HathiTrust Strategic Advisory Board
Meeting Notes—Phone Conference
Monday, August 24, 2009

Present:

Ed Van Gemert (chair)	University of Wisconsin, Madison
John Butler	University of Minnesota, Twin Cities
Patricia Cruse	California Digital Library
Robin Dale	University of California, Santa Cruz
R. Bruce Miller (recorder)	University of California, Merced
Sarah Pritchard	Northwestern University
Paul Soderdahl	University of Iowa
John Wilkin	HathiTrust, Executive Director
Paul Conway	University of Michigan, School of Information

Announcements—John Wilkin

- The Executive Committee has approved appointment of the Chair of the SAB as an ex officio member. This will enhance communication and coordination.
- Be sure to sign up to the RSS feed at http://www.hathitrust.org/ to receive the most current information about HathiTrust activities.

1. HathiTrust development update—John Wilkin

- Full Text search—Full text search over the entire body of material will be tested during the next few months. Basic capabilities will be implemented in October. Additional features will be added incrementally. Partners will be called upon to help with testing. A press release via CHE has been planned to be published prior to the Google settlement hearings. Sarah asked for text that could be used in campus announcements that would show how the HathiTrust search would differ from GBS, and what the scholarly value would be (e.g., more exhaustive, more precision, more sophisticated tools, etc.) This would be a good story to share with our faculty. Care should be taken to clarify copyright limitations.
- Shibboleth—The ID management folks are developing rudimentary services for authenticated individuals that can be instituted for full book downloads. At the moment this service is not being readied for special needs users due to the policy and legal issues that need to be better understood and resolved.
- Page turner software—Collaboration with UC is the first example of extended distributed development.

- Work progresses on improving the technology infrastructure. To date, there has been no loss of data or any extended outages.
- There is desire to develop embeddable and mobile page turner software.

2. Validation of Digital Objects in HathiTrust—Paul Conway

Paul has prepared a grant proposal to the Mellon Foundation requesting support for a one-year (8/09-7/10) collaborative planning project to:

1) define the relationship between the characteristics of digitized books and serial volumes deposited in HathiTrust and prospective uses of those volumes;
2) establish stakeholder consensus on these definitions and uses; and
3) prepare and submit a funding proposal to the Institute for Museum and Library Services (IMLS) to test and evaluate routines for validating the functional capabilities (uses) of deposited objects through a combination of manual inspection of statistically valid samples and machine processing of portions of the HathiTrust corpus, and then branding the trustworthiness of some significant subset of deposited volumes for particular uses.

The primary unmet needs for most preservation repositories are strategies and mechanisms to assess the quality and usefulness of deposited content in terms of a set of purposes whose fulfillment will provide additional investment incentives for (existing and new) stakeholders beyond the public good of preservation itself. HathiTrust will serve as a large-scale test bed for resolving a set of challenges that face all preservation repositories, particularly those that include digitized book and serial volumes.

Quality is not absolute and must be defined so that an end user can readily understand the suitability of a particular digital object for a given usage, i.e., to clarify expectations in an effective manner. A primary issue is that assessment of all of the volumes in the repository cannot be realistically done with an item-by-item visual assessment, but the framework should allow single-item inspection and register that. Data points needs to be found that can be used to predict the probability of a certain level of quality and usefulness. The next step will be to request funding from IMLS or NSF to put the definitions into practice in a test bed.

The SAB is very supportive of this project and wishes to be involved. We will rely on Robin to keep the SAB informed. The SAB will review and comment on definitions as they are developed and will provide support for the grant applications.

3. Organizational discussion

There is consensus for a once a year face-to-face SAB meeting to focus on the setting of strategic priorities. Possibly the first such meeting can be held in conjunction with the

Google summit in California on 21-22 October, perhaps before of after. Ed will follow up.

4. Access for print disabilities—Paul Soderdahl

One of the issues is the difficulty of determining which agency (if any) can legitimately declare that an individual has a print disability. There are no standards or central offices for the designation of disability status. Yet, good faith following of copyright laws creates an expectation for validated disability status to allow for the creation of digital access. Shibboleth technical issues come into play in authenticating and authorizing a user with a validated print disability. Ed will frame the next steps.

5. Error Rate and Ingest working group—Paul Soderdahl

The group is finishing the position paper.

6. Google summit agenda

De-duplication and error quality characterization are two potential agenda items that are of particular interest to the SAB.

7. October SAB meeting in California

See item 3 above.

8. Announcements and new business

Trisha noted that UC Berkeley will host Google Books Settlement and the Future of Information Access Conference on 28 August. Bruce and Trisha will attend.