Zephir Advisory Group
November 18 2015

Present:

Todd Grappone
Kathryn Stine
John Rothman
Ryan Rotter
Timothy Cole
Chew Chiat Naum
John Mark Ockerbloom (recorder)

Absent:
Gary Charbonneau

Notes:

1. Zephir Update

Zephir team plans centralizing access to contributor setup data (administrative metadata elements like identifiers for configuring processing content streams, and identifying digitization sources, etc.). They want a central registry of identifiers on a Hathi website so that everyone can see them and inconsistent use of these is avoided.

Also planning on streamlining submissions processes (in the new year). For example, looking into not depending on the ftps protocol, which is cumbersome for some to use.

AWS migration: They are now running a test instance of the Zephir processes and the database on AWS (includes ingest, XSL transforms, etc.), in parallel with current production environment. They're now debugging some inconsistencies between AWS test and current production instances.

Also very close to working with bib corrections group of Hathi user support to see how they're using Zephir data and how they can better meet their workflow needs.

They have a couple of auxiliary services that may support their workflows, e.g. update trackers to flag changed records that are on a watchlist. Another service: an "item-level" API that serves the MARCXML records from each of the contributors for a bibliographic entry. This may be useful for reference and review, for determining which records may need to be changed, and to verify that appropriate changes are made.

The current HathiTrust user API represent records at bib level, which may represent volumes from multiple institutions; the Zephir “item-level” API lets you see multiple records from a given “cluster”, not just the one algorithmically chosen as the "best" record from a cluster of records each representing duplicate items.

Proposal: Open this API to bib-corrections group without having any technical restrictions on the API, like login. They don't see a lot of risk. (They could get hit with a volume surge, but they don't see this as a likely user problem.) This API isn't intended for wide access, or to be distributed as such. At this point, it's a read-only API with targeted applications for internal HathiTrust work. They still need to discuss this with the bib-corrections group, but some early users find it useful. (It may also provide a way to verify whether the "best record" algorithm is actually picking the best records.)

Tim: Notes that most of these records are tied to OCLC records, which are tied to work identifiers. (Caution: OCLC often clusters together different editions that have different copyright statuses.) Tim hopes these work clusters could be used to pull in more LC classes (currently only 45% of records have an LC class associated with them.)

Kathryn: There's been some interest in making composite records with the best data from the various records in the cluster, but we're a ways off from going in that direction.

Tim notes that in a BIBFRAME world, the notion of a discrete record gets somewhat tenuous anyway, and it may be more natural to compile assertions from various sources. It'd be good to construct the API with that in mind, so that it can support multiple technologies.

Kathryn thinks it might be worth discussing the potential for item data API functionality in future meetings.

Todd: Be prepared to shut down certain IPs if too much access.

Ryan had a question on notifications from the update tracker service: how does that work? Kathryn: it's been dormant while we've migrated to AWS, but it can report on metadata for items that have been tagged as needing investigation, or reporting on, and notify when those records have been touched. It can send a comment back to a Jira record that notes a change has been made (though the change may not necessarily be due to a contributor-submitted correction).

Tim had question on clustering in the context on Zephir: how does the system determine clusters? Kathryn: The means of identifying perceived duplicates as they come in is to match on the contributor's local ILS system number, and also to check on an OCLC number match. They've discussed using a bit more subtlety, though clustering has implications on rights data.

Tim notes that digital humanities researchers also need suitable clustering for their analyses (e.g. they might or might not want 25 instances of Jane Austen's _Emma_. Some care about work level, edition level,manifestation level, depending on what they're researching.)

Kathryn thinks HTRC may have done some research on this matter, and other user studies, and wonders if ZAG might want to look into this. Some researchers might want access to the various-library bib records via the API discussed above, so they can tell whether they want to distinguish their copies or not.

(**ACTION:** Kathryn asks Tim to share links with list about what HTRC researchers are interested in. Tim sent us email with some links after the meeting.)

Recent Zephir data inquiries:

-- one internal inquiry from collection committee (PSC) to revisit duplicates analysis report they originally made ~3 years ago on impact of duplicates & deduplication on user experience, storage costs, etc. (2012 report: https://www.hathitrust.org/documents/hathitrust-collections-duplicates-report-201204.pdf) Kathryn thinks the queries they need should be straightforward, but wanted to share this request with the group.
-- another duplicate-detection inquiry came from a campus that wants to compare their holdings with what's in HT. This might not be something they need Zephir to analyze; they might be able to do it from currently available data sets (e.g., the hathifiles).

So for both of these, the pre-ZAG procedures criteria seem to work for inquiries.

