Zephir Advisory Group
July 6, 2016

Present:

Kathryn Stine
John Mark Ockerbloom
Gary Charbonneau
Sandra McIntyre
Tim Cole
Jon Rothman
Naun Chew (Recorder)

Absent:

Todd Grappone
Ryan Rotter

Agenda:

1. **Welcome Sandra McIntyre, HathiTrust Director of Services and Operations, and Introductions**

Sandra will be joining future calls. She is a member of ZAG as well as MUSAG.

2. **Zephir Update**

- HathiTrust Camp
    - In early June, day-long camp in Ann Arbor, MI. Meeting with new staff including Sandra, Heather Christenson (US Federal Documents program), and Lizanne Payne (Shared Print program).
    - 90 minute session on Zephir, and 90 minutes on HTRC
    - Zephir session focused on:
        - Where Zephir fits into the HT metadata ecosystem and where it fits with new HT initiatives
        - History and role of CDL's/UC's involvement in HathiTrust
    - Brainstorming and networking outside of regular meetings; hope to have similar meetings again in next year or so.
    - Better appreciation of what individual libraries have contributed, and what teams like this one have done. Moving forward, hope to have better integration of what Zephir team, HTRC, UMich library do.

- Metadata Survey
    - Reaching out to contributors. Working on draft of survey questions to share with metadata contacts. Goal of improving submission process. E.g. opportunity of

web-based interface for getting metadata in and out. But waiting for input from contributors before taking major steps. Also feedback on reports. Sharing questions this week with UX and huddle group. Will inquire into: who are metadata contacts, submission (including troubleshooting), how people are using documentation, information about local workflows including automation, managing multiple metadata streams. Interested in learning from experiences with other aggregators, e.g. DPLA, OCLC. Hope to launch survey by end of month and do analysis at end of August. May pursue focus groups for deeper dive.

- Collections Committee Duplicate Request
    - Request from Collections Committee about duplicate copies. Revisiting a 2012 report. At the time the hope had been to do this annually. Caveats with the data remain: hard to get at volume duplication with serials and multi-volume works. Focussing on single-volume monograph duplications. Also looking at data points to see if suspected duplicates are in fact duplicates. Providing links to HathiTrust copies. Estimated 1.78M volumes of single-volume monographs. Working to tailor queries. More analysis than development work. Also working on improving duplicate detection as an outcome of this process.
- Record Tracking (Metadata Provenance)
    - Data modelling being done on this. Looking to improve reporting out of Zephir. Possibly building data warehouse on top of bib metadata. Working with U of M on common needs.
- Submissions Administrative Data Modeling
    - Have been working on sharing common administrative metadata values (e.g., identifiers) between Zephir and HT/UM staff that factor into both metadata and content submissions. We'll be coordinating as well on drafting a shared data model and data dictionary to further confirm common ground in managing submissions.

3. MUSAG Update

Tim shared that:

- MUSAG met end of May.
- Environmental scan first draft shared in May; will be revised this summer.
- Discussed concerns about metadata as well as content quality.

Kathryn talking with Zephir ops team and looking at getting Zephir and MUSAG together. She and Todd will coordinate around how/when it makes sense for the Zephir Ops team to share operational perspective on policy discussions with MUSAG, and how/when/where MUSAG and ZAG business overlaps.

4. Confirm the Zephir Request Procedures DRAFT document

Under 2f of this document: discussion of Tim's suggestion about how to assess the benefit of a request. It may not be necessary to expect the requester to anticipate all the benefits of a request. Costs of meeting request should be balanced against benefits. Kathryn put some draft language into the document: "What benefits will this request have on the existing requestor? The Zephir Operations team will also consider benefits to other Zephir stakeholders as appropriate to the resources required."

Suggestion by Sandra to define entities at the start of the document. Kathryn added footnoted definitions.

Tim Cole: does not affect other HathiTrust policies regarding data access, e.g. rights of institutions to their own data. Jon Rothman: add language in 2g to cover this point.

Sandra: state explicitly that the operations team will make determinations relating to adherence to HathiTrust data governance policy in consultation with the HathiTrust repository manager.

To the criteria regarding policy (2g), Kathryn added: "Requests for data will adhere to HathiTrust policy governing data access, in consultation with the HathiTrust Repository Manager as needed."

CONFIRMATION: ZAG approved the Zephir Request Procedures

5. **Discuss findings from the MUSAG environmental scan report draft**

Document is still open for comment until the end of the month.

Naun asked if this document provides opportunity to continue considering having a "HathiTrust record".

Kathryn suggested that the report cites opportunity areas, and does not make specific recommendations on metadata management, such as that addressing our previous discussions about maintaining HathiTrust copy/version of bib data.

Tim Cole: There is a need to add additional data not represented in MARC bibliographic data, e.g. gender data in authority records.

Kathryn noted that there are opportunities for Zephir engagement with HTRC, especially in considering metadata enhancement. Tim suggested that the HTRC's work on enhancing metadata (e.g., with VIAF data) would be at a good point to share/discuss with the Zephir team in about six months or so.

