Notes on Meetings at ALA Annual, New Orleans, 2018 with a Focus on MARC and Other Encoding Standards, and Linked Data (Plus a Couple Other Meetings of More General Interest)
Submitted July 12, 2018, by Jim Soe Nyun, Chair, Encoding Standards Subcommittee

MARC Advisory Committee I
Session 1: 6/23/18
Sessions 2-3: 6/24/18

Meeting agenda with links to papers: https://www.loc.gov/marc/mac/2018/2018-02.html

Announcements of latest fast-track changes:

- Added $4 to 730
- $r redefined in Field 382 (MLA request)

Proposal No. 2018-02: Subfield Coding in Field 041 for Accessibility in the MARC 21 Bibliographic Format

- Amended to clarify wording of new subfields $p and $r, passed unanimously

Proposal No. 2018-03: Defining New Fields to Record Accessibility Content in the MARC 21 Bibliographic Format

- Amended to drop $0 and $1 as there was not a direct use case that people could see; added $8 to both fields; revised wording of the field's scope and for indicator 2
- Wrinkle for music: No clarifying statement that musical notation is to be understood as text was written into the final text for 341/$a: "Mode required to access the content of the resource without using assistive features (one of textual, visual, auditory, or tactile)." Best practices could help us.

Discussion Paper No. 2018-DP10: Designating Access to Online Resources in Field 856 in the MARC 21 Formats

- Roy Tennant presented for OCLC. Some general background information:
    - 50 million records in OCLC could likely be processed by algorithms to supply an indication that a resource in an 856 is open access
    - Only .12% of bib records have multiple 856 fields
    - Setting a value for "hybrid open access" could require record maintenance
- There was interest in seeing this paper return as a proposal, but the room felt that the discussion paper was not well enough developed to convert to a proposal on the spot
- There was also general agreement that the 856 field is full of ways to record ancient technical information that hasn't had any relevance for many years. There was a hope that someone would undertake this, but it won't likely come out of a proposal based on this discussion paper

Discussion Paper No. 2018-DP11: Open Access and License Information in the MARC 21 Bibliographic and Holdings Formats

- This paper has many relationships to the preceding one (DP10), and once again the current not-useful, not-current state of the 856 field came up. That task would be too large for any revision of this discussion paper, however.
- In the end it appeared that developing Field 540 had the most support, and the paper may return as a proposal that develops this option

Proposal No. 2018-04: Versions of Resources in the MARC 21 Bibliographic Format

- Amended to add $3, $6 and $8 to Option 2, define new Field 251. Passed unanimously, with the acknowledgement that wording is not perfect and that the field may need to be developed as terminology and supporting vocabularies evolve or are established.

Proposal No. 2018-05: Multiscript Records Using Codes from ISO 15924 in the Five MARC 21 Formats

- The committee preferred the option to use $6 and to develop the language in Appendix A of all five MARC formats.
- One abstention, with the rest of the committee voting in favor of passing the proposal.

Discussion Paper No. 2018-DP07: Designating Sources for Names in the MARC 21 Bibliographic Format
Presented by Chew Chiat Naun

- There was general consensus that the basic idea was sound, but many different opinions on what a proposal based on this paper should look like. Some of the issues:
    - The paper looks at fields where names are recorded, but these same fields are often used to record "titles" or "works"
    - The PCC Task Group on URIs in MARC was mainly interested in dealing with names, but it's hard to separate just names from what we could encounter in these fields
    - There are potential problems with a name coming from one authority and the title from another (or not at all, e.g. 100/240)
    - Should $2 be repeatable?
    - "$2 local" can be problematic
    - "$2 viaf" is not authoritative in citing an authority since VIAF aggregates multiple authorities, and might make more sense as a real world object
- The paper may return as a proposal or further developed paper, but there will be a lot to figure out.

Discussion Paper No. 2018-DP08: Use of Field 024 to Capture URIs in the MARC 21 Authority Format

- Option 2 was favored (e.g., developing $0 and $1, but leaving $a for non-URI and non-RWO identifier strings)
- Will likely return as a proposal

Discussion Paper No. 2018-DP09: Improving Subfield Structure of Field 245 in the MARC 21 Bibliographic Format

- Jay Weitz presented this paper, which was developed mainly by OCLC's Leiden office
- General agreement that the 245 is not clear in expressing all the elements that are placed there, with the inability to repeat subfields causing many problems. However there was little agreement on how to improve it. Discussion revealed many issues, including:
    - Neither of the two options presented had general acceptance
    - Examples oversimplified the reality of how complex titles can be
    - Confusion over how to treat parallel title statements versus titles that consist of multiple title, other title, and responsibility statements
    - We could not rely on machines to reliably modify existing 245 statements retrospectively
- Ultimately much reticence to take on this giant task. A straw poll in the room showed 8 people who were interested in developing this paper further, versus 12 who thought this was too big or complex to take on at this late stage of MARC.

