HathiTrust Digital Library

Update On November Activities

December 11, 2009

In This Newsletter

Top News

Release of Large-scale Search Application – On November 19, HathiTrust launched a new service enabling full-text search of all volumes in the repository. Indexing of newly ingested volumes is ongoing, but the release of the first production index (containing approximately 4.6 million volumes) is the culmination of more than a year of research and benchmark testing conducted by staff at the University of Michigan. This new service dramatically changes the way researchers are able to use our collections and, along with the release of the bibliographic catalog in May, demonstrates HathiTrust's commitment to providing sophisticated ways of accessing and using collections preserved in the digital repository. The official news release is available at http://www.ns.umich.edu/htdocs/releases/story.php?id=7426. More can be read about large-scale search in the Development Updates section below.

Development Opportunities – This month we provide the second in a series of announcements about development opportunities in HathiTrust. These are opportunities that have been identified by HathiTrust partners, and are available to HathiTrust partners, to create key systems or services that will benefit the partnership as a whole. Each month we will provide a brief description of one of these opportunities, give a sense of the level of priority that it has, and provide additional information about what might be involved in developing and supporting it. The opportunities are also listed on the HathiTrust website at http://www.hathitrust.org/projects. The opportunity described this month is usage reporting.

Usage reporting

Description: A clearer sense of the level of use of library materials in HathiTrust will help shape extended activities such as collection management and further digitization. Volumes in HathiTrust may, in some cases, be read in their entirety, while in other cases they may only be searched. To what extent are search-only materials viewed? Which works that are fully viewable are displayed? Where does that access originate? As HathiTrust introduces authentication, to what extent do users authenticate to get access to a fuller array of services? How frequently is the HathiTrust catalog searched, and how does that use compare to the use of full text indexes? These are some of the questions that an improved service for usage reporting will help to answer.

Resources available: HathiTrust retains raw log data and registers some uses through Google analytics.

Priority: moderate

Additional details: An institution that undertakes this work must:

- clearly outline a commitment to undertake appropriate measures with regard to user privacy (e.g., with regard to IP addresses and, at such time that HathiTrust implements Shibboleth, user authentication information). Such efforts should include secure storage of sensitive data, appropriate aggregation of data so as to anonymize use by specific individuals, and a commitment to not transfer private user data to a third party;

New Growth

Number of volumes added:

	November	Total
Indiana Univ.	32,427	116,559
Penn State	108	4,783
Univ. of California	105,864	892,278
Univ. of Michigan	11,729	3,428,993
Univ. of Wisconsin	12,511	255,216
Total	115,890	4,697,829

15,980 public domain volumes were added in November, bringing the total number to 717,941 (about 15% of total content).

HathiTrust Digital Library

Update On November Activities

December Forecast

Top News (continued)

- outline a process for design and specifications with a group of interested partner libraries;
- give consideration to producing reports consistent with appropriate library community standards (e.g., COUNTER and SUSHI).

Working Group on Computational Research Center – Working group members provided final feedback on the *Call for Proposals for a HathiTrust Research Center* that will be distributed to HathiTrust institutions in December. The Research Center will make textual and image data in HathiTrust available for a wide variety of computational research and analysis purposes, including research in areas of digital humanities, linguistics, automated translation, and searching and indexing techniques.

Working Group on Collaborative Development Environment – Additional effort devoted to the release of Large-scale Search in November delayed further progress on the development environment, but it is now a prime area of focus. The first milestone, a preliminary proof-of-concept environment that supports current development efforts will be ready for developers at the University of Michigan and the University of California to begin testing in the first half of December. Once this milestone is reached, the working group will be re-engaged to discuss the current provisions of the development environment and explore next steps.

New Programmer For Non-Google Ingest – Staff at the University of Michigan received and reviewed applications for a position to aid in the transformation and modification of non-Google content for ingest into HathiTrust. Five candidates were interviewed by phone in November and three were invited for in-person interviews. After a period of review, the search committee decided to continue the search and repost the position. An additional avenue, involving hiring one or more part time student employees operating with close supervision, is also being considered.

