HathiTrust Digital Library

Update On March Activities

April 8, 2011

Top News

TRAC Certification

HathiTrust has been certified by the Center for Research Libraries (CRL) for compliance with the Trustworthy Repository Audit and Certification (TRAC) criteria for digital repositories. This important certification has been a key aim of the partnership since the repository's founding in 2008, and one we intend to uphold in coming years. The full audit report is posted on the CRL website. HathiTrust posted a news release on the certification and updated documentation on HathiTrust's compliance with TRAC elements. In conjunction with this announcement, we have included a spotlight on HathiTrust technology below, posted also at http://www.hathitrust.org/technology.

HathiTrust Webinar

Partners from across the country attended the HathiTrust new partners webinar on March 23. A variety of topics were addressed, including HathiTrust's organizational structure and costs, our collections and services, and future directions. Partners also had an opportunity for Q&A with the presenters. The webinar will be offered on two additional dates: Tuesday April 12, 12:30-2:00pm, and Friday April 15, 12:30-2:00pm (both Eastern Daylight Time). If you would like attend, please RSVP to Jeremy York as soon as possible before each webinar: jjyork@umich.edu. Please also include any questions or issues you would like the presenters to address.

Open Webinar

Due to a high level of interest expressed by non-HathiTrust partner institutions, an open webinar will be held on May 3 and May 5 from 11am-12pm Eastern time. This webinar will be open to the public. As above, if you would like attend, please RSVP to Jeremy York as soon as possible before each webinar (jjyork@umich.edu), and include any questions you would like the presenters to address.

IMLS Quality Grant

In 2010, the Institute of Museum and Library Services granted the University of Michigan and Associate Professor Paul Conway funding to research quality in large-scale digital repositories. The grant project is using HathiTrust as a test-bed for review. Work on Phase One of the project commenced in late January 2011 with the creation of a project team to focus on defining error types and levels of severity, statistical analysis processes, a web application for data entry, and project management procedures. By the end of March, the team had identified initial project needs and accomplished the following: identified twelve initial error types including scales of severity, hired and trained two data coders, coded an initial random sample of 15 volumes from the public collection, analyzed variance in coding within the sample, and produced a first draft of procedures for quality evaluation. The team also connected with project members at the University of Minnesota who

April Forecast

- Deploy new version of PageTurner with BookReader
- Complete storage Replacement work
- Draft a specification for Data API security enhancements

Papers

Shane Beers and Bria Parker, "HathiTrust and the Challenge of Digital Audio"

All HathiTrust papers, presentations, and reports are available at http://www.hathitrust.org/papers

You can follow HathiTrust on Twitter at http://www.twitter.com/hathitrust

Update On March Activities

will be participating in the grant, sharing initial documentation and results. For further information regarding progress and updates, please see the HathiTrust grant projects webpage.

Print Holdings Database

HathiTrust has been working to design and populate a database of information representing the print holdings of all partner institutions. This database will serve a number of important functions:

- It will support analysis of the overlap of institutions' print holdings with digital holdings in HathiTrust – this information is required in order to implement the new financial model.
- It will form a foundation for the expansion of legal uses of materials in HathiTrust (e.g. services to users with print disabilities) by partner institutions.
- It will facilitate collaborative collection development and management operations.

Total Volumes Added

	March	Overall
Columbia University	15	58,480
Cornell University	31,371	280,381
Indiana University	1,093	182,988
Library of Congress	71,418	71,418
NYPL	126	258,691
Penn State University	1,824	38,998
Princeton University	8,758	228,224
University of California	62,804	2,367,215
University of Chicago	1,945	5,172
University of Illinois	73	14,501
University of Madrid	2,774	88,311
University of Michigan	15,038	4,318,394
University of Minnesota	4,068	83,566
University of Wisconsin	12,206	443,370
Yale University	21	161
Total	213,534	8,430,230

Public Domain (~26% of total)

Total	231,171*	2,204,521

*This count includes volumes already in the repository to which rightholders have newly opened access

To date, approximately 119.5 million rows of data have been received from partners, with each row representing one copy of a single volume monographic print item that is (or previously was) held at a partner institution. At this point, we have outgrown the hardware where initial database testing and development took place. When new HathiTrust development environment hardware becomes available in early April, a new version of the database will be created, all of the data we have received will be loaded, and we can begin generating statistics and preliminary cost modeling data. At the same time, we will be working toward a near-term production release of the database to support services to users with print disabilities at partner institutions.

