HathiTrust Digital Library

Update On September Activities

October 9, 2009

Top News

HathiTrust participates in grant from NSF – Sayeed Choudhury of Johns Hopkins University, John Wilkin of the University of Michigan, and Amy Friedlander of the Council on Library and Information Resources (CLIR) are co-PIs in an NSF EAGER grant to determine the needs and requirements for developing an open-access repository for publications arising from NSF-funded research. The PIs will leverage Johns Hopkins' experience in evaluating digital repositories, HathiTrust's experience with large-scale infrastructure and ingest of digital objects, and CLIR's experience and facility in bringing together groups of experts to determine next steps and directions on targeted issues. CLIR will host a series of workshops focusing on technical requirements, business and policy concerns, and organization and operations issues relating to the open-access repository. Johns Hopkins and HathiTrust will evaluate various technical systems based on the recommendations from the workshops. The creation of a sustainable, efficient, and scalable model to deliver the products of NSF-funded research to users at no cost will have a transformative impact on the dissemination and use of this valuable work.

University of Michigan Press Backfile and "Buy a Reprint" Links In HathiTrust – HathiTrust has begun ingest of the majority of the published backfile of the University of Michigan Press. More than 350 volumes are now available in the temporary catalog and the HathiTrust PageTurner, with an option to purchase print copies of many of the volumes in the PageTurner. The collection is the first of what is hoped will become many collections or bibliographies in HathiTrust that are maintained by official sources such as organizations, faculty, and librarians. The partners are still working on a name for these types of collections. More information about the Press partnership, including links to the official press release and the collection itself, are available at http://press.umich.edu/digital/hathi. Full-text search is available inside of the UM Press collection, and all other HathiTrust collections (see the Collection Builder home at http://babel.hathitrust.org/cgi/mb?a=listcs;colltype=pub).

Returned Duplicates — The University of California, the University of Wisconsin, Indiana University, and the University of Michigan have undertaken a review of volumes returned by Google as duplicates to better understand how duplicate determination takes place. During the month of September, staff members evaluated materials that were rejected by Google in August, identifying matches and potential mismatches. Results are currently being compiled and analyzed, and will be presented at the Google Partner Summit.

Working Group on Computational Research Center – The Research Center advisory group has completed their initial round of discussions on the demand, structure, content inclusion, legal considerations and funding of the Research Center. A report on that work will be submitted to the Executive Committee in the coming

In This Newsletter

New Growth

Number of volumes added:

	Sept.	Total
Indiana Univ.	1,036	19,518
Univ. of California	64,210	521,704
Univ. of Michigan	81,829	3,210,981
Univ. of Wisconsin	7,230	222,275
Total	154,306	3,981,227

36,400 public domain volumes were added in September, bringing the total number of public domain volumes to 641,170 (about 16% of total content).

Top News (continued)

weeks. The group identified the need for additional strategies to gather specific information about the composition and ongoing use and support of the Research Center. A plan to assemble and incorporate this information should be in place in October as well.

Working Group on Storage – A series of teleconferences have led to the construction and refinement of a table defining the important decision criteria for adding a third instance of HathiTrust storage. By mid-October the group will develop a version of these criteria with institutional-specific weighting factors. It will then work to reconcile the weightings and develop a final recommendation.

Working Group on Collaborative Development Environment – Michigan staff have completed operating system installs on the initial development environment equipment. Staff will next configure one of the development servers with the base set of software required to support known demands on the environment, including shared development with staff at the University of California on the HathiTrust PageTurner. The initial configuration will be documented and discussed with the working group for further revisions and enhancements.

New Programmer For Non-Google Ingest – In the near future the HathiTrust partners will hire a developer dedicated to receiving non-Google materials from their respective institutions and preparing them for ingest into the repository. The new hire will speed the addition of these materials to HathiTrust and develop specifications and processes that will be applicable to content from new partners in the future.

Internet Archive Ingest – Staff members from the University of California, the University of Michigan, and the University of Illinois held a teleconference in late September to discuss the file formats for Internet Archive-digitized content that will be included in the HathiTrust book package. The partners are working to build consensus on a package that will meet the needs of all institutions contributing this content. The University of Michigan and University of California held two teleconferences in September to discuss issues surrounding ingest itself, such as book package identifiers and ways of preparing ingested OCR for use in full-text searching and viewing applications.

