HathiTrust Strategic Advisory Board
Conference Call notes, 11/24/09

Participating: Ed van Gemert (chair), Paul Soderdahl, Bruce Miller, John Butler, John Wilkin, Ivy Anderson (for Patricia Cruse), Sarah Pritchard (recorder)

1. **Minutes** from the 10/23/09 meeting were approved and will be posted at the web site. It was agreed that SAB members could designate a substitute to represent them at a given meeting, as long as the person has the required background or expertise. SAB discussions are not confidential, nor do they normally require a board member to be able immediately to commit a partner institution to a given recommendation, thus the presence of a substitute is not seen as a problem for normal business.

2. **Operations update** (John Wilkin): The October (and now November) updates are posted at the web site, with much more detail. John highlighted the total number of files ingested and the priority being given to UC/CDL content while CIC continues at a lower rate. By the end of 2009 Hathi may be current with all already-scanned files and will then be growing at a steady state, with about 2.5M volumes for 2010. Columbia will soon be added.
   The Mellon grant-funded project on certification is going well, and a related IMLS proposal may soon be ready for submission.
   The working group on a computational research center will have a RFP to send to partner institutions early in 2010.
   There has been extensive discussion about development work with OCLC, see item later in this agenda.
   Development work to support ingest of non-Google content will begin early in 2010; a pilot with CDL to test ingest from the Internet Archive will start in January.
   The launch of the large-scale search function for users is a major accomplishment although it has not yet been heavily used.

3. **Large scale search and what next**: The launch of the large-scale search function for users is a major accomplishment although it has not yet been heavily used. This led to a discussion of the long- and short-term functional objectives, most of which are almost completed. Accomplishments in several areas are about to be posted with progress related to: Shibboleth, fixity checking, born-digital content, nonbook/non-journal content, an API for data, and the VUFind extension to support OCLC.

   There are usability and design issues with large-scale search; should the SAB take an oversight role for this issue, and charge a group and help direct the development effort? It is related to the OCLC WorldCat Local development, there are interrelated and overlapping aspects to the two.

4. **Administrative matters for the SAB**: The next conference call will be January 14 (the December 11 call was cancelled subsequent to this meeting).
   The HT Discovery Interface Working Group has been transferred to the SAB (see #5 below).
   Validation work is proceeding; the focus initially is on Google content
   A nonpublic collaborative work/document sharing space will be set up for the SAB via GoogleDocs.

5. **Hathi Trust working group on Discovery Interface**: Working Group co-chair John Butler reported. An MOU with OCLC was signed in the fall of 2008, with work commencing in the winter of 2009. The work leading to a version 1 implementation of a HathiTrust catalog using the WorldCat Local (WCL) software (due April 2010) has played out in three phases. Phase 1 has been proceeding to determine metadata, functional, and interface for HathiTrust within WCL. Phase 2 has been the e-content synchronization of HT and OCLC records. Phase 3is interface design, including usability

testing, which is estimated to be completed in early winter 2010. There are a number of complex issues to contend with, for example "scoping" of the WCL; i.e., how to enable searching across Hathi and then across all OCLC while still retaining branding and user awareness of "where" they are.

The Discovery Interface working group and the corresponding OCLC team met in Chicago in late November 2009 to discuss respective visions for a post-version 1 implementation (the vision statement that the HathiTrust group presented at this meeting is available at: http://www.hathitrust.org/documents/hathitrust-discovery-vision.pdf). John Wilkin has sent around to the SAB his notes on the statements he made at the meeting, which are included at the end of these minutes. HT wants to ensure the pursuit of its vision and the independence of its goals; it does not just want to be folded into a WorldCat model. More open architecture and data sharing is required; and the architecture/interface need to be able to be open to searching from or interoperability with external services

6. **Error rate and Ingest working group**: An update document from Paul Soderdahl was sent around prior to the meeting and is appended here. Discussions with Google during the partners' summit in California were productive. HT is at this time still using an error-screening methodology.

7. **Semantic concepts**/Wikipedia research: Bruce Miller reported on interesting work being done by Ian Wittens of Waikato University that results in enhanced metadata and improved search results; there could be useful implications for HT. Bruce will post the paper in the SAB wiki when he gets a copy.

ADDENDA

Informal comments presented by John Wilkin at HT/OCLC meeting:

The HathiTrust vision, from the beginning, has focused on shared curation of the published cultural record. Notice that I didn't say "digital … record" or "historical … record." The mission statement we articulated made clear that we wanted to leverage the digital to deal with issues of shared print, and that we intended to focus not only on past publishing, but publishing in the future.

On this goal, we've made tremendous progress. In the first year, we reached nearly 5m volumes online. In the coming weeks, we will surpass 5m volumes, and by the end of next year we will reach nearly 8m volumes. We have the potential of having, in HathiTrust, more than 75% of all of the content in Google Book Search, and we will have a significant percentage of the content in the Internet Archive, as well as much local digitization. The size of the repository will soon rival all but the largest research libraries.

In the coming 2-3 years, we will be:

- adding digitization partners and content, probably surpassing 10m volumes online;
- adding library partners with a desire to support curation, even though they are not engaging in digitization on a grand scale. Our exploration with NYU is focused on how we can bring in these partners who will help us shoulder the cost of curation, and we have interest from others such as Arizona State, Maryland and MIT;
- we will give attention to issues of governance and finance;
- and we will of course continue to focus on preservation (e.g., the TRAC review) and access (e.g., the large-scale search).

