HathiTrust Digital Library

Update On January Activities

February 12, 2010

Top News

New Cost Model – The HathiTrust executive committee approved a new cost model for partnership in December that will be adopted by all partners beginning in 2013. In the new model, partners will share in the cost of public domain and open access volumes preserved in HathiTrust, and in the cost of in copyright volumes that they hold, or have held, in their physical collections. The model will distribute the costs of curating and managing the digital collections in a way that more accurately reflects the benefits each partner receives from deposited volumes. It will also allow institutions to join HathiTrust who do not necessarily have content to deposit, but who wish to support and benefit from the long-term curation and access services that HathiTrust provides. Such institutions are eligible for partnership effective immediately, and do not need to wait for the 2013 general adoption. Details of the new cost model are available at http://www.hathitrust.org/documents/hathitrust-cost-rationale-2013.pdf. Please contact hathitrust-info@umich.edu for additional information and inquiries about partnership.

Disaster Recovery Planning – Following an evaluation of disaster preparedness performed last summer by an IMLS-funded intern, and the hiring of a preservation librarian in November, the University of Michigan is taking steps to formalize and expand HathiTrust's policies and practices relating to disaster recovery. The UM preservation librarian is leading a process to form a Disaster Recovery Planning Committee and, with support of a winter intern from the UM School of Information, has begun to gather key inventory, personnel, and workflow documentation. Guided by industry standards such as TRAC and best practices in the digital preservation community, the committee will ensure a high level of preparedness for known and unknown risks to the long-term integrity and use of materials in the repository. A preliminary meeting of key staff will occur in February, and membership in the Disaster Recovery Planning Committee will be finalized soon thereafter.

Digital Library Profile – As part of its participation in an NSF EAGER grant awarded in September 2009, HathiTrust completed a technological profile of its repository based on two frameworks developed by Johns Hopkins University. The profile can be found at http://www.hathitrust.org/technology.

Working Groups

Quality – In July 2009, the Strategic Advisory Board (SAB) assembled a working group to investigate issues surrounding the quality of partner institution volumes downloaded from Google. The working group was asked to research and provide recommendations on a quality threshold HathiTrust uses to limit ingest of poor quality volumes. The working group presented its recommendations to the SAB in January and the SAB decided to continue the working group with a revised and expanded charge. The new charge is to a) develop a set of quality principles for HathiTrust, b) monitor quality control as related to user experience, c) track developments in a separate quality working group established by Google and Google library partners following the Google partner summit in October, and d) evaluate HathiTrust practices with regard to thresh-

In This Newsletter

New Growth

Number of volumes added:

	January	Total
Indiana Univ.	38,344	151,816
Penn State	0	5016
Univ. of California	972	1,156,339
Univ. of Michigan	71,094	3,730,968
Univ. of Wisconsin	691	268.044
Total	104,342	5,312,183

5,384 public domain volumes were added in December, bringing the total number to 764,331 (about 14% of total content).

Update On January Activities

olding or limiting ingested content. Membership in the new group, called the HathiTrust Quality Ingest and Error Rate Working Group, is currently being determined.

Discovery Interface – With the version 1 catalog beta release only a few months away, the Discovery Interface Working Group is turning its focus to the usability of the catalog and its integration with existing HathiTrust Digital Library services (Collection Builder, Page Turner, and Full-Text Search). The Working Group formed a usability subgroup, which will collaborate with staff at OCLC to begin usability testing of the catalog before it is released. Testing will also be performed in post-release phases. Aspects of the pre-release analysis will include verifying accurate functionality and fulfillment of agreed-upon requirements.

In preparation for loading HathiTrust volumes into Worldcat for the version 1 release, staff at UM provided an API that will allow OCLC to display HathiTrust volume information in Worldcat records.

Collaborative Development Environment – UM staff have been gathering specific topics for the working group to discuss when it reconvenes (now planned for late February), and have developed a draft timeline for the steps ahead. A message to reassemble the group was sent in early February, and scheduling is underway. The area the group will address first is the design of a version control system. UM staff have also begun to research the GlusterFS cluster file system as a storage back-end for the environment.