Key Points (ABBREVIATED by KS)

Trends and themes noted that intersect across multiple areas:

- Enumeration/chronology is problematic.
- Duplicate detection could be improved.
- Bound-withs are a problem for discovery, for copyright review, and for duplicate detection.
- Reuse and sharing policies exist explicitly only around bibliographic metadata.
- The HathiTrust metadata environment is not well integrated because systems and applications were developed independently at different points in time, by different individuals and teams. Due to the lack of integration across systems and applications, updates to metadata go through a waterfall-type process (i.e., updates get pushed from one application to another), resulting in time lags for updates appearing, or are siloed.
- MARC metadata does not fully match user needs and expectations and does not meet the needs of applications.
- Bibliographic data is treated inconsistently across applications.
- Multiple applications may benefit from integrating with linked data initiatives that make available implement widely accepted authority data and controlled vocabularies (e.g., VIAF, LCNAF, FAST, LCSH, etc.).

Easy Wins

- Acquire data from OCLC about merged and duplicate OCLC numbers. This could be used in the Zephir and Federal Documents Registry clustering processes and in the Print Holdings Database.
- Develop consistent policies for sharing, updating, and enhancing HathiTrust metadata.
- Develop policies and ways forward for integrating with linked data initiatives.

6. **Kickoff metadata watch group**

This item was deferred to the next meeting. Kathryn asked members to review the watchlist template, by 1) making suggestions to the format of this tracking tool and 2) add/refining issues that would be useful to track for their potential impact on or significance to HathiTrust bib metadata management:

Next call: July 20th (Kathryn will reconfirm schedule)

Zephir Advisory Group
July 20, 2016

Present:

Kathryn Stine (Recorder)
John Mark Ockerbloom
Gary Charbonneau
Sandra McIntyre
Tim Cole
Naun Chew

Absent:

Todd Grappone
Jon Rothman

Agenda:

1. **Zephir Updates**

Zephir 2016 Q2 load counts

shared on the ZAG Google site

Zephir 2016 Q3 roadmap

forthcoming

Collections Committee duplicate reporting

KS: We've been working with the PSC Collections Committee to provide reporting from Zephir that documents overlap in the corpus. At the title/manifestation level, we detect duplication based on OCLC number and local bib system number.

JMO: Duplicate detection is the clustering that you're doing. Feedback: has seen over-clustering. (e.g., same year, diff publisher, or pub place).

KS: Thanks for this feedback. The Zephir team has been considering how to improve clustering in the system, and in particular how we might do this given access to current OCLC number relationship data, either through a concordance table and/or by dynamically checking OCLC data sources to confirm OCLC number relationships and getting that into our processing.

TC: With any cluster changes, look out for losing information that could be useful for end-users. Would be a shame to throw away a link between the records/titles/manifestations. HTRC users need to deal sometimes with ~20 results, but need to wade through all of them because relationships between titles may be legitimate and useful. If Zephir does engage in changes to

duplicate detection/cluster changes, could we retain links or breadcrumbs that may have indicated relationships between different manifestations?

NNC: How to distinguish between deprecated-current master number relationships and erroneous relationships? OCLC runs de-duplication on their records and maintain both manifestation ids and work ids. A while ago Cornell got a table of work IDs against manifestation IDs from OCLC (this was a one time occurrence and unusual) - clustering related editions in their catalog. Data is not squeaky clean, and they sometimes encountered strange results. Think that it's been of benefit to their users. Want to have a better mechanism to feed corrections into OCLC.

JMO: Some interest in reconciling OCLC numbers - getting a full copy and evaluating hasn't happened yet, but has engaged in using some homegrown relators (in Online Books), the HT "related" on the sidebar is somewhat useful as well.

KS: Another active project right now for the Zephir team is launching our plan to learn from metadata contributors about their submissions experiences. We have a data gathering plan in the works and are honing in on questions and potential respondents.

2. **Continued discussion of findings from the MUSAG environmental scan report draft**

- Addressing enumeration/chronology parsing

NNC: Aware of work to see how feasible it would be to normalize this data - item-level metadata, has anyone attempted to normalize?

TC: When multiple copies are in the UIUC catalog, even for a given title, there are variations in how holdings statements have been recorded over time, also binding issues.

NNC: Comparing across libraries - can assess that there are many variations. Are there folks in the serials cataloging community to connect with on this? Perhaps projects that digitized journals for back-runs have dealt with these parsing issues?

TC: CIC has a "last copy" print archive project which involved work on who had which parts of the run. It's been quite manual. Can inquire about the shared repository's approach on this.

JMO: Are there canonical sources to match against?

KS: These are great ideas. We aren't quite at a point to dig in on serials (and multi-volume) holdings parsing, but are gathering leads and have some other projects at CDL/UC that will be tackling holdings parsing (both at the summary holdings and volume statements levels).

- Improving duplicate detection

See earlier discussion re OCLC number relationships.

- Establishing policy that addresses metadata maintenance: enhancement, updating, correction

Did not actively discuss.

3. **Kick off the metadata watch group (please review/provide feedback on this watch list template)**

Several people self-nominated to track particular issues. Please review the template and claim areas with which you are familiar or have an interest in tracking. We'll continue to shape our tracking approach at an upcoming ZAG meeting.