OCLC Linked Data Roundtable
6/23/18

[...meeting more than ½ over after a long bus ride and walk to opposite end of convention center...]

Arrive at the end of the second presentation:
Jeremy Myntti: Presenting on the Western Name Authority File
https://www.imls.gov/sites/default/files/lg-72-16-0002_proposal_narrative.pdf

DPLA API being queried
Exploration of where to go next

Non-Latin script resources in BIBFRAME (Sally McCallum)

Currently LC practice in bib records has been:

- Latin only in non-880 fields
- Non-Latin entirely in the 880
- Transliterated data for Main entry, titles, imprints, notes, etc.

Authority has Latin in 100 but non-Latin may be used anywhere else. Problems: Much duplication of effort, transliteration issues including what transliteration scheme was used.

LC's BF WORK practice: transliterated title, names are transliterated or linked; optionally may include non-Latin variant titles in work record or non-Latin script.

All instances link to work, no redundancy for INSTANCE, local data kept with instance description. Some savings in not having to Romanize. Full Unicode available, more scripts.

There will be a presentation later today on this topic. Reference ALCTS Non-English Access Working Group on Romanization Report 12/15/2009

Questions:
Jessica [?, first presenter] on using wiki data input for non-Roman scripts. Actually pretty easy using her interface. The tool is good; problems are conceptual.

Problems with Olmeca-S: Can publish in JSON-LD but only a record at a time. Some issues with searching simply. Reconciliation issues, ways to work with OpenRefine? Currently separate datasets are being exported and worked on.

Any mapping to mapping to MARC, particularly en masse? No work yet. Want to look to implement in the future.

To Jeremy: What will be the model for regional authority hubs, going local to national? Now they send local name candidates to someone to add NACO records. Have done about 20 names but have about 500 more they could think to contribute. (Possibly an issue here of scaling up.)

Work records capture variants, question on whether instance variants could feed up to the work record? Not expecting to automatically populate work records from transcribed instance information.

Open question to Sally McCallum about outgrowths from BF work: They're developing efficiencies with building out modeling, displaying information from linked records.

Question to Jeremy: How would they integrate multiple AFs? Query both LCNAF and Western Names File. Federated search. Sally thought adding URIs to MARC could lead to enriched BF content. People mentioned models with less duplication, parallel statements with and without parallel information.

[Western names question:] After names are reconciled will EADs across repositories be synced up with the unified forms? Not clear whether partners are willing to do this synchronization. Grant runs through end of October. Maybe a subsequent grant.

Question about linked data pilot project, creating descriptions for translators, where they would not transcribe authors, only translators? Author name might be in a fingerprint, but not in the description, proper. Some questions about whether a translation would be a new work under WEMI.

MARC Formats Transition Interest Group
6/23/18

Chew Chiat Naun
PCC Task on URIs in MARC

- Background/history of group
- Basic concept was that some of our structured data is available in external resources, e.g. VIAF. It would help to include these links if we made them available in our records
- Faceted vocabularies have become more available and important
- Steven Folsom's 2015 Best Practices for Linked Data in MARC, was an instigator
- Worked with various players including Casalini, RSC, implementers of URIs in MARC, others
- National (BL, German libraries) libraries have had an interest in this
- Did not intend to make MARC a carrier for linked data. MARC can only go so far, and the intention was not to give MARC eternal life. Still, "MARC is not going away any time soon."
- Some work looked at existing MARC to see how well it could move easily towards linked data. 7XXs are fairly compatible, but others not. Talked about RWOs versus authority control. We can't assume the contents of MARC fields have a unitary vocabulary or authority, so $2 in these fields could be useful. Figured out later in to the work that the headings don't point to entities. Added the use of $4 for a relationship
- Work identifiers, those algorithmically generated by OCLC's project, other sources
- Papers will be heading to MAC to address some of the above
- Look at outcomes page for useful documentation:
    - Real World Object, $1, $0, position paper on expanding the use of URIs in authority records. Best practices group being formed
- Terry Reese pointed out that he can make some of the MarcEdit work available to other developers

- Next: reach out to practitioners to see what they're doing and what they need. Look at use of linked data vocabularies. How can WorldCat identifiers be used and preserved? Where will this work live once done?