2. Authority data – maintenance and policy issues

background: Collett/Stine code4lib 2015 talk on Zephir/VIAF data considerations:
http://code4lib.org/conference/2015/stine

First exploratory approach: Can we just pull VIAF identifiers and associate with record? Well, we'd also have to get relationship between VIAF identifier & OCLC number. One use case: getting author's death dates for clearing works for non-US access. Bigger policy issue: how would we move towards enhancing the metadata using VIAF?

Naun noted a PCC group is looking into issues of getting URIs into bib records. (What kinds of identifiers? What does reclustering do to this?) OCLC's non-library links to things like DBPedia/Wikipedia provide another way of getting death dates.
ACTION: Naun can get us some links to relevant resources.

Tim: A lot of name strings don't match authority identifiers. (In UI catalog, only 60% of unique name strings did; sometimes this was due to obscure people not in authority files; sometimes it was due to "bad" strings for people who did have authorized headings.) In some sets (like German national identifiers for Renaissance books) they found DNB identifiers more useful than LC IDs as there may be additional data elements tracked by DNB.

Q: What did you do with names you couldn't reconcile against VIAF? OCLC is apparently minting additional identifiers based on association with a work, rather than occurrence in a national authority file. There are a lot of these, potentially. There are also low match levels with institutional repository content; they're looking into ways to find or mint identifiers for these.

Kathryn: It might be worth having a subgroup look into issues that may have bearing on this group, or on HTRC. But before we go down that path: Is there an operational impetus here to address underlying policy issues (e.g., can there be a “HathiTrust” record/collection of statements that can be enhanced/corrected/edited)?

Todd reports MUSAG group is looking into forming a metadata sharing policy based on an environmental scan. It's hard to say what it will look like at this point, but their tendency is to want to make the data as open and reusable as possible. There's still some concern in community about IP issues related to data in bib records that might not just be raw facts, but require judgment. Also, they might want to formulate terms of use/expectations for scholars who want to contribute metadata (so it's clear they know what others can do with it). Jonathan also notes this is related to interest of some in having an enhanced "HathiTrust" record in addition to contributor records. (This is different from the merely "corrected" record we've also been discussing for HathiTrust.)

Kathryn asks whether the MUSAG metadata use and sharing policy work in intended to be comprehensive of addressing both what uses can be made of metadata contributed to HathiTrust and how contributed metadata might be enhanced and/or corrected. Todd replied that this MUSAG work will address both policy considerations.

3. SLA 4.3.1-4 policies and procedures

Work in progress document:
https://docs.google.com/document/d/1pJwrASi3T0pus5x7-oAHjLij0FWnfspA7VBVqavQ3Zk/edit?pli=1

Some of these issues are also of interest to MUSAG.

We'd like to time-bound this work, so we can make some draft recommendations by early next year.

One question: Is it sufficient just to refine the Pre-ZAG document? (and if so, can we go ahead and start revisions) Or do we need to start afresh?

Group seems to think Pre-ZAG is a good starting point. Though it's a fairly broad document, and maybe we should figure out what to prioritize in terms of attention. The incoming DSO (when Hathi gets it) might have some ideas about priorities.

One early question to consider: Where do requests come in, and then how and to whom do they route it?

We might be looking for something like a triage guide to whoever is designated as first responder.

Consider the previously mentioned examples of data requests, one internal, one external.

Another pertinent question: Do we need to do development to handle requests, or is process improvement enough?

ACTION: Kathryn offers to share some notes from current Zephir staff as they are developing tools to help make decisions, moving from issue identification to requirements gathering on through to development; they would be useful for us to consider.)

Jonathan notes that scope within capacity determines how much we have to prioritize requests.

Some useful distinctions for prioritizing anticipated work:

- hours needed
- maintenance vs enhancement
- output based on what exists vs reports which require functional or structural changes

One example: introducing identifiers (VIAF etc.) would be important part of creating a "HathiTrust" record but would clearly require a go-ahead for making enhancements.

Next steps:

1. **ACTION:** We can all go back and annotate pre-ZAG procedures document based on our later notes. Look especially at procedures table in the work in progress document to identify where we may.

2. We're a bit in limbo without a DSO in place (we hope to have one early in the year) but it could be a good goal to have a revision of the pre-ZAG procedures document for the team to review, even before a DSO is in place, and identify questions/clarifications to share with the DSO.

3. **ACTION:** Kathryn will consult with the Zephir Operations team, highlight parts that could use early attention and will share this with the ZAG team.

4. **ACTION:** Kathryn will keep the ZAG apprised of developments in the Zephir Operations team’s engaging with the corrections team (scheduling in progress), as they're a potential key client of the Zephir metadata services.

GOAL: Have key parts of pre-ZAG procedures document needing attention identified by next meeting, December 16, with comments and draft language proposed.