

Metadata Policy, Strategy, Use and Sharing Advisory Group (MUSAG)  
December 11, 2015

Present: Tim Cole  
Kristina Eden (recorder)  
Todd Grappone  
Valerie Glenn  
Martin Kurth  
Patti Martin  
Shana McDanold  
Angelina Zaytsev

Absent: Steven Folsom

### Agenda

1. HT Member Meeting update (Todd)
2. Environmental Scan WG update
3. Policy Development WG update
4. Timing of our January call
5. Late agenda additions

### Action Items

Angelina - contact UM Tech Services regarding their data on out of print book searches

Angelina - get Zephir metrics for choosing the best record

Todd - ask the ZAG group how solid they feel the Zephir metric is; do they feel there are areas where it could be improved

Martin - arrange a conference line for our next call on January 22 at 9PT/10CT/11ET

### HathiTrust Member Meeting update

Todd gave an update on MUSAG at the Program Steering Committee (PSC) meeting, including the work of the environmental scan and policy development subgroups. They discussed questions about ownership of metadata and bibliographic records in HathiTrust. PSC and Mike Furlough are interested for MUSAG to work on use case scenarios for the metadata.

Todd reported that PSC discussed the following items:

- Collection survey results - particularly in developing out new formats beyond print. This point is still under debate. Angelina asked whether the discussion included accepting other forms of metadata. She periodically gets questions from partners about accepting Dublin Core. Answer: other forms of metadata have not been part of that discussion.
- Quality within the collection - mostly about scan completeness and error rate, some about metadata error corrections. Regarding scan completeness, Google decided not to digitize foldouts such as maps and there are an estimated 500,000 items where additional scanning

could improve completeness. The conversation on metadata error corrections was a hot topic. What should the role of members or HathiTrust be in making corrections to bibliographic metadata - the bylaws aren't as helpful in clarifying this. PSC is looking to MUSAG to start moving this conversation forward.

For more information from the general Membership Meeting - see notes linked from the agenda <https://www.hathitrust.org/hathitrust-2015-member-meeting>. Of note regarding metadata, the question arose of what the membership position should be regarding metadata enhancement or sharing, and what would the impact be on programs and services.

### **Environmental Scan Working Group update**

The environmental scan group finalized a template and distributed it to most of the identified services. The documents are being collected in a Google Drive folder.

[https://drive.google.com/drive/folders/0B\\_2SAs8abBpGdHg3NzY2akw5ZGs](https://drive.google.com/drive/folders/0B_2SAs8abBpGdHg3NzY2akw5ZGs)

Many are partially filled out already. We've asked for people to meet a deadline of January 6, however some are likely to require more time. Miao Chen is coordinating who at HTRC will respond to the surveys. She is waiting upon a decision from the PIs. Tim will prompt them to see if we can get an answer.

Angelina raised the idea of data HathiTrust has which isn't managed in any systems we're surveying. She is aware that the ad hoc quality group is investigating quality visibility, and feedback. Some of this data is stored in JIRA from user feedback. Also, some partners store data about quality from their own scanning activities. Google also stores quality metadata. Another kind of data not stored currently in HT systems is data that is the result of U-M Tech Services investigations of market availability of certain titles. Their data may only be available in a spreadsheet held locally, and other institutions are interested in having it. Angelina will contact U-M Tech Services. (HT documentation is here: <https://www.hathitrust.org/out-of-print-brittle>)

External communication - We voiced concerns about communications and products of working groups going outside of MUSAG without due feedback and knowledge of group members. Martin summarized that when a working group communication goes beyond the group we should give MUSAG a heads up about it. When we do output, if it reflects the committee's thinking about something, we should talk about it as a group before sharing publicly.

### **Policy Development Working Group update**

Tim, Todd, Patti met and shared a document that had been previously prepared for CDL by an intern to look at what should be contained in a metadata policy. It looked at DPLA and others institutions similar to HathiTrust. It also discussed the dispersed nature of HathiTrust information about metadata and corrections; that it should be brought together into one place for people to understand. This document has been posted to the MUSAG group site. The policy group will use this as a template to address language of a proposed policy and will next look for gaps in the document and places where it is off base. They will meet again in January.

The group reported that they had discussed what constitutes metadata, ownership, rights and responsibilities of participating members, and what can we do as HathiTrust and what needs to be done by members. In assessing existing models for metadata policy, the group is leaning towards those that say metadata should be freely available for reuse. It will be interesting to see what the environment scan

turns up in relation to this. DPLA uses a CC0 license required of all contributors. OCLC has a non-Creative Commons license on WorldCat right now. This is an issue which is undergoing a lot of evolution right now. We have heard from developers that even with CC licenses, attribution is problematic for display of data coming from different sources. It becomes confusing to interface users. Individually facts are not subject to copyright. The policy group acknowledged that any final policy will have to go through PSC and the Board to assess legal implications.

The policy group also discussed whether there should be a HathiTrust bibliographic record which is different from contributed metadata. This is part of the metadata quality discussion.

A question came up about how records contributed from various sources are selected as to which is the best quality. Angelina confirmed that Zephir uses a scoring metric in their process which identifies the best record automatically. MUSAG members would like to see the metric; Angelina will get it from Kathryn. Also we'd like to ask the Zephir team if they feel this is the best algorithm or could it be better. Todd can raise this question at a ZAG meeting.

### **Timing of our January call**

ALA Midwinter causes a problem for several people to attend next call. We agreed to move the call to January 22 at the usual time. Martin will arrange the conference line with Melissa Stewart.

### **Late agenda additions**

There were no additions to the agenda.