Nancy Fallgren
A long and Winding Road to Linked Data
Experimenting with Linked Data at NLM

- NLM's work to add machine dereferencable URIs in their records. Showed some vignettes in their efforts to implement the above in relation to the PCC task force
- They joined the BF experimenters group, then an early implementer group. NLM formed a multi-dimension linked data initiative. Published MESH in RDF. Experimenting with SEO in their digital repository. Involved in PCC linked data working groups
- How has the TF's work intersected with NLM's? Collaborated on some MAC papers. Terry Reese had started work on MARC Next tool, including tool to add URIs to MARC. It proved you could do this programmatically. Resulted in MAC paper to add $0. Terry added MESH RDF to the lookup
- Looked at fields to which $0 applied in data fields, which subfields were part of the label
- NLM wanted to move from theory to actually doing something in RDF. They added identifiers for MESH into MARC records. Was a harder task than originally though. 2017 NLM added MESH URIs to their distribution records
- Digital repo work: Didn't want to work with strings. MARC Next tool opened up working with NLM records. All their repo items start out life in MARC. Schema.org>>JSON LD serialization added to their repo resources
- Casalini had interesting projects going on. Share VDE hs good components, can pull data from various places such as Wikidata, has a working interface, etc. Casalini became a project participant, was interesting that they had MESH. Looking for ways to make MESH URIs more easily available

Considerations related to changes:

- Making changes through the MARC Advisory Committee is slow, and it can take a while to make a change to MARC
- Institutional priorities don't always align with moving ahead
- Technical limitations, e.g. not everyone has a triple store
- FUNDING

Lack of funding is not entirely an impediment. NLM has developed MESH in RDF. Gets some education for funding. Their government mandate says they can't put money towards developing tools. None of their work above was done with external funding. They look for possible collaborations, e.g. PCC URI TF, looking at projects out there that they could work with. They limit what they take on. Other resources: cheap coursework, lots online, work with interesting people.

Selected questions:

- Still working on going from MARC to Schema/JSON-LD. SEO is capricious. SEO-enhanced tools have received more links, and it could expand exposure
- Answer to question about doing things without funding: SHARE VDE had no cost for Phase I. They still have a voice in development even though not part of subsequent formal stages, however

- Are vendors being part of the discussion? Not really at this point. The work was to see what vendors can do. We might need to be clearer with where things are headed before committing to developing products. OCLC has been involved, partly to build on what they can provide. Some vendors were part of information sharing early on. Most vendors now offer URIs in records for free now
- At least one vendor (for Primo product) apparently uses identifiers in records to mine references. Apparently is causing some conflicts, [though I was not totally clear of the problem being stated]

Big Data in Libraries: Friend or Foe?
6/23/18
Intellectual Freedom RT

William Marden, NYPL
Peter Brantley, UC Davis
Erin Bergman, San Jose Public Library

100 seat hall, all full...

How do you define BIG DATA?

- Peter, terabytes to petabytes of data, big amounts of data from systems or sensor arrays. Not all problematic data is only BIG. Erin, echoing Peter's comment, lots of information run through various algorithms.

Is there a diff between big data and surveillance?

- Sometimes it can be the same. When collected anonymously maybe less of a problem. What's collected now may in the future turn into more focused, less anonymous data collection. "Reidentification" is something to be guarded against, mainly coordinating one dataset against another. "Trusted data environments" could be a safer way to go. Safest practice is to collect only information that is being collected; look at how data could be used or misused. Consider the dataset immediately compromised. GDPR is a good direction towards what can be collected.

How do you decide how much to track?

- Best not to track individuals. Be careful about trusting that something is being made anonymous. Be careful about what data from services you take advantage of, e.g. credit scores. Maybe you can ask a user to opt in to doing some analytics on their online behavior. Lessons from the recent European implementation of GDPR: toss what you don't absolutely need. Campuses might do analysis looking at library users. Don't use data collection as a substitute for interactions; connect with the users. Requiring opting in to agree to user terms is NOT an option for public libraries. It's not true sampling if we're working with opt-in strategies, population versus sample analysis. Some data will still end up being collected.
- Libraries are among the most trusted institutions right now. People want control of their information. Every state has a statute protecting library information.

Is it time for different legislation that protects library data?

- Control data ourselves as much as possible; you don't know what a 3rd party vendor would do with data. UC doesn't have a system policy on patron data, probably could use one. Don't track the reading habits of patrons.

Who should be the guardian of patron privacy?

