HathiTrust Digital Library

Update On May Activities

Top News

Formation of Working Groups on Research Center and Development ‘sandbox’ – HathiTrust issued calls for names last month for participants in two new working groups: one to develop a proposal for a Research Center to be created under the terms of the Google Settlement, and one to create a development environment for HathiTrust partners to build and test repository applications and services. It is expected that membership of these two groups will be finalized in June.

University of California and University of Michigan Collaboration – With ingest of content from the University of California started in April and progressing well, staff at the University of California and the University of Michigan have broadened their sights to other areas of collaboration. One of these is the creation of an updated PageTurner application for viewing volumes in HathiTrust. Staff at the California Digital Library took a close look at the open source book reader GnuBook and outlined the steps for integration into the current HathiTrust PageTurner. Staff from both institutions discussed implementation possibilities in a conference call in May, and development is expected to begin in June.

New Blogs for Large-scale Search and the HathiTrust-OCLC Catalog Project — Two new blogs have been launched on the HathiTrust website (http://www.hathitrust.org/blogs). One will provide up-to-date information on HathiTrust’s efforts to enable full-text searching across the entire repository (http://www.hathitrust.org/blogs/large-scale-search), and the other will track the development of the permanent HathiTrust catalog, proceeding in collaboration with OCLC (http://www.hathitrust.org/blogs/hathitrust-oclc). RSS feeds for the blogs are available at http://www.hathitrust.org/blogs/large-scale-search/feed and http://www.hathitrust.org/blogs/hathitrust-oclc/feed.

HathiTrust-OCLC Catalog Project — In May, the HathiTrust-OCLC Catalog Implementation team (chaired by John Butler, Minnesota, and Lee Konrad, Wisconsin) developed a detailed communication plan for collaboration on the project. This plan includes biweekly meetings of the newly formed metadata subgroup, which will focus on metadata questions including display of access rights, faceting, and sorting of volume information. The team recently welcomed on board several new members with cataloging expertise to contribute to the subgroup, and its major goal for June is to finalize many of the metadata requirements for the catalog. To inform the setting of requirements for the catalog, user evaluation of the HathiTrust beta catalog is underway.

The HathiTrust-OCLC team also recently agreed upon a schedule for the project, including milestones and tentative deadlines for finalizing metadata and overall requirements, beginning ingest, and implementing the catalog. The projected date of Version 1 delivery remains at April 1, 2010. To read more, visit the new project blog at http://www.hathitrust.org/blogs/hathitrust-oclc.

In This Newsletter

New Growth

Number of volumes added:

	May	Total
Indiana Univ.	2,625	12,745
Univ. of California	86,106	86,106
Univ. of Michigan	21,794	2,658,911
Univ. of Wisconsin	4,696	173,865
Total	115,221	2,931,627

19,475 public domain volumes were added in May, bringing the total number of public domain volumes to 465,981 (about 16% of total content).

Update On May Activities

Development Updates

Storage – New storage was installed at Indiana University in May as planned, and is operating with no issues. This installation completes the second capacity expansion of both sites, bringing each to approximately 325 TB of storage.

Large-scale Search – A hardware configuration for servers to support large-scale search in HathiTrust was finalized in May, and data center space is nearly ready for use, pending installation of network gear. An order for the servers will be placed in early June for a planned deployment in July. Deployment of common-grams phrase searching in HathiTrust's experimental full-text beta search (http://babel.hathitrust.org/cgi/ls) was delayed in May, but was completed as of press time for this newsletter. In search benchmarking tests, the use of common-grams in phrase searching reduced average query response over a representative set of queries by more than 85%. The response time for the slowest query in the set decreased from 2 minutes to only 8 seconds. Up-to-date information on large-scale search benchmarking will be posted on the new large-scale search blog (http://www.hathitrust.org/blogs/large-scale-search).

Temporary Beta Catalog – Citation tools were introduced into the temporary beta catalog search interface, including export to Endnote, and functionality to email a citation.

Ingest – Several small legacy collections of the University of Michigan digital library were ingested into the repository in May. Development staff are using this work as an opportunity to refine the process of ingesting material not produced by Google, and to aid in the development of a validation tool institutions will be able to use to evaluate their own legacy collections against HathiTrust ingest requirements.

Data API – California Digital Library staff started working with the HathiTrust Data API. Michigan is working on a response to the feedback already received.

Outages – There were no HathiTrust outages in May. As of the last update, the Michigan and Indiana sites were in fully redundant operation (ingest occurs in Michigan, and data flows to Indiana via automation). We are now ensuring that users do not feel the effects of single-site outages, such as routine maintenance, by taking advantage of site redundancy. We will thus no longer report routine maintenance and outages at a single site unless performance was degraded sufficiently to warrant inclusion.

June Forecast

- Continue to explore the complexities of indexing with Solr, especially the selection of common words and the impact on index size.
- Possibly begin working with facets in large-scale search.
- Establish a basic method for a CDL programmer to work on HathiTrust Page Turner code with the ability to commit changes to the CVS repository. The first coding task will be to integrate GnuBook page turning functionality.
- Finalize many of the metadata requirements for the HathiTrust-OCLC catalog.

Presentations

Digital Library Federation	May 4
Archiving2009	May 5
CIC-Center for Library Initiatives	May 18
Research Libraries Group	June 2

Please see http://www.hathitrust.org/papers for all presentations, papers, and reports.

There's an elephant in the library.

www.hathitrust.org