HathiTrust Digital Library

Update On February Activities

March 12, 2010

Top News

In the last several months, the HathiTrust partners have made steady progress in expanding the repository's ability to support the variety of digital outputs produced at their local institutions. While the bulk of content in HathiTrust currently is the result of Google's digitization efforts, preserving and delivering content from libraries' non-Google sources is an important part of HathiTrust's mission to meet the needs of libraries broadly, and assemble a comprehensive collection of published materials that is co-owned by libraries themselves. Three items in this month's update highlight our efforts in this area: our progress in ingesting materials digitized by the Internet Archive, the hiring of two new programmers to focus on the transformations and normalizations involved in bringing in diverse content, and the creation of a demonstration application that uses the HathiTrust Data API to deliver master repository content from non-Google sources to users. We will be highlighting developments such as these in the coming months.

Internet Archive Ingest – Ingest of UC volumes digitized by the Internet Archive was delayed in late February due to a validation error that UM staff encountered, but ingest of more than 200 pilot volumes was begun in early March. Following quality review of the volumes by UC staff and the resolution of any associated issues, download of UC's Internet Archive-digitized volumes will begin in earnest. Staff at UC and UM are in the process of compiling technical and procedural documentation related to Internet Archive ingest to share with partner institutions and the community at large.

New Programmer for Non-Google Ingest – UM has hired two new programmers, for a total of 1.7 FTE, to concentrate on developing ingest routines and common workflows for non-Google-produced materials. These will include materials digitized by the Internet Archive and through local digitization efforts at partner institutions.

Data API - The interface to the Data API demonstration application that was undertaken by Michigan in January is available at http://www.lib.umich.edu/two-over-threehundred/. The goal of the application was to use HathiTrust's Data API to facilitate the location and download of complete book packages for public domain volumes not digitized by Google. The code used to produce the demonstration is also available. The application is still processing the HathiTrust data files, and so will only display a subset of the full data.

Working Groups

Quality, Ingest, and Error Rate – The working group kicked off activities under its recently revised charge in February, and will be meeting on a monthly basis. At this stage, the group is undertaking information gathering and doing planning for work items, including building a framework for defining quality principles and developing a varied set of scenarios under which content would be gated from entering HathiTrust. This work will help to spur discussion and identify larger issues that are play. Members of the group include Paul Fogel (California Digital Library), Peter Gorman (University of Wisconsin), Bryan Skib (University of Michigan), and Paul Soderdahl (University of Iowa).

In This Newsletter

New Growth

Number of volumes added:

	Month of Feb.	Total
Indiana Univ.	23,066	174,882
Penn State	128	5,144
Univ. of California	5,976	1,162,315
Univ. of Michigan	50,873	3,781,841
Univ. of Minnesota	64,966	64,966
Univ. of Wisconsin	35,683	303,727
Total	108,977	5,434,537

54,555 public domain volumes were added in February, bringing the total number to 818,886 (about 15% of total content).

HathiTrust Digital Library

Update On February Activities

Discovery Interface – The HathiTrust-OCLC team made significant strides in February towards the version 1 catalog beta implementation, with some adjustments to the projected timeline. Due to changes in OCLC's product release cycles, the catalog is now expected to be complete in May 2010. The HathiTrust library team is now exploring strategies and requirements for the catalog's public release, with the guidance of both the HathiTrust Strategic Advisory Board and Executive Committee.

The load of HathiTrust bibliographic metadata into WorldCat remains on schedule. OCLC is currently testing the first batch of records, and large-scale loading will take place throughout the month of March. Preliminary user testing is currently underway at Penn State and will be complete in mid-March, thanks to the collaborative efforts of OCLC and HathiTrust's usability group.

Collaborative Development Environment – The working group reconvened via conference call in February to discuss strategies for version control. All agreed that the version control tools used should facilitate development at local sites as well as within the environment itself, and allow public availability of the source code. Modern distributed version control systems, including some third-party systems such as GitHub, fit well with these needs, and UM staff will propose an architecture to the group at their next meeting in early March for approval. The group also discussed building logical divisions in the environment to segregate its use for various purposes, such as active code development, integration testing and staging for production release, the presentation of relatively stable "beta" versions of software systems, and replicating and troubleshooting issues live in production.

Ingest

University of Minnesota – Ingest of content from the University of Minnesota began in February, with nearly 65,000 volumes being deposited. All of these volumes are government documents, and are part of a larger effort of the Committee on Institutional Cooperation (the Big Ten plus the University of Chicago) in partnership with Google, to digitize more than 1 million U.S. federal documents from their combined collections. The Minnesota documents themselves can be found by clicking on the University of Minnesota facet in the HathiTrust Catalog.

Development Updates

Shibboleth – UM is in the process of finalizing Shibboleth attribute release requirements for HathiTrust applications in coordination with partner institutions, and is registering HathiTrust as a service with the InCommon Shibboleth federation. The release of this enhancement to HathiTrust applications is still planned for a March timeframe.

Large-scale Search – The large-scale search index grew to the point in February that it exceeded the Solr/Lucene limit of 2.1 billion unique terms. Core Lucene developer Michael McCandless graciously provided a patch raising the limit to 274 billion unique terms. Michigan continued performance tests aimed at identifying optimal shard sizes. Staff at Michigan also led team members at CDL on a walk-through of the large-scale search implementation in mid-February.

March Forecast

- Deploy the new page image server and related changes to the HathiTrust PageTurner
- Release Shibboleth authentication support
- Continue large-scale search performance monitoring
- Complete quality assurance processes for pilot ingest of Internet Archive-digitized materials
- Begin ingest of all UC's Internet Archive-digitized volumes

Update On February Activities

Four new redundant servers for large-scale search indexing arrived at Indiana and will be installed once additional power and networking infrastructure work has been completed, probably in late March. Two new servers for index building arrived in Michigan and are tentatively scheduled for March installation as well, pending staff availability.

PageTurner – Michigan revamped the PageTurner code that generates PDFs from the repository in February, optimizing it for high performance delivery of full-book PDF files containing full-resolution page images. The ability to download full PDF files of HathiTrust public domain volumes will be available to partner institutions when Shibboleth is implemented. Michigan also explored pipelines for fast on-the-fly generation of scaled, rotated, and watermarked page images and developed a prototype image server. Once completed, it will serve all individual page images not encapsulated in PDF.

Outages – There were no outages in February.