4P* - tool development and linked data production experiments
- *British Library* - British National Bibliography ([BNB](#)) is CC0 RDF with SPARQL
- *UC Davis* - developed and following this plan