Internet Archive Ingest – Much progress was made toward the ingest of content digitized by the Internet Archive in November. The University of California shared specifications for a preferred set of files to be downloaded into HathiTrust with the broader community of Internet Archive digitization partners, and received constructive feedback from the group. In continued weekly calls, staff at UC and UM discussed procedures and conventions for content transformation, file-identification, and preservation and technical metadata, as well as error logging, exception handling, and policy issues surrounding the deposit of digital objects. The ingest team is working to have practices surrounding many of these issues finalized by mid-December, when UC will deliver bibliographic metadata for an initial set of IA-digitized volumes to UM. Once the transformation and validation processes for ingest have been finalized and coded, UM will conduct a pilot test, downloading and ingesting this initial set of volumes. It is hoped that the full pilot, including quality review of ingested volumes, will be completed by mid-January.

- Refine indexing methods, including frequency of complete index optimization and best index shard size
- Develop processes for rebuilding the entire index
- Finalize specifications for content digitized by the Internet Archive and prepare for ingest pilot
- Add Collection Builder functionality to the HathiTrust full-text search interface

Update On November Activities

Top News (continued)

Changes to Tab-delimited Metadata Files – As of December 1, rights determination reason codes are included in the metadata files available for download at http://www.hathitrust.org/hathifiles. Please see the file specification at http://www.hathitrust.org/hathifiles_metadata for updated information.

Development Updates

Large-scale Search – The launch of HathiTrust's large-scale search application was postponed in October in order to acquire additional hardware to accommodate new index growth. Due to a variety of factors including a delay in hardware delivery, staff at the University of Michigan altered their index storage strategy and reconfigured the Solr index servers at Michigan to use the Isilon storage system as a back-end. In addition to solving issues related to the size of the index, moving from existing direct-attached storage to the Isilon network-attached storage more readily accommodates the significant index growth that occurs during routine index optimization. The move to Islion is a temporary strategy, however, and staff at UM will be investigating alternative options for storing the large-scale search index over the long-term.

After the storage reorganization, a small backlog of indexing was completed and a new automatic daily indexing process was developed. The University of Michigan launched the full-text service in mid-November and it is performing well.

With an eye toward achieving full redundancy of the search service, staff at UM implemented a nightly synchronization of the index to the Indiana site. Work toward redundancy is ongoing, however, and will involve further research to determine the optimal size of index shards. The size of index shards will help to determine the optimal number of index servers to deploy to guarantee adequate search performance, as well as the additional server deployments and workflows needed to support continuing testing of the search system, routine indexing, and volume re-indexing. Once complete, additional equipment will be purchased and installed at both the Michigan and Indiana sites as appropriate to establish full redundancy.

In additional ongoing work, staff at UM performed analysis of post-release query logs to improve performance testing and cache warming.

HathiTrust/OCLC Catalog – On November 20th in Chicago, the HathiTrust Discovery Interface team met with the corresponding OCLC-WorldCat Local implementation project team for a productive visioning session of the HathiTrust catalog beyond version 1 due in April 2010. Each group shared its long-term vision for the project, and together began to identify areas of common interest and commitment for the year of work following the release of version 1. The HathiTrust team's draft vision document is available for review and comment at http://www.hathitrust.org/documents/hathitrust-discovery-vision.pdf.

Update On November Activities

Development Updates (continued)

Ingest – The University of California sent shipments of bibliographic data from its Santa Cruz and San Diego campuses to the University of Michigan for ingest in November, totaling approximately 400,000 volumes. Ingest of these volumes, in addition to 200,000 more that are expected from UC's North Regional Library Facility, will bring HathiTrust to more than 5 million volumes by the end of the year. UM received an initial shipment of bibliographic metadata from the University of Minnesota in November as well. As these and subsequent records from Minnesota are loaded into HathiTrust, ingest of the digital volumes will begin.

A lower number of new volumes were ingested into HathiTrust in November than expected because of a large number of volumes that were re-processed by Google. Google continually re-processes images and OCR of volumes to make improvements and corrections, and these volumes enter a single queue with newly processed volumes for ingest.

Collection Builder – Following the meeting with OCLC staff in Chicago, the focus of Collection Builder integration in the temporary catalog has shifted to integration in the full-text search application. This move sidesteps cross-site linking issues that were encountered, and will provide useful experience on which to build Collection Builder inclusion in the HathiTrust Catalog at a later time.

Outages – There were no outages in November.