Upcoming development work will focus on improved duplicate detection and clustering mechanisms on two fronts: we are working with OCLC on the development of tools that will provide improved identification of potential duplicate bibliographic records; and we will be ramping up our work on duplicate detection/matching mechanisms for the parts of multi-part works to allow expansion of the print holdings database to include serials and multi-part monographs.

Update On March Activities

Ingest

Local Digitization Ingest

Staff members at the University of Michigan are currently investigating a sample of rare volumes digitized from Universidad Complutense de Madrid for deposit. Staff are also performing final evaluation of approximately 600 locally-digitized volumes submitted by Northwestern University.

Library of Congress

Ingest of an initial set of more than 70,000 volumes from the Library of Congress, digitized in partnership with the Internet Archive, was completed in March.

Working Groups

Collections

The Collections committee continued to work on recommendations regarding duplicate volumes in HathiTrust, coordinated print management, and responding to users requests to contribute volumes to the repository.

Communications

HathiTrust figured prominently in the news in March, and the working group was in high gear to disseminate announcements regarding the Google Settlement ruling, HathiTrust's agreement with Summon, and the positive outcome of the TRAC audit. The group also began setting up a HathiTrust Facebook presence and conducted the first of three new partner webinars.

Development Environment

New, more powerful MySQL servers were installed in the development environment to support the additional performance requirements of the partner holdings database. The new servers are being synchronized in real time with the old in preparation for a cutover planned for early April.

Discovery Interface

The WorldCat Local Prototype usability test reported in last month's update ran for a few weeks in March. User experience experts from the Discovery Interface Working Group (DIWG) and OCLC are analyzing the data and drafting a report of findings for review. The Full-Text Search Subgroup, charged to "identify and prioritize features and functions anticipated to have immediate high-impact value to users presented it recommendations that can be reasonably afforded by the existing technology framework," presented its analysis and recommendations to the DIWG, where it received full endorsement.

Update On March Activities

Usability

Usability Working Group members continued their work as liaisons in other HathiTrust committees in March. The group also began to develop a set of personas and use cases to inform development and policy-making surrounding HathiTrust applications and interfaces. The Usability Group is still looking for people to join the new User Experience Special Interest Group (UX-SIG), reported in last month's update. Please contact Suzanne Chapman (suzchap@umich.edu) if you are interested in joining this group or have any questions about participation.

User Support Working Group

The charge of the User Support Working Group was approved by the Executive Committee and is posted online. The group plans to schedule its first call in April, and will become the primary body responsible for addressing user inquiries submitted through HathiTrust interfaces and the HathiTrust contact address.

Development Updates

Bibliographic Data Management

The Metadata Management System development team at California Digital Library (CDL) continued development of the core database system in March. The team continues to review workflows for receiving bibliographic data from HathiTrust content-contributing partners, and has responded to changes in bibliographic processing at the University of Michigan by adjusting processes in the new system to mirror those changes. Team members continue to benchmark data loading performance and adjust computing resources for optimum results. Interviewing continues for a Principal Metadata Analyst. The position opening is posted on the CDL website. Interested individuals are invited to apply.

Collection Builder

Staff at Michigan completed modifications to Collection Builder, enabling it to support the creation of permanent, full-text-searchable collections of arbitrary size. Details on the modifications were reported in the February update. The first real-world test, the creation of a collection of more than 50,000 volumes, was completed without issue.