- E.g. no business use for computer logs. Libraries sometime shave non-library staff who aren't up on all our privacy issues. Some organizations have a chief privacy officer, requires deeper pockets. Library needs to be involved in privacy issues for larger entity (e.g. city).
- Bête noir: 3rd party vendors. "Deidentified" data may be in many contracts. No great answer here. Try to constrain use of data as well as to see what if being collected. Understanding what is baseline. Libraries aren't always the paupers they self-describe as being. Vendors can often be asked to change their policies. Overdrive is one vendor that has been problematic. It's okay to expose vendors that aren't complying. NISO has a privacy policy. Ask a vendor if they say "patron privacy is job #1," what do they mean by the statement?
- Beware of ILSs. Beware of profiling or recommending systems. Point to ALA privacy guidelines, a national standard for library vendors. Authenticating via services like FB or Twitter or including like buttons can compromise patron data. Many cool toys, but we can't always be cool. Who are the data beneficiaries when a company folds?—be sure to have that condition negotiated in a contract.

Faceted Subject Access Interest Group
6/23/18

FAST update (Jody DeRidder, OCLC)
Survey from various institutions went out last year. Most responses from US, mainly from tech services and cataloging Interviewed 10 institutions to follow up on survey (BL, Brown, Columbia, Cornell, Harvard, Penn State, Stanford, Yale, Analysis & Posicy Observatory and Royal Melbourne Institute of Technology, the latter two in Australia).

Survey findings:

- Needs FAST supported long-term.
- OCLC is loading FAST terms into the the authorities database.
- Needs to improve current tools, enable batch conversions, and APIs.
- Provide ongoing FAST support.
- Don't make FAST too complex.
- Some confusion about when to use fields.
- How to request new and alternative terms. What is the relationship to LCSH? How to deal with terms that are from different domains.
- Time periods are US-weighted, and may not be clear to other users in other countries.
- Different terms have different meanings internationally. (E.g. "Health & Hygeine" in Australia).
- If we have new terms how should the relationship to LCSH be viewed. Sole source? Work through SACO? Break ties? Treat LCSH as one vocabulary.
- Not ready for a break. But LCSH is often inadequate. What are out options?
- SACO? Community input mechanism? Use external vocabularies?
- Incorporating terms could make FAST easy to use.
- Requests for parallel culturally-appropriate terms, multiple languages. WikiData might have some options, particularly multi-language capabilities.

- Complexities of adding multiple sources, mapping among terms.
- Need more community engagement, a transparent process, a fast process.
- Community voting? Working groups to address various parts of implementation and development. Establish editorial policies.

Next steps to develop editorial policy committee: scope of work, structure, requirements for participation, selection, terms...

Batch Authority Searching with Python (Kelsey George, UNLV) [Recorded presentation] "Very much a work in progress"

Worked with ETDs from ProQuest; destination was their IR.
PQ sent Excel spreadsheet with metadata.
The above works with PQ MARC records that have their own scheme of assigning terms.
Went from records using MarcEdit to OpenRefine. Cut and pasted the facets into a single column then de-duped. Ended up with 257 unique values that she fed into OCLC's search for authorities. Came back with 11 successful searches, 20 errors and 226 too many matches. The tool searches all of the OCLC record. Ended up doing a manual browse of the terms, and correcting records based on the results.

Took above to FAST converter.

Would it be useful to develop a Python-based toll for long-term work?
What features?
Send Kelsey.george@unlv.edu if you have answers or ideas.

Open discussion on the use of vocabularies designed for faceted subject access (moderated by Lynn Gates and Lucas Wing Kay Mak) [No comments to moderate. Meeting adjourned]

Metadata Interest Group
6/24/18

PROGRAM

The Cataloging Lab: Encouraging Collaboration for a More Effective and Inclusive LCSH Proposal Process (Violet Fox, Dewey Editor at OCLC, as of Monday; this is a personal project, not something from OCLC)

Slides will be up at ALA Connect

Grew out of a Google Doc process to propose a new LCSH term.
Wiki platform grew out of the above.
Talked about the current SACO program and process. It takes 6-10 weeks for one round; may extend if a proposal needs to be revised.
Cataloginglab.org is a WordPress site that enables submitting LCSH headings. A template heading is prepared and then put up on the wiki for comment.
The process is an opportunity for us to be responsible with the vocabulary we use. A group process can help reduce individual biases.

Questions:

- Revisions? ("Transsexuals" came up as a problematic term.) Showed example of headings in process, including one where a term "ableism" was added as a reference.
- Demo'ed site briefly. She has de-emphasized MARC to break down barriers to contribution from communities outside cataloging.
- Who submits? Violet Fox right now. But everyone is currently empowered to submit to LCSH.
- Comment from Paul Frank: He's very encouraged to see projects like this.

