Linked Data for Production Report, April 2016 (Partner Meeting #1)

Linked Data for Production

Report

April 2016, Washington, DC

Present:

Columbia: Kate Harcourt, Melanie Wacker

Cornell: Naun Chew, Christina Harlow, Jason Kovari, Dean Krafft, Jim LeBlanc

Harvard: Steven Folsom, Marc McGee, Randy Stern, Robin Wendler

Library of Congress: Judith Cannon, Paul Frank, Sally McCallum, Beacher Wiggins

Princeton: Joyce Bell, Peter Green, Tim Thompson

Stanford: Nancy Lorimer, Darsi Rueda, Rob Sanderson, Philip Schreur

Linked Data for Production Goals and Needs

The meeting began with a review of the goals of the Linked Data for Production (LD4P) Program. According to the grant proposal, the LD4P Program will be the overarching project tying together the linked-data representation of the inter-institutional workflows, protocols, and data sharing of the individual members of LD4P. In addition, the Program will be the bridge between LD4P and the broader metadata community. The goals of the project are three:

- Reinvent metadata production for any resource a library curates making use of linked data in a cooperative, distributed environment
- Essential tool and system configuration to pilot this metadata production
- Develop a community/collaborative framework for coordinating the ontology and technical work

The group agreed that the development of the cooperative, distributed environment would be both the most difficult and potentially most important contribution of the project. The discussion of this topic was set for later in the day.

The group moved on to discuss the tool sets we anticipate using. The tools ranged from Vitro and the LC tool suite to WeCat/ALIADA (developed by Casalini) and the Cataloger's Workbench Editor being developed at Princeton using the W3C XForms standard. A formal assessment of these tools will be part of our efforts and we will want to speak with LD4L-Labs about the sequencing of any enhancement requests. The MARC to BIBFRAME converter will be of great importance to all. LC will be developing a new version of theirs based on BIBFRAME 2.0. LD4P/LD4L-Labs may develop one as well. LC would be interested in a BIBFRAME to MARC converter if LD4L-Labs or OCLC took this on. There would be a need to create operational records out of BIBFRAME but more fully fleshed out MARC as well for those libraries that are still dependent on MARC. We would need to be cautious, however, not to artificially contort BIBFRAME so that we could create better MARC records in the conversion process.

We will also need a tool for ontology development. The tool will need to allow for community involvement as the ontologies are developed and we will need to resolve how these ontologies are preserved and maintained. Two initial possibilities would be with the societies that help us develop them or the Program for Cooperative Cataloging providing a central service that would allow for their access and preservation.

As the group discussed our production of linked data for discovery, a number of issues surfaced such as what would be our metadata of record. We also realize that the transition period from MARC to linked data will be a lengthy one for some libraries. How can we ensure that our staff will not have to work in both worlds for an extended period? What will be the relationship between this discovery metadata and the operational records we presume will drive other functions in the ILS? Although we may not care to keep the metadata in the operational record in synch with the linked-data discovery metadata, how will we handle updates that continue to come in as MARC? Much of this must be resolved by close collaboration with the three ILS vendors represented in the project (SIRSI-Dynix, Ex Libris, OLE). OCLC will also hold an important place as a central distribution point and perhaps a role in identifier management and BIBFRAME to MARC conversion.

The morning session closed with a discussion of the place of PCC standards and RDA in the production of LOD. Both will need to change to adapt to the new environment. We will be moving into a world of greater variability. The focus should be on how to best make our data shareable. Do we need to share everything? Some data will be private. Other metadata currently in MODS is not routinely shared. LC will be in a special position as they will need to be able to support BIBFRAME and MARC customers for a long time. What is the group's responsibility in furthering the transition?

Communication Channels

LD4P will develop a public and private wiki space on Duraspace. Both will be managed by the Program Manager. We would like to include the following types of things on the public wiki:

- Public version of the grant
- Common procedures and tools
- Updates on the individual projects
- Speaking engagements with copies of the presentations
- Update on support efforts from the PCC

We will keep the LD4P Google Group for ourselves and make use of other, already established lists to disseminate information. We will establish three formal groups that will meet regularly via Webex to discuss developments across LD4P and LD4L-Labs: an Ontology group, a Technology group, and a monthly all-hands call. Various smaller groups will be developed on an ad hoc basis for particular issues.

We will also need to develop a non-intensive way of keep the six institutional projects aware of what the others are doing. Perhaps the private wiki would be the best place for that.

LD4P and LD4L-Labs

The afternoon meeting began with a discussion of the relationship between LD4P and LD4L-Labs. D. Krafft outlined the main foci of LD4L-Labs:

- Infrastructure needed for both projects
- Tools and techniques to:

Make use of the linked data created in the first two years to enhance discovery and visualization

Linked data creation/editing tools for annotation, organization and crowdsourcing; the conversion of non-MARC data; and the connection/integration with other resources such as the Hydra stack and Vitro

Ontology work, reconciliation, and persistence

Service to establish "same as" relationships

Metadata creation tools (to help with geospatial data sets, archival film, converters)

Reconciliation and persistence (and what that means) will be of major importance, including reconciliation against what (local data, group data, all data). We will need to make use of various international identifiers and local ones as well. We have not guaranteed the persistence of local identifiers for this project but what will this mean moving forward. There is reluctance to lose the intellectual work needed in the reconciliation process.

