HathiTrust Digital Library

Update On March Activities

April 9, 2010

In This Newsletter

Partner Ingest

Internet Archive – Staff at the University of California completed quality review of the pilot set of Internet Archive-digitized volumes in March, and submitted a set of final issues to team members at the University of Michigan. These have have largely been resolved. Michigan staff also determined the cause of the validation error reported in last month's update. Correcting the error led to further revisions of the preservation metadata schema and re-evaluation of the validation routines put in place for Internet Archive-digitized content. Updates to these routines are currently being implemented. California sent bibliographic records for a set of 97,000 Internet Archive-digitized volumes to be loaded into HathiTrust. As soon as the updates to the ingest process and fixes for issues raised in QA are in place, download of these volumes will begin.

Local digitization – The University of Michigan has begun to receive locally-digitized content from several partner institutions for ingest into HathiTrust. Two programmers hired by Michigan in February have started to evaluate the material, determining needs and requirements for ingest, both in terms of digital package specifications and content transformation routines. Jessica Feeman, a programmer at Michigan and the original developer of the data validation and ingest system for HathiTrust, left her position at the end of March to start a family (congratulations, Jessica!). A new position will be opening in April.

Working Groups

Discovery Interface – OCLC loaded test batches of HathiTrust bibliographic records into WorldCat in March. After the batches were reviewed by OCLC and the HathiTrust team, OCLC initiated full-scale loading. At the end of March, 1.1 million HathiTrust records had been added to WorldCat through OCLC's eContent Synchronization mechanism, and the loading process continues.

HathiTrust and OCLC recently completed a first round of usability testing for the version 1 HathiTrust catalog, involving five participants in individual one-hour sessions. Members of the OCLC and HathiTrust teams are currently analyzing the results of the testing, particularly in relation to HathiTrust's requirements for the version 1 catalog. Special thanks in this effort are due to the HathiTrust colleagues at Penn State University, where the testing took place, as well as to OCLC for providing gift cards as incentives to participants.

Collaborative Development Environment – Michigan staff are in the process of designing the architecture for the new development environment according to the general direction set by the working group. The design incorporates practicalities such as directory naming conventions that will be compatible with the version control strategy. Staff are also discussing initial provisions for virtualization within the environment, including one virtual web and database environment for each developer, one for pre-release integration testing, and numerous instances for public "beta" exhibition and review of new features. The group is working to transition active HathiTrust development at Michigan to the new environment in April.

New Growth

Number of volumes added:

	Month of March	Total
Indiana Univ.	138	175,020
Penn State	1,469	6,613
Univ. of California	1,940	1,164,255
Univ. of Michigan	72,227	3,860,817
Univ. of Minnesota	880	65,876
Univ. of Wisconsin	11,923	315,650
Total	88,798	5,588,311

35,904 public domain volumes were added in March, bringing the total number to 854,790 (about 15% of total content).

HathiTrust Digital Library

Update On March Activities

April Forecast

- Complete and deploy Collection Builder integration with large-scale search
- Depoly redundant hosting of large-scale search service at Indiana site
- Register HathiTrust Shibboleth Service Provider with the InCommon Federation.

Development Updates

Shibboleth – Staff at the University of Michigan staff have been discussing the most appropriate set of attributes to request for release to HathiTrust applications via Shibboleth, consulting with experts at partner institutions, including Michigan's central Information and Technology Services, which will coordinate Shibboleth federation interactions for HathiTrust. Shibboleth will be a mechanism by which HathiTrust is able to provide specialized services, such as full-PDF download of repository volumes, to partners. The final attributes to be requested are eduPersonAffiliation, eduPersonScopedAffiliation, eduPersonTargetedID, and displayName. Registration of the HathiTrust Service Provider is in progress and we hope to release the service in April.

Large-scale Search – Programmers at the University of Michigan continue to investigate queries taking longer than 30 seconds to execute. The present theory is that certain components of the hardware (network cards) are causing intermittent problems that disrupt communication with the Solr server. The focus is on isolating and replacing the problematic cards.

HathiTrust team members from Michigan and Indiana are coordinating on the installation of new servers in Indianapolis to make the large-scale search service redundantly hosted at the Indiana and Michigan sites. This work has required the installation of new electrical and networking capacity in Indiana, which is almost complete. The setup and configuration of the new servers is expected to be fairly simple, as it is a near-replica of the architecture already in place in Michigan.

Collection Builder – Users will soon be able to add HathiTrust items to personal collections in batches through the large-scale search interface. Currently, items can only be added one-at-a-time via the HathiTrust PageTurner. Full-text searches can be executed on the subsets of volumes that are saved to collections, so the new functionality will additionally improve users' ability to narrow searches to selections of materials. Testing of the new service is in progress and it is expected to be released in April. Software and firmware upgrades were performed on storage systems at both sites in March.

PageTurner – University of Michigan developers continued to improve the performance of a new service that will deliver full-volume PDFs of public domain materials to users at HathiTrust partner institutions. The service will be available to partner institutions via Shibboleth authentication. Michigan also began development on a new method of delivering individual page images to the HathiTrust PageTurner, that will scale, rotate, and watermark images on the fly. Development is about 75% complete, and the new method is already being used in the collaborative development environment as part of the University of California's work to integrate GnuBook into the HathiTrust PageTurner. Michigan and California are working together on enhancements to the existing PageTurner interface to incorporate the GnuBook improvements.

Outages – Large-scale search service was unavailable from 10am-1pm EST on March 25 while software and firmware upgrades were applied to the storage systems in Michigan and Indiana. The upgrades did not result in outages for production systems. The large-scale search application is considered beta pending redundant hosting of the service in Indiana. In the future, we will work to communicate planned outages for services like large-scale search despite their beta status.