HathiTrust Digital Library

Update On August Activities

Top News

September 11, 2009

Working Group on Computational Research Center – The Research Center proposal planning group has made great progress in the last month. The group has continued discussions on the types of research that could utilize the centers, how results might be shared, and what environments/datasets are best suited to which types of research. In bi-weekly calls, subgroup meetings, and individual interviews, the team has been working through difficult issues such as defining non-consumptive research and recognizing hurdles related to the management and publication of research results. Next steps include developing a draft plan for the infrastructure of the centers and marrying legal and security restrictions with that infrastructure. The group aims to have a draft proposal prepared in early October and a full proposal completed later that month.

Working Group on Development 'sandbox' – Based on a general conversation with the working group and a useful discussion of potential use cases with UC staff during their Ann Arbor visit, staff at the University of Michigan have gathered enough information to start building the development environment. The initial goal is to support all of the current development projects in a single place, and provide a large subset of content with which to work. The new environment will be a substantial improvement over current conditions, and should be a building block for additional capabilities later on, including significant partner development. Michigan has racked, cabled, and started operating system installs on the equipment set aside for the project. When further progress has been made on the base installations the full working group will assemble to discuss the provisions of the environment.

Internet Archive Ingest – The University of California will be the first partner institution to contribute content digitized by the Internet Archive to HathiTrust. UC has taken the lead in providing a set of guidelines and best practices for ingest of this content that will be applicable to other partner institutions. UC's current activities include the creation of guidelines for a preferred file set to download from the Internet Archive for ingest into HathiTrust, analysis of bibliographic metadata issues and subsetting of objects, and the development of an approach for authoritatively identifying an institution's materials in the Internet Archive.

University of Michigan Press Backfile and Reprint Purchase Links in HathiTrust – HathiTrust is collaborating with the University of Michigan Scholarly Publishing Office and the University of Michigan Press to open access to the majority of the published backfile of the UM Press in HathiTrust. The volumes, which are being digitized by the Press, will be available in HathiTrust with an option to purchase a print-on-demand copy in mid to late October.

HathiTrust Disaster Preparedness – Over the summer, an IMLS grant-funded intern in digital preservation performed an in-depth evaluation of disaster preparedness in HathiTrust. The report provides detailed information about the strengths of HathiTrust's

In This Newsletter

New Growth

Number of volumes added:

	August	Total
Indiana Univ.	--	18,482
Univ. of California	148,810	457,494
Univ. of Michigan	58,878	3,129,152
Univ. of Wisconsin	--	215,045
Total	207,688	3,820,173

23,434 public domain volumes were added in July, bringing the total number of public domain volumes to 604,770 (about 16% of total content).

Update On August Activities

Top News (continued)

current disaster recovery planning, as well as recommendations for improvements in the short-, intermediate-, and long-term. It is available at http://www.hathitrust.org/technical_reports/HathiTrust_DisasterRecovery.pdf.

Prototype for New HathiTrust PageTurner – Staff from the University of California and University of Michigan held two teleconferences in August to discuss deeper integration of the UC prototype PageTurner into the existing application. Team members discussed strategies for offering full development capabilities on a limited amount of HathiTrust content in advance of the development 'sandbox' environment. A working strategy has been reached and a development space should be available in October. UC has continued in the meantime to improve GnuBook functionality with thumbnail views of page images and the ability to display full-text OCR. Staff at UM are investigating ways to alter current processes that make access-quality images available to the PageTurner, to produce images that can be used by the GnuBook.

METS Profile Available — Staff at the University of Michigan have created a version 1.0 METS profile for HathiTrust content, which can be downloaded at http://www.hathitrust.org/preservation. The profile currently applies only to Google content in HathiTrust, but will be updated to reflect requirements for locally-scanned content and volumes digitized by the Internet Archive.

Returned Duplicates — For several years, Google has been working on ways to reduce duplication in its digitization workflow. In August, it implemented processes that use metadata to detect volumes that have been scanned previously at other institutions so identical volumes will not be scanned again. The number of volumes rejected in this de-duplication effort has raised concerns among HathiTrust institutions about the accuracy of Google's detection processes. The University of California, the University of Wisconsin, Indiana University, and the University of Michigan have undertaken a review of volumes returned as duplicates to better understand how duplicate determination takes place. The four universities have identified a target set of materials to review and are finalizing methodology to perform a manual evaluation. It is hoped that the results will be available for the Google library partner summit later this month.

Mobile Interface – Michigan made significant progress on the development of a mobile interface to the HathiTrust Catalog in August. The work continues, and staff will next turn their attention to the PageTurner application. Initial development will be followed by user testing for both applications.

September Forecast

- Test large scale search performance on new dedicated server hardware.
- Begin working with facets in large scale search and continue testing performance variables including commongrams and punctuation.
- Add reprint purchase links to the HathiTrust interface for UM Press items.
- Continue development of mobile interfaces for the temporary catalog and PageTurner
- Establish a collaborative development environment for the HathiTrust PageTurner

HathiTrust Digital Library

Update On August Activities

Development Updates

Large-scale Search – After additional search performance testing in August, an improved index configuration was established by staff at the University of Michigan using a punctuation filter and a list of 400 common words (see blog post for details: http://www.hathitrust.org/blogs/large-scale-search/tuning-search-performance). This index configuration will be put into production on the new dedicated server hardware, which was installed in August. Michigan also completed additions to the indexing control software (SLIP) to support distribution of indexing across several servers, each with multiple Solr index shards. A continuous indexing strategy for this distributed system and corresponding requirements for storage configuration and scripting has been implemented, and the first indexing tests will have begun by the time this report is published.

Ingest – The number of volumes ingested dropped significantly in August as ingest rates caught up with the rate at which partner content was made available from Google.

Data API – Ed Summers provided insightful and constructive feedback on the HathiTrust Data API in a blog posting in mid-August (http://inkdroid.org/journal/2009/08/13/open-to-view/). The comments are being reviewed by University of Michigan staff.

Collection Builder – Two new APIs for Collection Builder are being tested by staff at Michigan. The first returns the list of collections owned by a user. The second adds multiple items to a collection. These APIs will support future integration of Collection Builder functionality into other applications, such as the HathiTrust temporary catalog.

Outages – On Wednesday August 5 from 8:15pm to 9:30pm EDT, service was degraded (service may have been unavailable to some users) due to a storage system problem at the Indiana site. On Sunday August 23 at 6:30pm EDT to Monday August 24 at 8:00am EDT, Wednesday August 26 from 5:00pm to 6:00pm EDT, and Friday August 28 from 7:25pm to 8:35pm EDT, service was degraded due to network connectivity problems to database servers.

Software and firmware upgrades were performed during the weeks of August 10 and 17 at both sites without incident or interruptions in service. The upgrades conducted during the week of August 17 were preventative in nature, and addressed a hardware problem discovered by the storage system provider, and which was the underlying cause of the service disruption on August 5.

The cause of the other outages has been thoroughly researched but is still not known; workarounds that eliminate any service impact have been put into place, systems are being monitored, and investigation into the problem continues.