We would like to have the 6 month LD4P and LD4L-Labs meetings run back-to-back. The LD4P meetings will be in DC and the LD4L-Labs meetings will be in Cornell following them. We will also need to identify and prioritize the specific support needed from LD4L-Labs for LD4P.

OLE

The Kuali Foundation reexamined itself and made a serious transition, moving development to a commercial entity. OLE therefore also reexamined itself and has decided they need to redevelop, partnering with EBSCO and IndexData (Denmark). Their concept will be to go back to the original concept of OLE with a modular framework that can plug in various elements and have services built on top of it. There will be a viable replacement product by 2017. OLE would like to include linked data into its new structure but will not include any developments that happened in connection with BIBFLOW as they were developed around the old code base.

Common Technology

The group discussed various common tools we would all use:

- MARCtoBF Converter: This converter is mentioned in both the LD4P and LD4L-Labs grants. There will also be one developed by LC that will be in line with BIBFRAME 2.0 and another developed by Casalini. The LD4* converter may be developed in a modular way so that it could have the potential to convert multiple formats and produce various RDF outputs. There will also be the need to convert non-MARC metadata to BIBFRAME.
- BIBFRAMEtoMARC Converter: Most projects would need this tool eventually to create the MARC operational record for their ILS if they create new discovery metadata in BIBFRAME. LC would like a more complete MARC record that they could use to fulfill the metadata needs of those libraries that will still require MARC for their processing systems. The tool could be developed by LD4L-Labs but may also be of interest to OCLC for development.
- Lookup Tool: All projects will need to lookup and retrieve metadata from their triple store. This may not be a common, modular need, however, as some systems have it bult in (Vitro, WeCat/ALIADA)
- Triple store: Each institution will need its own triple store but we will possibly need one for the group as well. It is the question of aggregation or federated search. We will need it for discovery, to locally answer calls from each other, capture new assertions, batch updating via notifications, etc. How will we handle updates to each other's data? Actual correction or new assertion?
- Linked data fragments and index (eg solr): This would be needed to support discovery and visualizations in real time.
- Infrastructure to support URIs (id.stanford, id.cornell, etc. maybe), including identity management system.

Individual Project Updates

Columbia: Columbia will be making use of VIVO/VITRO and will also be looking at the BIBFRAME Editor to see if it can be extended and used locally. They have mapped a subset of art properties to MARC and discovered that some of the changes they made to the data for BIBFRAME were the same as for MARC. The collection set is between 10K-12K resources but not all of those have been cataloged at this point. The initial test set used for modeling will be about one hundred descriptions. Once a model has been developed additional descriptions may be converted or created from scratch. They have MARC representations of their data so they will not need to generate operational records. The group will use KARMA for visualization. An important consideration will be how to separate private vs public data. Columbia did a literature review of how art objects are modelled and will make that available to the project.

Cornell: Cornell has two projects, cataloging LPs from the Bambaata collection and the rare materials ontology extension. The cataloging project is currently in the discussion phase. There are two simultaneous projects; LD4P which involves BIBFRAME and a separate NEH grant for processing and digitization. Cornell will not be doing double cataloging for these projects. Some resources will be done in BIBFRAME and others in MARC. They will be speaking with their curators about the inclusion of annotations. The rare materials ontology extension will be done in conjunction with RBMS and will be discussed at the PCC participants meeting at ALA Annual in Orlando. Preliminary work is being done with RBMS on what new predicates would be needed. If necessary, Cornell will be creating local authorities or pulling identifiers outside of the LC NAF.

Harvard: Harvard will be working on traditional cartographic resources such as maps, atlases, and geospatial data. They don't expect a radical move away from BIBFRAME but there will be a lot of traditionally coded MARC data that will need to be included. The working group has been formed through outreach to ALA MAGIRT and the Open Geometadata communities and has already held their first virtual meeting (3/23/16). Members of the group include the LC Geography and Map Division catalogers, Kim Durante (Stanford), Kathy Weimer (Rice, geohumanities); MAGIRT members Paige Andrew (Penn St), Louise Ratliff (UCLA), and Shannon Erb (Minn.). Biweekly working group meetings will be scheduled. A wiki has been started at Harvard but will be moved to the new LD4P wiki. The group has already done a poster presentation (3/30) with project overview at American Assoc. of Geographers meeting, gathering use cases from geographers. The group will need support for an RDF editor (including shareable web version) and ability to support specific cartographic descriptive properties widgets such as a bounding box generation tool.

Library of Congress: LC's initial BIBFRAME pilot ended on March 31st but the AV/Sound Recordings part of it will be extended through June. LC is now analyzing the outcomes of the pilot and will report at ALA Annual in Orlando. Staff will maintain their skills in working with BIBFRAME by devoting one day to BIBFRAME cataloging per week until the next pilot begins in October. LC's LD4P projects will include AV/Sound recordings, Prints/Photographs, General Collections (with emphasis on RDA), and BIBFRAME 2.0. LC will work with other members of LD4P when their domains overlap.

