HathiTrust Digital Library

Update On June Activities

Top News

New Working Group on Computational Research Center – June was an exciting month for HathiTrust, both in terms of repository development and in terms of deepening collaboration among the HathiTrust Partners. Calls sent out in May for participation in two HathiTrust working groups were answered, and membership in both the Research Center and Development 'sandbox' groups was finalized. Members of the Research Center working group, which will develop a proposal for a computational Research Center to be created under the terms of the Google Settlement, include Steven Abney (University of Michigan), Jack Bernard (University of Michigan), Geoffrey Fox (Indiana University), David Goldberg (University of California Irvine), Robert McDonald (Indiana University), Qiaozhu Mei (University of Michigan), John Ober (California Digital Library), Beth Plale (Indiana University), Scott Poole (University of Illinois), Sarah Shreeves (University of Illinois), and John Unsworth (University of Illinois). The group will be coordinated by Kat Hagedorn, with project support to be provided by Jeremy York from HathiTrust.

Working Group on Development 'sandbox' – The Development 'sandbox' working group, which will work to create a development environment for partners to build and test repository applications and services, includes Stephen Abrams (California Digital Library), Albert Bertram (University of Michigan), Lynne Cameron (California Digital Library), Kaylea Champion (University of Chicago), Stephanie Collett (California Digital Library), Steve DiDomenico (Northwestern University), Bill Dueber (University of Michigan), Mike Durbin (Indiana University), Phil Farber (University of Michigan), Paul Fogel (California Digital Library), Eric Hetzner (California Digital Library), Sebastien Korner (University of Michigan), John Kunze (California Digital Library), David Loy (California Digital Library), Andy Mardesich (California Digital Library), Mairéad Martin (Pennsylvania State University), Jon Miller (University of Chicago), David Minor (San Diego Supercomputer Center), Bill Parod (Northwestern University), and Cory Snavely (University of Michigan).

Prototype for New HathiTrust PageTurner — As plans for the Development environment moved forward, the University of Michigan and California Digital Library (CDL) continued to explore possibilities for integrating the GnuBook reader into the current HathiTrust PageTurner to expand PageTurner's features and capabilities. The California Digital Library created a prototype GnuBook-integrated page turner application with repository code and a sample volume made available by the University of Michigan. Staff at the University of Michigan are currently testing the functionality of the prototype and will work with CDL in July to determine the next steps for development. This collaboration is exciting not only because of the enhancements it will bring to the existing PageTurner application, but because it demonstrates the way that shared development will enhance the services and capabilities HathiTrust is able to offer.

In This Newsletter

- New Working Groups Formed
- Prototype Created for New HathiTrust PageTurner
- CDL Staff to Visit Ann Arbor
- HathiTrust-OCLC Catalog Update
- HathiTrust.org Website Reorganization
- Stragegic Advisory Board Meeting Minutes Online
- Ingest Rate Increases
- Data API Testing and Development
- Changes to Google Metadata

New Growth

Number of volumes added:

	June	Total
Indiana Univ.	5,136	17,881
Univ. of California	113,139	199,245
Univ. of Michigan	223,460	2,882,371
Univ. of Wisconsin	37,473	211,330
Total	379,208	3,310,827

68,357 public domain volumes were added in May, bringing the total number of public domain volumes to 534,338 (about 16% of total content).

There's an elephant in the library.

www.hathitrust.org

HathiTrust Digital Library

Update On June Activities

Top News (continued)

CDL Staff to Visit Ann Arbor – Collaborating to enhance services and capabilities is the major theme of a visit that HathiTrust team members from the California Digital Library will make to the University of Michigan in July. Staff from both institutions will discuss a range of topics including the ingest of Internet Archive and other non-Google content, development of the HathiTrust PageTurner, communication about HathiTrust, and future development directions in a series of focused meetings from July 20th to 21st.