It is only natural that the effort with OCLC has focused on discovery. That was the stated intent. But notice that, from the outset, the HathiTrust team has been concerned with issues of:

- bibliographic clarity and identification;
- bibliographic relationship—of one entity to another;
- digital to print relationships;
- provenance;
- volume-level (and not title level) information.

That our team has been focused on these issues in addition to discovery is not an accident. Unlike GBS, this effort focuses on *curation*, even when it is tackling discovery.

There is now emerging, particularly in our leadership, a discussion about the need to tie together a *collective* sense of our print holdings: what in HathiTrust has corresponding print, and where is that print? Why is there this focus? For many reasons:

- for legal reasons (Section 108, Section 121 and ADA, and Section 107);
- for fiscal reasons (a desire to explore new cost models and expand the partnership);
- for collaborative print management.

Bill Carney asked me to try to address the question "How can OCLC help HathiTrust reach its goals, and not just through WorldCat Local, but using OCLC enterprise resources?" The effort we've undertaken with OCLC to create a catalog **may provide** the foundations for this elaborate "holdings" record-keeping effort, an effort that is essential to our future. But perhaps it won't. Recognizing that this is not what OCLC set out to create (rather, it set out to create a record-sharing vehicle), we need to ask:

- Can OCLC become a reliable record of what is held and what was held?
- Can OCLC extend its current model from the record to the volume?
- And Can OCLC move from a model of central development in Dublin, OH to a model that uses shared, open development and capitalizes on member strengths?

These questions feel very much like questions we should explore over the coming year.

HathiTrust/OCLC Team Members for HathiTrust/WorldCat Local Implementation

HathiTrust Team
John Butler, co-chair, AUL for Information Technology, University of Minnesota
Lee Konrad, co-chair, Director, Memorial Library, University of Wisconsin – Madison
Julia Lovett, HathiTrust/OCLC Project Manager, Special Projects Librarian, University of Michigan
Adam Brin, Programmer Analyst, Bibliographic Services, California Digital Library
Lisa German, Asst. Dean of Technical and Collections Services, Penn State University Libraries
Suzanne Chapman, Interface & User Testing Specialist, Digital Library Production Service, Univ. of Michigan
Kevin Clair, Metadata Librarian; Penn State University Libraries
Jon Rothman, Senior Systems Librarian/Analyst, University of Michigan
Christopher Walker, Serials Cataloging Librarian

OCLC Team
Jeff Penka, Portfolio Director
Phil Norman, Director, Reference and Resource Sharing, Global Engineering
Mindy Pozenel, Director, WorldCat Discovery Services
Cheryl Snowdon, Product Manager
Bill Carney, OCLC/HathiTrust Project Lead, Content Manager, Business Development

Google Book Search Book Quality and Error Rate Reporting

Since release of "new" GRIN in February 2009, it has become apparent that Google's Overall Error Rate statistic is no longer an accurate reflection of individual book quality. As Google has made changes to its algorithmic error detection software, the meaning of this threshold has been altered in a way that is now less useful to the Library Partners1.

In the interest of presenting the highest quality product to end users, Library Partners need to make item level quality evaluations for purposes such as suitability for download/ingest, selection of materials for rescanning, etc. While the manual audit data that Google delivers seems to be accurate and reliable, it is presented in a manner that only allows for collection level assessment. This leaves Library Partners with no reliable means for determining the quality of individual books.

There are five main issues that the Library Partners would like to see addressed by Google with respect to quality improvement, assessment and reporting:

1) For local management purposes, the Library Partners need a reliable, useful quality assessment metric that more closely approximates the human experience of the book. We would like Google to develop such a measure in place of, or in addition to, the current form of algorithmic error reporting. This will allow Partners to make item level decisions based on quality data. Google should separate its own internal metrics from the one provided to the Partners so that these measures are not conflated. It is also requested that in undertaking this work, a differentiation is made between processing and acquisition errors to the extent possible.

2) We would like Google to present manual audit data in GRIN on an individual item basis, much as the Material Error % and Overall Error % are reported. This data should include the manual audit date, as well as an audit "score", and these parameters should also be usable in the advanced search interface.We also suggest that Google increase the scale of the manual audit process, to provide a greater number of quality assessment comparison points.

3) Now that Google is rejecting duplicate items for scanning and returning matching duplicates to its Library Partners, rejection decisions should take into account the quality of a previously-scanned item. This also assumes the availability of volume-specific quality metrics that are accurate and reliable. We would like more information about Google's plans in this area.

4) We would like Google to make an effort to improve communications about existing methods, workflows and processes so that the Library Partners can clearly understand the information that accompanies the digital versions and the implications for the library's management of its digital copies. This includes communications regarding any modifications or updates to those processes. It would be helpful to receive regular communications about Google's quality improvement processes and plans.

5) In light of statements that Google has made about diminishing returns in improving quality, we strongly encourage Google not to give up on attempts to improve quality; it is important that this remain a high priority of GBS.

Google has recently explained that as error rates go down the automated detectors are 'over-tuned' to detect potential issues, resulting in a high number of false positives. As a consequence, the reported Overall Error Rate is no longer a reliable measure of individual book quality as it would be experienced by a human reader.

10/13/2009