BUSINESS MEETING

Introductions

MLA Report (see below, at end of the report on this meeting)
CC:DA Report: Currently light in their activities because of the RDA freeze. RDA beta site is up; materials will be put up on the RSC site. Midwinter will have another iteration of Toolkit sessions. Toolkit webinars are coming up. RDA in its original form is frozen; no further updates.

Tomorrow's CC:DA will look at impacts on MARC21 development.

LITA/ALCTS Metadata Standards Committee: Recently adjusted its charge from the original more amorphous charge.
This year the business meeting will look at the transition from two current chairs to new ones.
This afternoon there will be a program on inclusion and diversity issues as they relate to metadata, followed by a discussion where attendees will be welcome to share their ideas.

Voting for officers:
On ballot, Darnell Melvin, vice-chair and Chair Elect Ming yawn Like, programming co-chair; Kelsey George and Rachel Tilley Blog Coordinator; Rachel is the winner.

Scott Carlson from DLF
DLF AIG Metadata: Google "Metadata Clearinghouse" an attempt to collect application profiles at a point in time. An attempt to help share profiles for a lot of profiles to look at how others have dealt with the same materials / the site has things like code and best practices

Brief Report on Metadata Activities of the Music Library Association Report prepared by Jim Soe Nyun for the June 24, 2018 meeting of the Metadata Interest Group

Background: The Music Library Association has one overarching Cataloging and Metadata Committee, plus three subcommittees, one of which is the Encoding Standards Committee, for which I'm Chair until February of next year. The other subcommittees have some folio assignments related to link data, particularly vocabulary development and maintenance. The Encoding Standards Subcommittee has its core assignment MARC and any other schemes that might be used to encode metadata, and as chair of the subcommittee I am designate by MLA as a liaison to this group, as well as to the MARC Advisory Committee.

MLA's Cataloging and Metadata Committee also may establish working groups devoted to a special project. Falling into that group has been the Linked Data Working Group, LDWG, which I co-chaired

along with Kirk-Evan Billet until last fall to help with the Music Library Association's participation in the Performed Music Ontology (PMO) component of the larger LD4P project. As the Music Ontology project began to issue documentation on GitHub and Biblioportal, LDWG was resuscitated with Hermine Vermeij as Chair to review the project documentation and provide comments from MLA as the constituency of primary interest in the work of the PMO group. Several papers were reviewed by the group this past winter and spring, and comments submitted back to Nancy Lorimer of the PMO project.

In more old-school metadata work, MLA proposed several changes to the MARC format for bibliographic data that we hope will help to better associate some parts of the record with other data elements of the same record. The changes to validate subfield $3 for several 3XX data element fields were passed by the MARC Advisory Committee at the time of my last report to this group, and the changes have been published in the recent MARC Update 26, and will be available for use in a few days after the 60 day review period concludes. Finally, the Encoding Standards Subcommittee and other members of MLA provide feedback on papers and proposals being considered at the MARC Advisory Committee meetings, and a consolidated response to the nine current MAC proposals or discussion papers was issued to the MARC listserv in preparation for MAC's meetings this weekend.

LC BIBFRAME Update
6/24/18

Introduction: Sally McCallum
Introduced speakers and presentations

LC BIBFRAME 2.0 Pilot progress report:
Beacher Wiggins

- Working on a more formal report with responses from people working on BR 2.0, will be reviewed and put out soon. Worked with LD4P participants to include their work in the BF test. The BF catalog includes records from the LC catalog that have been converted for testers to consult. 4billion triples from ca. 20 million MARC records in the MARC database. Re the BF Editor, it can deal with in-process records, can save. Now can deal with many formats done under BF 2.0. Their work includes rare materials which don't rely exclusively on RDA. Participants at LC worked with their triple store, which was continuously update as cataloging from outside the Pilot was converted for inclusion.
- **NEWS: The Library is pursuing BF as a viable alternative to MARC.**
- Creating and Updating a BF Database Jodi Williamson Loc.gov/BIBFRAME
- 17+ million converted records, 1.2 million uniform title authority records, records added every day from ongoing MARC cataloging.
- Showed editor tool, includes a left-anchored search
- Match and Merge process looks at 130/240 titles indexed as nametitle, also 1XX+245 for title matching. Merges work infor from "bib records" and creates instance based on other data in the records. Some types of works, like music, have special needs, where works should be related to the larger work, but not necessarily expressions of it. Music got a shout out for all our good name/title records created, making merging easier. Not all work records link to LC resources have instance records since they don't have everything. More records will be matched and merged. Name/title works pretty well until you encounter things works titled "Untitled."