HathiTrust-OCLC Catalog Project — June was an important month for discussions about the HathiTrust-OCLC catalog, particularly regarding metadata and holdings information functions and display. At each juncture, the project team has prioritized meeting the unique needs of HathiTrust's all-digital catalog while maintaining consistency across the entire WorldCat database. For example, the team recently discussed how to accommodate viewability levels (e.g., search only, full-text, or a mix of the two in multi-volume sets) that do not occur in any other WorldCat records. The team has also focused on strategies for displaying and faceting on HathiTrust's many contributing institutions, in a way that would differentiate this information from print holdings. In striving to create a consistent user experience of HathiTrust, the team has turned to user feedback on the temporary beta catalog (http://catalog.hathitrust.org/). Future months will see increased focus on display and interface concerns as well as functionality issues.

HathiTrust.org Website Reorganization — Due to the evolving nature of HathiTrust and the additional information it has been necessary to incorporate on the HathiTrust website, a comprehensive reorganization of the website was undertaken by staff at the University of Michigan. The website has an identical look and feel, but information about areas such as preservation, rights management, partnership, and access have been more clearly separated and defined to be easier to locate. As part of the changes, additional information about the requirements and benefits of becoming a partner, how to become a partner, and the costs of partnership have been added. All of these changes can be viewed at http://www.hathitrust.org.

Strategic Advisory Board Meeting Minutes — The HathiTrust Strategic Advisory Board (SAB) met for the first time on June 17. The minutes of this meeting are posted on the HathiTrust website at http://www.hathitrust.org/sab. Future minutes of the SAB will be posted here as well.

Development Updates

Large-scale Search – University of Michigan staff ordered additional servers to support large-scale search in June, and prepared space for them in the MACC data center in Ann Arbor. UM also continued to explore the use of common-grams in large-scale search with a focus on refining the set of common terms in order to strike a balance between Solr index and performance.

July Forecast

- Determine next steps for collaborative development the HathiTrust Page Turner.
- Work to resolve indexing problems with the beta large-scale search.
- Begin working with facets in large-scale search and continue testing performance variables including common-grams and punctuation.
- Work on enhancements to the HathiTrust interface, most likely in Collection Builder.
- Begin loading Pennsylvania State University bibliographic metadata in preparation for ingest.
- Prepare for further collaborative work with CDL on the range of issues to be discussed during their visit to Ann Arbor.

Update On June Activities

Development Updates (continued)

Performance testing that was conducted in the process generated unexpected results that led to the discovery of a bug in a custom Solr punctuation filter. The bug was fixed, and tests will be conducted again in July. The large-scale search team has also encountered a problem when building full-text indexes for the beta large-scale search (http://babel.hathitrust.org/cgi/ls), in which indexing stops when memory errors are encountered after about a day and a half of indexing. This problem will be investigated further in July.

Ingest – As the numbers for New Growth show, more than 375,000 new volumes were added to the repository in June. This large amount is due both to an increase in digitized volumes available from partner institutions and an increase in ingest capacity gained by bringing a server online that had been held as a spare, boosting ingest rates by up to 25%.

Data API – University of Michigan staff have completed a response to feedback received from California Digital Library on the Data API following the release of the Data API specification in April (http://www.hathitrust.org/data_api). This response will be shared with CDL in early July. In the meantime, CDL continues to test end user functionality of the Data API alpha release. The Data API allows metadata of volumes in the repository, as well as OCR text and images of volumes themselves to be retrieved from the repository. Although it may have many uses, the Data API is intended to facilitate the development of custom applications by HathiTrust partners and others for delivering and using content in the repository.

Changes to Google Metadata – Over the last several weeks, Google library partners have worked with Google to incorporate improvements Google has made to the metadata it returns to partners with their digitized volumes. These include the addition of descriptive metadata, manual image auditing information, calibration information, and more. HathiTrust has accepted much of this new information into the HathiTrust METS metadata package that accompanies volumes in the repository. The changes occurred seamlessly and had no effect on the delivery of volumes through the PageTurner application.