SAB discussion 5/17/12

Analytics for the Hathitrust:

Prepared by Todd Grappone/UCLA

Problem Statement: Every collaborative library project needs analysis if we understand what we have and what others would like us to share. The OCLC analysis tool, though good is incomplete. The ability for OCLC to serve as a robust collection analysis tool is limited, it's all a one-off. How can the HT grow and bring in new members if we can't do sophisticated content analysis? We should look at the current state of tools (OCLC, PAPR?) and perform a gap analysis based on our needs. How do we compare HT content to analog complementary information (books linked to journals linked to newspapers etc).

Questions for SAB:

1. Is there interest in HT creating an analytics program similar (perhaps in competition with) OCLC?
2. Is there interest in creating a service that would link HT content with similar (e.g. by subject area) to a knowledgebase like PAPR?
3. What would this tool look like beyond subject and metadata aggregation?
4. What are the questions that Collections Council can't answer?
5. What books in the public domain are not yet in the HT?
6. Within the UC or compared to other institutions where are our collection strengths and how does that relate to the published record?

Call Notes from April 13, 2012

Call with York and Wolven: Conclusion is to take the problem statement to the Collections Committee.

W: Who is doing similar and how are they doing it? OCLC is frustrating.

JY: Is this something only some partners would pay for (in response to Todd's introduction)?

W: The Collections Committee has wrestled with waiting to ID holdings within current membership which meets gap needs, Public Domain and not yet digitized. Working with currently supplied membership. Possible as a recruiting device? Need connections with OCLC.

Y: Current plans include providing cost information (Holding plot) graph of how many other partners hold given partners holdings. Report of number of records submitted in different types (single-part monographs, multipart monographs, OCLC overlap to HT holdings). Have had requests from partners for additional information, e.g., how many volumes held by my institution are not held by others. We did not fulfill this request at the time because of the additional

resources needed. When partners approach HT they provide holdings records to receive a fee estimate for 2013.

W: We can only compare to data you have. HT would need the ability to submit a data set to compare.

Y: Zephir a bibliographic management system for HT. Will manage all submitted records from institutions. HT has all the records. Not sure what the scope is beyond existing needs. At CDL it's Lynne Cameron, Kathryn Stine, Stephanie Collett, Michael Thwaites, Lena Zentali. Hathitrust.org/htmms. What about higher level connections?

W: We have the same problem at Columbia. Using backlight. Well not sure what HT brings to the table that gives us a leg up or aligns.

Y: Specific value that HT has with print holdings data in relation to the print and the digital.

W: DPLA is moving in this direction. Idea of aggregating discovery across different data.

Y: talks and discussion around aggregating the data in some kind of fashion (OAI and then discovery).

W: Challenging is the overlap with commercial databases. We need a tool that offers quick and flexible analysis with print collections. Comparing HT to digital offerings. Not sure commercial vendors would want to play.

Y: Discussed the keepers registry. What that lacks is we are working with OCLC for a multiple holdings numbers. Anything compelling;

Y: greatest possibility is what does it allow you to do with decisions points around digitization and collections. Lay out the possibilities that analyzing print holdings could provide. What are the current effort and how are libraries trying to resolve these problems? What does HT provide and what information and computation on the print holdings and do a gap. Thoughts about what a proposal around analytics might include (restatement of some of paragraph above I had written some thoughts out prior to the call and read them quickly):

.

Possibilities that analyzing information about print holdings affords in general, in relation to large goals libraries have

.

Current efforts of libraries

What the efforts are trying to achieve

.

What services are currently available (such as OCLC)

How they are used

What shortcomings there might be

.

Discussion of what HathiTrust is currently using print holdings information for and what its goals are for using this information

This would include what we are providing by default to partners

What kinds of additional information might be provided and how it would be used

Some information on the resources required (after doing some investigation of the kinds of queries that might be desired)

Who would do the work (make sure no one assumes we have staffing for this)

Next steps:

W: Asking collections committee to flesh out the things we¹ve been talking about.
Seeing the service committee to flesh out the case as well for value. See a development proposal.

Y: at some point in talking about strategic ideas then it would go to development initiatives group (Constitutional Convention proposal). reiterating the value that the analytics would be to partners, particularly when institutions begin making preservation commitments on the print volumes. The GLMRS (sp?) was probably about the fact that with data we are getting from OCLC currently (all OCLC numbers for a given manifestation) we are able to do very robust matching. This may increase or become easier to do if the promise of OCLC GLMRS (a single identifier for these manifestations) is realized.