- Editor currently includes work with various catalogers at LC, sound recordings and rare materials, e.g.
- Showed instance for a map

Things to improve:

- 7XX related works aren't dealt with well. Now they're stubs but should link to related works
- Contributors from 7XX tags are sometimes lost in conversion
- Load sequence and system control numbers impact conversion; revisions assign new numbers and links can get broken
- They still need to develop more profiles
- Need to develop a base BSR description that can be expanded
- Need to be able to easily clone a description
- Need to accept different serialization schemes
- Working to see how Casailini's data could be loaded
- Ingesting CIP and ONIX data
- LC's BF file is open for exploration
- Keep developing editor

From MARC to BIBFRAME in the SHARE-VDE project (Tiziana Possemato, Casalini Libri) Background on SHARE-VDE, project to move LAM data into linkable data, including MARC to RDF using BF model. Includes editor/interface. A collaboration with the library community.

- Enriching MARC with URIs
- Conversion
- Data publication in BF
- Batch data updating and dissemination
- Development of use cases from the user community
- Workflow:
    - MARC conversion relies on a Similarity Score, and then "Authify" to identify each entity and enrich with external data. Creates data clusters. Starts with authority records feeds to POSTGRES dB which is processed goes to CLUSTERS KNOWLEDGE BASE.
    - Bib records go through similar process for enrichment.
    - KB goes through "Lodify" processing then into triplesore with different output.
    - Process tries to extract different entities in each record. Reconciliation works with automated or manual processes. The former allows for large datasets. The result is a "cluster." Showed Vivaldi examples as a result of going through the process.
    - The Lodify process parses MARC, so e.g. a 300 field can go into BF properties (e.g. "dimensions")
- Phase 2 deliverables:
    - Datasets in BF 2.0 format, enriched with "tuples" derived from MARC Clusters in RDF Datasets in BF 2.0 URIs table for external sources
    - Followed LC's work to see what they need. Includes LC BF extensions.
    - Lodify is the framework that automates the conversion and publication of bib and authority data in RDF according to BF 2. LODIFY maturity level test shows good results.
    - Part of the goal: "To create a new ecosystem."

Using BIBFRAME in multi-institutional projects (Jeremy Nelson, Colorado College)

- A Colorado Wyoming DPLA service hub, includes working with various cultural heritage organizations including Denver Public, Colorado College, University of Wyoming, History Colorado, American Heritage Museum, others.
- Input formats in CSV, MODS XML, JSON (not according to a standard). RML (RDF Mapping Language) was used to go from these sources into RDF. They simplified the BF 2.0 model for the mapping work. RML is RDF, RDF to create RDF. They had to make custom mapping's for custom data models, but MODS XML was more standards compliant and reusable. Uses XPATH to map XML to BF RDF.
- Output front the triple store can be exported as JSON in the DPLA data model.
- Above is from the BIBCAT open source project. Used, at first, Blazegraph RDF Triplestore but performance was an issue. Later used RDFFramework, RDF to Elasticsearch with RML mappings.
- Planestopeak.org service hub, using ResourceSync, a replacement for OAI PMH.

OCLC research with BIBFRAME (Nathan Putnam, OCLC)

- What OCLC has been doing with the BF 2.0 converter.
- They reviewed the style sheet to see what would need to be modified for the OCLC business model. Converted 11 million to MARC XML and ran them through the BF converter.
- Findings:
    - Work IDs, cluster IDs helped
    - URIs are very important
    - Edited to deal with gaps re instance records/info

- They Populated 758 with Work/cluster IDs Looked for $0 and $1 Preferred URIs for VIAF and FAST instead of LCSH.

QUESTIONS:

- What will happen with 880s? 880s are paired with the Romanized fields. Instances are in non-Roman.
- LC will do both mapping and the converter tool. LC needs it and there's public interest. LC hopes they'll be out very soon.
- Please share slides, video would be nice.
- How will RDA entities be expressed in BF? Issues with difficulties in data modeling. In BF, e.g., bf:hasExpression can help clarify RDA work relationships.
- Item information for BF is currently being converted. It's an area where item extension vocabularies for vendors might be necessary.
- Are the enchanted Casailini records being exported and distributed? The deliverables were to go back to the partner library. One project shared broadly, one with just the contributing library.
- How did LC choose which records were part of their work sample? Records with LCCNs only.