Princeton: Princeton has a better idea of what the Derrida materials look like now. Over 6K items have dedications. The collection as a whole has an EAD and some of the monographs are already present in the University's collections and thus are represented with full MARC records. Princeton will do some augmenting of current metadata and some original description using BIBFRAME. They have been experimenting using WebAnnotation for their dedications. Princeton is also considering an EADtoBIBFRAME converter and will need to link their work to Aeon.

Stanford: Stanford has two projects, the Performed Music Ontology and the Tracer Bullets. The Performed Music Ontology (PMO) will be done in collaboration with domain experts (MLA and ARSC) and other music parties in LD4P. Some work on the ontology for medium of performance is already underway. A key element will be examining the MARCtoBF converter to optimize it for performed music. There is also much interest in event and technical data. LC will be focusing more on music in relationship to RDA, the PMO will take a broader perspective. Another important aspect will be how the PMO extension to BIBFRAME will be hosted and sustained. One possible output may be best practices for doing that. This new ontology should dramatically improve searching for music materials once discovery environments can take advantage of it. Stanford has four tracer bullet workflows planned: vendor-supplied copy, original cataloging, deposit of a single item to the digital repository, deposit of a collection of items to the digital repository. The first workflow will have two variations, one beginning with MARC copy and one beginning with BIBFRAME. Stanford is beginning by mapping out the first workflow and making decisions about what needs to be resolved at this first pass and what can be deferred as they are working with a tracer bullet model.

Technology needs

Following the project updates, the group laid out their technology needs to help prioritize the LD4L-Labs development queue. The identified needs are:

- Place for local authorities
- VITRO support for public vs private information (need for other implementations as well)
- Hosting of community efforts--what technology do they need
- Visualization tools
- BIBFRAME to MARC Converter. For LC, this is the one tool that they would not have themselves. This converter would need to make two different levels of MARC, a brief operational record and a more complete representation that could be used as a surrogate for the BIBFRAME data for those libraries that are still dependent on MARC.
- MARC to BIBFRAME converter that supports extensive editing of BIBFRAME post-conversion.
- Ontology editor. Protege has some complexity and the workflow can be complex. Web Protege has many limitations. LD4L did successfully use Protege with Github and that workflow should be documented.
- BIBFRAME (and other ontologies) editor that can support extensions
- Ability to share local identifiers
- Bounding box tool to format data
- Auto-conversion to convert coordinates to decimals
- Ability to enhance converted data through a converter
- Validation
- Statistics
- Rdf to Solr mapping for BIBFRAME 2.0
- Annotation store; expertise on storing and serving that data
- Tool that supports annotation function
- Triple store to host data created

Timeline review and 6 month goals

The group reviewed the project's goals as a whole for the first six months:

- Hold first 2 day meeting for the core LD4P participants: This has been accomplished with this first meeting.
- Review, coordinate, and adjust the collective goals for LD4P for the first six months: This is the discussion we are having now.
- Design and implement the LD4P wiki: This is part of the Program Manager's JD but we may have to set up a brief version soon as it may take a while to fill this position.
- Scope and design the communal work environment: This will need extensive discussion and will be part of the duties of the Project Technologist.
- Share development of the MARCtoBIBFRAME converter with LD4L-Labs: This converter is the top development priority across all projects.
- Define the parameters for interaction with the Standing Committee on Standards of the PCC: PCC is well represented in the membership of this project. More intense interactions will take place through PCC TG work. PCC will have a place on the project wiki.
- Select core set of vendors to contact: The group discussed which vendors should be contacted very soon:

SIRSI Dynix

Ex Libris

OLE

Casalini

Harrassowitz

Backstage

OCLC

YBP

- Make initial contact with the selected vendors: The group will work on this over the next couple of months.
- Select core set of LOD projects to contact:

LibHub

BIBFLOW

UIUC's project - http://cirss.lis.illinois.edu/News/newsDetails.php?id=130

Getty Research Institute & the Museum community

BL, Europeana, DPLA, etc. -- other orgs deep in this space

Canadian group doing work around sesquicentennial

Next Steps

The group closed the meeting with identifying next steps that should be completed in the next few months:

- Reach out to LD4L-Labs to prioritize and spec-out development for tooling and other coordination
- Make initial contact with OCLC about mutual issues
- Make the initial public announcements about the grant and then separate announcements on various lists.
- J. Kovari will send out Doodle to LD4P and LD4L-Labs to schedule Ontology WebEx
- Set up Email lists
- Select an approximate date for the next meeting. It should be scheduled sometime in October/November in DC with an LD4L-Labs meeting to follow at Cornell
- Coordinate LD4P presentations at ALA Annual in Orlando
- Be part of the BIBFRAME Update session each ALA
- Set up a VITRO webinar together with a sandbox and Docker or AWS instance to drop-in and work within
- Set up an ALA time for vendors to meet with us
- Coordinate project management
- Identify point people at each institution
- Establish high level groups (Ontology, Engineering, All Hands Call)
- Decide on specifics for communication both internally and externally
- Formalize connections with the PCC