HathiTrust Digital Library

Update On July Activities

Top News

August 14, 2009

UC Staff Visit Ann Arbor – HathiTrust project leads from the California Digital Library joined staff at the University of Michigan for two days of intense and fruitful discussion and planning from July 20-21. The teams consulted on a variety of forward-looking topics including a roadmap for the ingest of content digitized by the Internet Archive, strategies for future bibliographic metadata management, the challenges of providing help and feedback to users in a virtual library with multiple constituencies and stakeholders, HathiTrust PageTurner development, and creating infrastructure for collaborative development efforts. Several new planning efforts were initiated as a result of these discussions and both partners came away believing the visit had helped them to further coordinate efforts and was instrumental to continuing their successes in the future.

New HathiTrust Working Group On Storage – A new working group has been convened to explore the possibility of securing a third instance of storage for HathiTrust in the western United States. The working group members include Stephen Abrams, California Digital Library (co-chair), John Kunze, California Digital Library (co-chair), Luc Declerck, University of California San Diego, Rob Lowden, Indiana University, David Minor, University of California San Diego, and Cory Snavely, University of Michigan. If a third instance of storage is recommended, the group will investigate a variety of technical, management, and organizational issues involved in implementation.

Working Group on Computational Research Center – The Research Centers working group has been hard at work over the last month. The participants (please see the June update) have been engaging in a series of conference calls discussing issues related to the creation of the centers, including the types of research that will be done, the environment needed to support such research, and legal restrictions surrounding the use of the data. The group will continue to discuss these issues and others, such as funding sources and derivative research resulting from HathiTrust data use, in calls throughout August and September.

Working Group on Development 'sandbox' – The Development Environment working group convened for the first time in mid-July via teleconference to discuss the scope of the environment, the contexts in which development will occur (remote development versus local, specific use cases and desired features), and working group logistics. The group identified current applications such as the HathiTrust PageTurner and Collection Builder, and GROOVE, HathiTrust's ingest mechanism as priority systems to be made available in the development space, and conferred about particular ways that work will be done, such as code versioning. The development environment was a focus of one of the sessions during the meeting between California Digital Library and University of Michigan staff mentioned above, where further discussion on these issues took place. In the coming weeks, team members at Michigan will prepare

In This Newsletter

New Growth

Number of volumes added:

	July	Total
Indiana Univ.	601	18,482
Univ. of California	109,403	308,648
Univ. of Michigan	187,903	3,070,274
Univ. of Wisconsin	3,707	215,045
Total	301,614	3,612,449

47,028 public domain volumes were added in July, bringing the total number of public domain volumes to 581,336 (about 16% of total content).

Update On June Activities

Top News (continued)

hardware that has been set aside for the project and do preliminary configuration of the environment on that hardware.

Prototype for New HathiTrust PageTurner – Collaboration between the California Digital Library and the University of Michigan to enhance the HathiTrust Page Turner with GnuBook functionality continued in July, primarily in the form of discussions about division of labor and the establishment of a basic collaborative work environment. A new planning and development team with staff from both institutions met in mid-August to kick off the next phase of GnuBook and PageTurner development.

HathiTrust-OCLC Catalog Project — The HathiTrust WorldCat Local Implementation team is nearing the completion of high-level requirements document for the version 1 catalog, with a target deadline of August 31, 2009. The team also began to document usability issues and suggestions for the proposed interface. OCLC has begun working on the e-content synchronization process that will bring HathiTrust's records into WorldCat Local. 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/).

HathiTrust Statistics — Member institutions have identified the need to make statistics about how HathiTrust is being used more broadly available within the partnership. As a provisional measure, access statistics gathered by Google Analytics are being provided to representatives at these institutions. While these analytics will be useful in the short-term, there is a need for a reporting tool that will provide more granular information, such as usage by institution and by format, in the future.

Development Updates

Large-scale Search – University of Michigan staff investigated the indexing problems with the beta large-scale search that were reported in the last update. The problems were due to a shortage of available memory. However, a decision was taken to wait for new hardware to be deployed before taking further action. The new hardware, purchased in June to support large-scale search, was received in July, and is currently being prepared for testing and use. With the new hardware in place, it is planned to have full text search of all volumes in HathiTrust by October 1st.

UM staff made refinements to the custom punctuation filter for large scale search, and ran tests only to discover the filter did not provide the performance boost anticipated. The punctuation filter has been set aside temporarily, but has potential for future implementation. Tests conducted by staff to compare response times for common-grams Solr indexes in various configurations resulted in a new emphasis being placed on the importance of a well-tuned list of common words. A new program that evaluates the total number of term occurrences for the most frequently occurring words in an index was created to aid in the selection of common words for this list. Additional details can be found on

August Forecast

- Establish a collaborative development environment for the HathiTrust PageTurner
- Test large scale search performance on new dedicated server hardware.
- Begin working with facets in large scale search and continue testing performance variables including common-grams and punctuation.
- Work to integrate Collection Builder functionality with the temporary catalog.
- Develop beta mobile interfaces for the temporary catalog and PageTurner to the point initial user testing can be conducted.

Update On June Activities

Development Updates (continued)

the HathiTrust Large Scale Search Blog (http://www.hathitrust.org/blogs/large-scale-search/). Four new posts were added to the blog in July.

Ingest – Ingest was slowed in July by the discovery that Google was making volumes available for ingest that did not contain the required descriptive metadata. Google addressed the problem and ingest continued as normal after these volumes were re-ingested.

Data API – University of Michigan staff responded to feedback received from California Digital Library on the Data API and discussion of the API continued when CDL visited Michigan. Key issues that have arisen are security and determining how much functionality should be built into the baseline API.

Collection Builder – Michigan explored solutions for integrating Collection Builder functionality into the temporary HathiTrust Catalog. Planned improvements would allow users to save multiple items to a public or private collection directly from a search results or bibliographic record listing in the catalog.

Outages – At 8:15pm EDT, Wednesday, August 5th, an incident (that we are currently investigating) at the Indianapolis data center caused HathiTrust storage at that site to be unavailable for 1 hour and 15 minutes. During that time the entire Ann Arbor node of HathiTrust as well as web servers at the Indianapolis node continued to be available for users. Our current load balancing and failover strategy does not adequately account for this sort of partial failure. In the worst case, a user whose browser was directed to the Indianapolis site may have been unable to view books in the repository during the period from 8:15-9:30pm EDT. For most users, however, load balancing would have directed their browsers to the Ann Arbor site during this period. In the coming year, we will be replacing mechanisms that currently handle load balancing and failover, and will devote attention to developing a more nuanced failover strategy.