CC:DA
6/25/18

Special session on RDA and MARC

(Notes available at: http://www.gordondunsire.com/presentations.htm)

The meeting began with three presentations to expose some of the issues related to moving the new RDA model into MARC

RDA and Data Encoding: The impact of the 3R Project (Gordon Dunsire)

- RDA doesn't talk about MARC though it does serialization in RDF
- Overview of 3R project genesis, and of current state of Toolkit in Beta (including development of entities etc. almost complete).
- New entities for agent, collective agent, nomen, place, timespan.
- Some other changes, e.g.: Attributes become relationships, e.g. date of birth > Related timespan of person; more inverse relationships, e.g. Date of birth of [related person of timespan]
- 13 entities, 1700+ elements. Balance changes include more elements for work than for manifestation previously.
- ISBD review will commence soon.
- RDA entities:
    - No change for Work, Expression, Manifestation, Place Item always exemplifies a manifestation.
    - RDA Corporate Body and Family are types of Collective Agent.
    - Person restricted to human beings, does not cover personae, non-human personages.

- Make explicit kinds of recorded data
- CLOSED WORLD ARCHITECTURES:
    - Flat file
    - Big/authority modeling split
    - Relational DB model
- GLOBAL, OPEN-WORLD ARCHITECTURE:
    - RDF
- Recording methods (formerly, 4-fold path) Unstructured description can only be mined for keywords Structured description (includes e.g., structured notes, controlled vocabularies) Identifier, local numbers, ISBDs, a code for a concept taken from a controlled vocabulary, as examples. You need to know the domain, the context to make sense of these.
- IRI/URI: globally unique identifiers. LRM says IRIs can be used as an identifier, RDA not.
- Data provenance is important. "Metametadata."
- RDA in RDF, RDF Resource Description Framework

MARC 21 in RDA Toolkit
James Hennelly

- MARC mapping RDA Registry RDF to output is currently very awkward; they've developed a CSV file that you can download from the site
- Inputs: tag, bib/authority, used to map RDA entities to MARC mapping
- Currently no RDF mapping for MARC 21.
- Approximately: RDF element>hasMARCmapping>MARC 21 element
- Currently there is a search that lets you go from RDA element to MARC.

RDA in MARC 21: Accommodating 3R
Thurston Young

- Context for change: 3R Project timeline, new RDA entities, new guidance chapters Context for change in MARC 21: existing, new MARC21 content designation; non-MARC data formats
- Basic principles:
    - Choice of MARC 21 format, separate bib authority, other Granularity Consistency?
    - Utility
    - Feasibility

- New RDA entities
    - RDA Entity: No place in MARC to capture current RDA definition. Probably no need to capture in MARC since this conceptan abstraction
    - Nomen: No content designation in MARC now. Split authorities into two records? Names plus attributes of entity? Big undertaking Now in the authority format X51, 370, 371, many fields with $z; location of meeting in X11
    - Place: Already fields in place in both authority and bib formats

- New RDA guidance chapters
- Recording methods: For structured descriptions: Focused on manifestation statement, MARC 21 is too granular, breaks up RDA element to pieces in 245 / series into subfields in 490 / No exact equivalence / How to fix? New indicators to existing fields or new subfields added to above OR new field
- Representative Expressions: There are many elements of a representative expression, such as aspect ratio or medium of performance. A representative expression record would look similar to expressions, but there are no ways to explicitly state that something is a representative expression / some of the elements though move up to the work level by virtue of being in the representative expression / New indicators or subfields could help, though designating the record as for the representative expression in the 075 / completely new tag. Current bib format has places for things like aspect ration, color content, duration, script, sound content of item, and these could move into the authority format.
- Data provenance: things might be of interesting for provenance: agent who publishes metadata, transcription standard used, content standard used etc. Some of this in the authority and bib formats. Source for authorities are in 670 / missing elements for validity of metadata, transcription standard for metadata. Greater scope for content designation may be available in the authority format.
- Diachronic works: Has a core concept of the PLAN, successive determinate or indeterminate; integrating determinate/indeterminate plans; static plan. Nothing in MARC21 for diachronic works
- Timespan: accommodated in various places in the bib format, no current method for the timespan element to be treated as a clump that can be pointed to outside of something like the subject vocabularies
- Options for incorporating 3R changes are limited / BF, Schema.org could be extended to work, but there are no definite signs of development in response to LRM / 1-2 full MAC cycles to amend MARC ; Toolkit is still in Beta ; Could RDA/MARC Working Group reconvene? New PCC task group? Other options?

