HathiTrust Digital Library

Update On October Activities

November 13, 2009

Top News

Development Opportunities – This is the first in a regular 'column' of development opportunities for HathiTrust. System and software development in HathiTrust is performed by contributions by HathiTrust partners. Although many HathiTrust systems and services must sit on central servers, our initiative relies on open systems and modularity, making it possible for partner institutions to develop key pieces of functionality. In this new column, each month we will provide a brief description of a system or service that has been proposed by HathiTrust partners, attempt to give a sense of the level of priority for that system or service, and provide additional information about what might be involved in developing and supporting it. These services will also be listed on the HathiTrust website at http://www.hathitrust.org/projects. This month we focus on an opportunity that has arisen directly from the expansion of HathiTrust day-to-day operations and the needs of new partners:

Ingest reporting

Description: The deposit of digital volumes and associated metadata into HathiTrust, referred to as "ingest," involves a significant number of updates to administrative systems — bibliographic records added, digital volumes ingested, and access rights established. Many data elements will be of interest to the contributing institution, and each institution may drive local processes based on the current status of content in the repository (e.g., the percentage of in-copyright works may highlight the value of performing copyright determination work, or a low number of items available in the Google Return Interface may stimulate exploratory discussions with Google). A system that combines all of the available streams of administrative data into a simple web-based reporting system may have considerable value not only for transparency but also for local decision-making.

Resources available: Staff at the University of Michigan and the University of California have assembled a table of relevant data feeds with a brief description of each in the following document: http://bit.ly/2Jk5mm.

Priority: moderate

Additional details: An institution that undertakes this work must:

- outline a process for design and specifications with a group of interested HathiTrust partner libraries.
- in consultation with partner libraries, give consideration to authentication and authorization needs for this system.

Upcoming Opportunities

- Usage reporting
- Print holdings database
- Ingest transformation

HathiTrust participates in grant from Mellon Foundation – With support from the Andrew W. Mellon Foundation, Associate Professor Paul Conway of the University of Michigan is leading a one-year research and planning project to find and test new procedures for validating the quality and usefulness of digital objects in HathiTrust. The short-term goal of the project is to prepare and submit a funding proposal to a federal granting agency to explore possibilities for validating

In This Newsletter

New Growth

Number of volumes added:

	October	Total
Indiana Univ.	64,614	84,132
Penn State	4,675	4,675
Univ. of California	264,710	786,414
Univ. of Michigan	206,283	3,417,264
Univ. of Wisconsin	20,430	242,705
Total	553,963	4,535,190

60,791 public domain volumes were added in October, bringing the total number to 701,961 (about 15% of total content).

Update On October Activities

November Forecast

Top News (continued)

these characteristics through manual and automated methods. The long-term goal is to develop criteria and methods to brand the trustworthiness of volumes in HathiTrust and other digital repositories for fulfilling specific purposes (e.g., reading, printing volumes on demand, and performing computational research). Such a branding or certification process would give assurance that content within a repository is worthy of preservation, and increase the value of that content in broader discussions about storage and management solutions for both digital and print collections.

Google Summit – At a periodic meeting between Google and partner libraries, HathiTrust members worked with Google on issues related to the ingest of materials digitized by Google. Some topics discussed included strategies for improved metrics with regard to the quality of materials, and volumes rejected as duplicates from Google's scanning workflow. The metrics discussed around quality could potentially be used to characterize or filter content that enters the repository (e.g., in the case of poor quality, to prevent ingest). The duplicate analysis conducted by HathiTrust partners is now being factored into Google's continuing development of duplicate detection and return. Evaluators at the University of Michigan will continue to examine volumes returned as duplicates throughout the semester.

Working Group on Computational Research Center – The working group submitted its final report to the HathiTrust Executive Committee in October, containing specifications for a HathiTrust Research Center and a request for proposals from interested HathiTrust institutions to build and host the Research Center. The Executive Committee has reviewed the document and pending final edits from the working group, will distribute the RFP to the partner institutions in November.

Working Group on Collaborative Development Environment – Michigan staff observed a problem with a hard drive in one of the nodes in the development environment cluster and spent time in October troubleshooting the problem and investigating other potential options for hard drive configuration on the nodes. As a result of this investigation, the system BIOS on all nodes will be upgraded and one of the nodes will need to be rebuilt. Work continues on setting up a preliminary development environment on the first node.