JY Notes:

Todd: Initiated conversation a couple of months ago, wrestling with problems info analytics regarding UCLA collections, questions we had from HathiTrust, West, ARL about our collection; we were providing same information to all these places; might be nice to have more robust analytics structure in place; had been talking a lot about governance, not so much on strategy in SAB; thought throw out as
something we might pursue;

Had been discussion about services HathiTrust might provide in addition to partner services; paid services; whether analytics might be higher-level service for partners as well

Bob:
Can see variety of angles - who else is doing it, what we might do with HathITrust; several of us trying to get data out of HathiTrust, trouble doing it; one of the factors I would want to think about before advising HathiTrust to go into it; what kinds of challenges we taking on here; 1 1/2 years back when talking about poss of shared print programs, data, analysis; at that point was thinking this is an area HathiTrust could do;

Something HathiTrust do that doesn't do everything

Todd: Also thinking of identifying future partners; with OCLC, find there services pretty good; depends on how you engage with them. One-off; don't know how OCLC;

JY: Something only some partners would pay for?

Todd: That is what was in the discussion.

Bob: Think about what HT do already; thoughts that come to me at different times: Collections Committee has wrestled with possibility of wanting to identify which holdings within membership are not digitized but potentially in the public domain or meeting other gap needs; what holdings we have to contribute that might be targets for digitization; that would be in data that should be supplied by members; possible to use as not so much a recruiting device (need WorldCat database to know who has content you'd lke), but if I were thinking of participating in shared print collection; what HT have that I could divest;

JY: Talked about what we plan to provide as a matter or course

Todd: We've had a hard time working with small college library in LA, determining what we hold and comparing to what they hold

Bob: What makes it difficult?

Todd: Matching our records and theirs at appropriate data points so as low a level of de-duplication as possible after we bring it in; trouble ensuring we match as close as possible what they hold and we hold. Mystifying a little bit why that's been so hard; been working with OCLC research; different cataloging standards? Just been difficult; took a number of months, our programmers and OCLC researchers;

Bob: those types of things are the inherent difficulties anybody, including HT would have 1) can only compare to data that you have 2) what would need from HT is ability for someone to submit holdings and compare; depends on what's in the data; doing that comparison works fine when standard data points like OCLC; gets pretty sophisticated after that;

Todd: Come up a few times when talking about deduplication; is there some way to enrich the metadata that makes de-duplication easier on our side; by-product on our side might be analytics process on our side;

Bob: know there's work going on CDL, JY?

JY: Said what know

Todd: Couple other thoughts, then brainstorm
1) Connecting HT content with non-monograph digital library content; if i'm developing a project, interested in both monographs, journals, ephemera; some mechanism HT might Look into other large digital collections and come back with concept or subject map that might be useful to researchers and libraries; At UCLA, have large archival collection of broadcast

news; large collection of LA news photographs; number of small journal publications that relate to certain ethnic and subject areas from local region; ability to provide a search and subject browse service that connects these together; beneficial to faculty members; everything that we can provide access to in one search space;

Bob: Not just UCLA issue; we've been working on solving for Columbia by building Blacklight interface on top of it; do think hit the right problem space; visualizations of this

JY: How relate to holdings as opposed to metadata?

Todd: was thinking what can HT do that others don't; trying to find some other killer app that could build from that resource that would be impactful on campus to scholars; This came to mind based on records and digitized content; OCLC not have digitized content; my catalog doesn't have, but we have other collections; one of the problems trying to solve; if have initial pusher, might move things along;

Bob: not sure what HT brings to table that gives them a leg up; still need to get the data; index data, need data farm, public interface; different space from HT;

Todd: If not HT, is ARL going to do it; if not ARL, who is going to?

JY: HT has relationship between print and digital; where that fall?

Todd: thinking both; all disparate formats together in one area

Bob: heard only rumors; DPLA moving in this direction too; Idea is aggregating discovery across digital; anyone know more?

JY: Just that

Bob: Also overlap with commercial databases; love to be able to do quick and flexible analysis of when large e-book becomes available, parse that in different ways; someone could make value out of comparing HT with commercial;

JY: have had questions about that

Todd: 2) Data visualization studio; look at holding strengths for libraries around the country

Ideas?

Bob: Next steps I could see; next steps of anything uncertain with new board coming in. in current structures could see SAB asking collections committee to flesh out things We have been talking about. Make case from focused HT collections angle; could see services committee taking it up and making a case for things like to see; can see it coming forward as development initiative proposal (under that umbrella);

Todd: see that happening at the same time? each sequentially? Which the one to explore? Don't see it as consecutive; seems like Collections Committee is good place to start;

Bob: In past year, a lot of questions have centered around Collections Committee. They have circled around some of these questions already. Idea would not be foreign;

Todd: Like to type up notes, send to SAB;