Thoughts on the above? Invitation to comments from the room, CC:DA and the room.

- Is it worth developing MARC more to accommodate RDA? But RDF is also difficult for some things as distribution and provenance change with each iteration. Hybrid records now are problematic as far as metadata provenance, with mixed content standards. Should we freeze where we are now? There are many half-solutions to in asking MARC to accommodate graph-like structures.
- Bib/Authority bifurcation is problematic. Bib record as stand-in for RDF graph...
- After all the comments no real consensus.

Would the room be interested in reconvening the RDA in MARC Working Group? Or should we go somewhere else more forward-looking?

- If a group is formed it should include practitioners, as well as some vendor input. No immediate plan for directions.
- Timescale of changes? MAC's process is deliberative, and a 1-2 cycle time plan is optimistic. Might MAC have a different cycle? And NACO node changes can take longer.

Subject Access Committee
6/25/18

Supporting Digital Humanities and LAM Data Access through Semantic Enrichment (Marcia L. Zheng, School of Information, Kent State University)

- Why digital humanities? It's the intersection between the humanities and computing disciplines. Librarians in one study argued that 6/10 in a survey thought the work belongs in a library. At a recent meeting, a high proportion had librarians as presenters. Http://dev.diggingintodata.org is a program with much grant support, and has some interests in metadata work.
- Data doesn't necessarily map into only digital data (think: tangible materials.)
- Data on the Web Best Practices (W3C 2017) encourage providing metadata.
- LAM world has many examples of structured data in things like archival finding aids, catalogs, curated research datasets. Often there will be unstructured portions of structured data (archival documentation not in EAD, TEI files...) Unstructured data is what we have the most of, and comes in the most varied forms, and are the most difficult to process.
- Data from LAMS and cultural heritage institutions are a rich source for humanities researchers.
- Trends in DH-LAM world: bigger data, more-structured data, machine-actionable data.
- Three perspectives on creating structured data:
    - Documentation/metadata: Descriptive / administrative / structural/technical metadata
    - Metadata: Descriptive metadata, indexing, markup, ontology Use Metadata, when I something cited researched tagged...
    - An aside on the importance of provenance information: The famous Kent State Massacre photo has been doctored to remove the pipe coming out of the head of the central figure. Going back through states of the image could help a user glean what changes had taken place, when.
- Semantic enrichment process
    - Analysis: pre-enrichment phase
    - Linking
    - Augmentation, selecting values from the contextual resource to enhance (See Europena Task Force on Enrichment and Evaluation report 29/10/2015)

- POSSIBLE APPROACHES
    - An enrichment task: a process that improves metadata bout an object by adding new statements “Enrichment” can mean several things: Alignment (Contextualize, create typed relationships between resources of different types, usually in controlled form, align to sources like DBpedit, GEMET... ; “massage” the metadata, align with existing vocabularies to harvest external data, “sameAs” can lead to harvested content ; connect to “real thing,” e.g. Weissbib project, which links to external resources as part of the project, added schema:sameAs, foam:focus, mapFAST project that can include a map.)
    - “Expansion” has starting point, existing metadata components that are unstructured. Showed an example of a hybrid record with expansions of the original. Getting structured data from less structured parts of a description. Used OPEN CALAIS in one example to paste semi-structured text into the tool to extract entities in the text, entity extraction from EADs (persons, corporates, geographic, events are pretty accurate; social tags, industry terms and products categories are not trustworthy). THese tools can be good for description and identification, but not so good for analysis of content (e.g. for subjects). There are taxonomies and ontologies behind the tool. New structured data from unstructured data, one case looked at ETDs. Focus on abstracts, titles, keywords and introductory paragraphs. Abstracts work better on titles, more useful for fields like music, where titles can be more fanciful. Major concepts were correct in most cases.
    - Ontology design for non-structured data: Linked Jazz project used oral history transcripts from 5 archives. Analyzed names mentioned to develop relationships between entities, e.g. mentors.
    - Look at wiki.numismatics.org/numishare:visualize.
    - How to deal with images: deep image/semantic analysis; look at annotations; semantic annotatioin goes deeper, enriches the unstructured or semi-structured tools.
    - Tool: Mix’n’Match, lists entries of some external databases over 1000 catalogs and allows users to match against Wikidata items.
    - Exposing our work: how do we get users to our sites?
    - DPLA and places like WorldCat using Schema.org can extract much information. These could be ways to get around bottlenecks in capacities to provide fuller subject access.
- In the end: If a vocab has embraced linked data we can extract much.