Data API

Now that the Collection Builder enhancements are done, Michigan staff will return to design and implementation of Data API security enhancements.

Full-text Search

Staff at the University of Michigan researched and estimated technical feasibility and implementation effort for potential new large scale search features for the

Update On March Activities

Full-Text Search Working Group's report. The Michigan implementation team began to mock up and prototype likely new features.

OAI

In conjunction with other changes to support Creative Commons licenses in HathiTrust, staff at the University of Michigan modified the Michigan OAI provider to include records for open access and Creative Commons-licensed items. Records for these items are available in the "hathitrust" and "hathitrust:pd" sets. Please see HathiTrust Data Availability and APIs for more information about OAI in HathiTrust.

PageTurner

Michigan staff reconfigured the single volume "Search in this text" mechanism to properly handle German double quotes. They also revised the rights algorithm to allow full book PDF download for Creative Commons-licensed volumes without authentication. For volumes where full-book PDF download is not allowed, an appropriately informative message is now displayed to the user.

More progress was made toward the release of an updated PageTurner with integrated BookReader functionality. Specifically, Apache2 and Plack were installed by staff at Michigan and tested in the Development Environment, and work is underway to do the same in production. Michigan remains on track for full production deployment of PageTurner with BookReader in April.

Storage Replacement Cycle

Michigan staff began the second half of storage replacement work at the Indiana and Michigan sites in March. All new equipment at the Michigan site is online and operational. The trip to complete replacement work at the Indiana site was delayed by several weeks while staff waited for a new database server and two new validation and dataset preparation servers to arrive; work on both projects will be combined and completed in a single trip.

Outages

HathiTrust Collection Builder was unavailable from 5:00pm to 6:20pm EDT on Wednesday, March 16 to change the underlying search engine to work against the full-text index.

HathiTrust Technology Spotlight

Cory Snavely, Head, Library IT Core Services, University of Michigan Library

HathiTrust is intended to provide persistent and high-availability storage for deposited files. In order to facilitate this, the partnership uses a storage architecture with a rich set of features designed for fault tolerance and long-term data retention.

Update On March Activities

Central to the storage architecture is the use of two synchronized instances of storage with wide geographic separation (located in Ann Arbor, MI and Indianapolis, IN) and an encrypted tape backup with 6 months of previous-version retention (located in a separate data center several miles from the Ann Arbor storage instance). All storage is physically secure, locked in racks within data centers that are accessible only to authorized IT personnel.

The need for continuous integrity checking is fundamental to HathiTrust's data management strategy and underlies the choice of online (spinning magnetic disk) media for primary storage. Internally, each storage instance uses N+3 Reed-Solomon parity redundancy, which is analogous to but more fault-tolerant than conventional RAID 5 storage due to the additional parity redundancy. The storage system internally performs in-flight data integrity checks as well as periodic integrity checks of all at-rest data, and makes use of parity redundancy to permanently repair any errors encountered. External to the storage system, HathiTrust also conducts periodic validation of data with stored checksums to ensure that data has been ingested correctly and remains intact.

Storage equipment replacement is an ongoing annual process and assumes that equipment has a useful lifetime of 3-4 years. The storage system is modular and virtualized, with files split into blocks that are distributed across nodes of a cluster and automatically redistributed as needed to balance storage utilization equally. Storage replacement therefore requires no manual movement of data, as this balancing is a normal housekeeping function of the system. Storage nodes that have reached retirement age may be removed from the cluster with an administrative command, and new nodes may be added, with all movement of data managed internally while employing the in-flight integrity checks described earlier. The remove and add processes neither disrupt services nor diminish the N+3 redundancy.

The following links provide more detailed information about our storage, backup, and disaster planning:

- Building a Future by Preserving Our Past: The Preservation Infrastructure of HathiTrust Digital Library
- HathiTrust is a Solution: The Foundations of a Disaster Recovery Plan for the Shared Digital Repository
- HathiTrust Trustworthy Repository Audit and Certiciation compliance