Storage – The working group tasked with making recommendations on a third instance of storage for HathiTrust presented its final report to the Executive Committee in January. The group concluded that although there were significant benefits to implementing a third instance of storage, given the high level of preservation confidence in HathiTrust and the absence of economic conditions favorable for acquiring and operating new storage, there was no urgency in establishing a new instance. The group noted, however, that HathiTrust should be prepared to establish a third instance of storage if such a course becomes more economically feasible.

The Executive Committee would like to solicit broader feedback from partner institutions regarding these recommendations (especially from a collection development perspective), and requests that thoughts on the report and a third instance of storage be sent by email to hathitrust-info@umich.edu. Those who wish to remain anonymous should indicate this in their email. The full report of the working group is available at http://www.hathitrust.org/projects#wg_storage.

Ingest

General – Ingest rates were low in January, due in part to challenges UC experienced in retrieving bibliographic records from one of its systems. UM loaded the first set of bibliographic records for Minnesota, but could not begin ingest because of problems with Google's delivery of the content files. Ingest numbers from other institutions were also low because HathiTrust caught up with the rate that partner volumes were made available from Google.

February Forecast

- Complete and deploy Shibboleth authentication support
- Complete quality assurance processes for pilot of UC's Internet Archive-digitized materials and begin ingest into the repository
- Continue large-scale search performance monitoring
- Make progress toward the integration of Collection Builder functionality in full-text search results

Presentations

NISO Webinar	Feb 10

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

HathiTrust Digital Library

Update On January Activities

Internet Archive Ingest –UM began testing validation routines on a batch of 200 volumes of Internet Archive-digitized volumes from the University of California in January. The teams are revising validation strategies based on the findings of these tests and the results of quality assurance performed by UC staff on transformed, but not yet ingested objects. UM and UC will proceed with the ingest pilot in February, testing all aspects of bibliographic and content loading, validation, and access. Completion of the pilot is projected for late February.

New Programmer For Non-Google Ingest – UM extended the bidding period for the new programmer position through mid-January, and several new qualified candidates have been interviewed. UM staff are in the final stages of selecting candidates, and expect to have a new full-time staff member and a new part-time staff member on board by the end of February.

Development Updates

Shibboleth – Shibboleth implementation in HathiTrust is nearly complete. Major portions of the code are in place and UM staff have begun to contact partner institutions to exchange information that will allow individuals from partner institutions to authenticate into HathiTrust. Initial benefits to partners will be increased facility in creating personal collections in Collection Builder and full-PDF download of all public domain volumes. Non-partners will still be able to create collections using the University of Michigan "friend account" system. Deployment of Shibboleth is planned for March.

Data API – In January, staff at the University of Michigan began work on a web application that will use the Data API to facilitate the location and download of complete book packages for public domain volumes not digitized by Google. The application is being created entirely with data and services available to the general public and is meant to demonstrate uses that can be made of the API. The first step of crawling the repository for eligible volumes is in progress, and release of a beta version of the application is expected in February.

Large-scale Search – UM improved logging and log analysis in January, enabling staff to monitor search performance in a way that more closely resembles the user's experience. UM staff documented changes to large-scale search hardware in a new blog post entitled "Scaling up Large Scale Search from 500,000 volumes to 5 Million volumes and beyond".

New index servers were ordered for the Indiana site and are scheduled to be in service before the end of March. The current index release process already synchronizes an updated version of the index to be stored in Indiana on a daily basis. Acquisition of the new hardware will provide full redundancy of the large-scale search application servers as well. Two additional servers that will be used exclusively for index building are on their way to the Michigan site, and one server originally purchased for production service is being re-purposed for testing and development.

PageTurner – PageTurner development was slowed in January but will pick up in February and March as staff time devoted to the ingest of materials from the Internet Archive decreases.

Outages – There were no outages in January.