Upcoming Changes to Tab-delimited HathiTrust Metadata Files – Beginning with the full metadata file produced on December 1, 2009, additional fields will be added to the tab-delimited HathiTrust metadata files that are provided at http://www.hathitrust.org/hathifiles (a description of the files is available at http://www.hathitrust.org/hathifiles_metadata).

Fields to be added include the rights determination reason code and the date of last rights determination. With this data included, the tab-delimited files will become an ongoing accessible source for information on how and when rights

- Create and maintain full-text indexing and search services in the new production environment.
- Continue to explore the addition of facets in full-text search. Facets have introduced metadata to the full-text index, and therefore new sorting options, including weighted relevance, will need attention.
- Continue to investigate potential solutions to the problem of dynamically serving images to GnuBook.

Presentations

iPRES	Oct 6
PASIG	Oct 7
NISO Forum	Oct 9

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

There's an elephant in the library.

Update On September Activities

Top News (continued)

determinations are made. The new tab-delimited fields will be added to the end of the current record structure in order to minimize any potential disruption for existing users of these files. More details on this change will be included on the website as they become available.

Development Updates

Large-scale Search Launch October 19 – In September, the University of Michigan worked to revise and debug production index-building routines to support a comprehensive index of HathiTrust volumes. This index is distributed across five servers with two Solr shards, or index fragments, on each server. In the process of running the routines it was confirmed that Logical Volume Manager (LVM) snapshots could be used effectively to deploy index updates. Concurrent testing of the indexes in the new search environment showed a significant improvement in performance over the current environment, as had been expected. The new full-text search service is targeted for release on October 19. When it is live, the full text of the more than 4 million volumes in HathiTrust will be searchable by anyone with a Web browser. At that time, a new portal interface will replace the current page at http://catalog.hathitrust.org, providing access to full-text search, bibliographic search, and linking to custom collections in the Collection Builder.

With the release of full-text search on the horizon, HathiTrust has begun exploring options for offering faceted browsing of content in conjunction with full-text search. The University of Michigan has built and performed preliminary testing on an index of 500,000 volumes that includes metadata suitable for faceting of search results. The tests suggest that the impact of faceting on full-text search performance will be tolerable in the new environment.

Principal developers for the open source Solr software integrated Michigan's contribution of common-grams code into the Solr code base. It is now a permanent feature of Solr and, of course, the HathiTrust indexing process.

HathiTrust/OCLC Catalog – The HathiTrust/OCLC Catalog team recently reached an agreement on metadata requirements for the version 1 catalog. To finalize these requirements, input was sought from catalogers both within and outside of the regular group. The team is also in the process of finalizing user interface requirements. A face-to-face meeting between OCLC and HathiTrust is being planned for November, where the group will begin to lay out a vision and timeline for version 2 of the catalog.

Ingest – Ingest rates were low in September as HathiTrust remained caught up with content made available from Google, and due to an issue of metadata encoding that required Google to reprocess a number of volumes before they could be downloaded. Ingest of metadata from Penn State is expected to begin in October, with ingest of content to begin immediately after.

Prototype for New HathiTrust PageTurner – Staff at the University of Michigan investigated ways of altering the current process by which images are transformed for access, in order to produce images that can be used by the GnuBook. A conclusion was not reached and investigation will continue in October. Development work at the University of California will also continue in October, as staff prepare a feature that will allow users to view thumbnail images of the pages in a volume.

Collection Builder – As mentioned above, the UM Press volumes will form the first officially sponsored collection in Collection Builder. Another new feature of the Collection Builder, and the PageTurner application as well, is that users accessing HathiTrust from partnering institution campuses will see the name of their institution in the bottom left corner of the screen. This note will let users know that their institution is supporting the effort to make this content available and ensure its preservation over the long-term. Staff at the University of Michigan continue to work to integrate Collection Builder functionality into the temporary catalog. Negotiating authentication requirements between the two applications has introduced some complications, but options continue to be explored.

Outages – There were no outages in September.