New Programmer For Non-Google Ingest – Applications for a programmer position at the University of Michigan to aid in the transformation and normalization of content to be ingested from a variety of digitization sources have been received and reviewed. UM has started the interview process and hopes to have the new programmer in place as soon as possible. The partners made the decision to centralize this ingest functionality initially in order to expedite the inclusion of non-Google content in the repository. Over time it is expected that individual partners will take a greater role in validating and preparing their content for ingest, leveraging tools and processes that result from this initial investment.

- Fully deploy comprehensive full-text search
- Continue to explore facets in full-text search
- Continue to research solutions for adding Collection Builder functionality to the HathiTrust catalog search interface
- Begin to develop HathiTrust METS specifications for content digitized by the Internet Archive
- Begin preparations to conduct usability testing on the HathiTrust/OCLC catalog interface

Update On October Activities

Top News (continued)

Internet Archive Ingest – Weekly conversations centered on the ingest of content digitized by the Internet Archive continued in October between staff at the University of Michigan and University of California. Particular focus was placed on determining the standard identifier scheme that should be use for the content when it is ingested into HathiTrust. The University of California's ARK identifiers, which exist for nearly all of its Internet Archive volumes, appear to be the most promising. Staff at UM have begun to test these identifiers in repository processes to detect any issues that may arise.

The University of California revised its set of preferred files to be downloaded from the Internet Archive for inclusion in the HathiTrust ingest package. The spec will be distributed to other IA partners in the near future for comments. UC also engaged in analysis of bibliographic data of IA-digitized files from its different campuses and continued development of an approach to authoritatively identify an institution's volumes in the Internet Archive.

Upcoming Changes to Tab-delimited HathiTrust Metadata Files – As reported in last month's update, 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 copyright determination reason code and the date the database entry was last updated. With this data included, the tab-delimited files will become an ongoing accessible source for information on how and when rights 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.

Development Updates

Large-scale Search – Staff at the University of Michigan successfully indexed all volumes in HathiTrust using the newly acquired hardware. However, the official launch of the large-scale search application was postponed in order to acquire additional hardware to accommodate new index growth. The original estimate of storage requirements turned out to be low once common-grams technology was introduced. Common-grams offer significantly better search performance but result in an increased index size. The very large number of volumes ingested into the repository in October contributed to the immediate need for more indexing space as well. Optimization of the index, a process occurring at regular intervals, requires as much as 3 times the size of the index shard being optimized.

Faceting of search results, a feature supported by Solr, was further explored in October. Faceting requires the addition of bibliographic data to the full-text index. A faceted index was built across two shards to look for potential problems in scaling.

Update On October Activities

Development Updates (continued)

Early indications are that performance is only affected slightly with the facets employed.

HathiTrust/OCLC Catalog – After finalizing metadata requirements for the version 1 catalog in September, the HathiTrust/OCLC Catalog team turned its attention in October to interface requirements. The team is currently finalizing interface requirements for version 1 of the catalog and has agreed to engage in collaborative usability testing during the first quarter of 2010. Meanwhile, OCLC's e-content synchronization work for HathiTrust remains on schedule, and is expected to be completed by the end of the calendar year.

Ingest – HathiTrust ingested a record 553,963 volumes in October. These included nearly 5,000 volumes from Penn State and initial loads of volumes from the University of California's Santa Cruz and San Diego campuses. Ingest of volumes from Penn State will continue in November. Subsequent shipments of metadata for up to 600,000 additional volumes from UC campuses are expected in November. Ingest of these volumes will begin shortly thereafter.

Prototype for New HathiTrust PageTurner – Enhancements to the HathiTrust PageTurner application and integration with the open source GnuBook were on hold in October as development efforts at Michigan focused on large-scale search and initial configuration of the collaborative development environment. The collaborative environment will enable staff at the University of California to fully test and troubleshoot GnuBook functionality in production conditions. Development of an "image API" is still needed to deliver page images from the repository for display in GnuBook.

Collection Builder – Michigan further explored integration of Collection Builder functionality into the temporary catalog search interface. Some difficulty was encountered due to cross-site linking restrictions, but options will continue to be explored.

Outages – There were no outages in October.