HathiTrust Digital Library

Update On December Activities

January 13, 2012

Top News

Changes to Tab-delimited Files

On February 1, HathiTrust will be adding three additional columns to the tab-delimited inventory files ("hathifiles") available at http://www.hathitrust.org/hathifiles. The files are frequently used by partners and non-partners as a means to obtain full bibliographic records for HathiTrust items to load into local catalogs (see HathiTrust Data Availability and APIs). The additional columns will identify the publication date and publication location of volumes in HathiTrust, as well as volumes that have been identified as U.S. federal government documents.

Ingest

Works Digitized by Internet Archive

Staff at Michigan continued conversations with staff at the University of Florida and Getty Research Institute regarding ingest of Internet Archive-digitized materials.

Working Groups and Committees

Working groups and committees in HathiTrust may have an operational or strategic focus. See http://www.hathitrust.org/working_groups for more information.

Operational

Communications

The Communications Working Group continued to work on a public services-oriented communications package, as well as announcements for new partners and the major milestone of 10 million volumes.

User Experience Advisory Group

The User Experience Advisory Group began reviewing the current home page and discussed additions and issues that will need to be addressed in a forthcoming redesign. Group member Jenny Emanuel contributed a "Perspectives from HathiTrust" blog post about the group's persona work that was completed in November.

User Support Working Group

The User Support group is still seeking nominations for new members. See the Update on November Activities for details. The table on the following page contains a summary of the issues received by the User Support Working Group in December.

Late Breaking News

HathiTrust passes 10 million volumes (view statistics and a timeline on the HathiTrust blog).

January Forecast

- Continue work on the advanced search feature for full-text search

Papers & Presentations

Jeremy York "HathiTrust: Reviewing Goals, Accomplishments, and Opportunities for Collective Action". Coalition for Networked Information, December 13, 2011.

See http://www.hathitrust.org/papers for all papers, presentations, and reports.

HathiTrust Digital Library

Update On December Activities

Projects

Bibliographic Data Management System

Team members from California Digital Library continued work on processes to compare bibliographic records in Zephir, the new metadata management system under development, with records in HathiTrust's existing system. Zephir team members continued to load and test new records as well, and refine the timeline for migration of bibliographic metadata management services to Zephir in coordination with staff at the University of Michigan.

HathiTrust Publishing (HTPub)

Staff at the University of Michigan revised the goal statement for HTPub (see the project web page) and plans for system architecture. Staff also began work on establishing a project timeline.

HathiTrust Research Center

Several changes were made to the HTRC leadership in December. John Unsworth, a key member of the Team at Illinois, accepted a position as vice provost for Library and Technology Services and chief information officer at Brandeis University. He will be leaving the University of Illinois but remain on the Executive Management Team. The Team will keep its base composition of 2 members from the University of Illinois and 2 from Indiana University, so this change will add one new member. Stephen Downie, Associate Dean for Research at the University of Illinois Graduate School of Library and Information Science, will fill the position left by John. Stephen's research has focused on music information retrieval and data mining. This work has involved building significant infrastructure for research, including grappling with issues of allowing computational access to in-copyright material. Finally, Marshall Scott Poole is stepping aside as co-director of the HTRC for personal reasons, though he will remain on the Executive Management Team. Stephen Downie will take his place as co-director of the HTRC with Beth Plale, who is co-director on the Indiana University side. Beth also chairs the Executive Management Team. The changes are in effect as of January 1, 2012.

IMLS Quality Grant

In December, project staff completed physical review of more than 90% of the volumes in the first 1,000 volume sample drawn from HathiTrust. Staff are working to

User Support Issues

	November	December
Content	**107**	**81**
Quality	102	71
Non-partner Digital Deposit	0	2
Collections	5	6
Cataloging	**43**	**30**
Access and Use	**103**	**107**
Copyright	55	59
Permissions	10	4
Takedown	1	2
Print on Demand	2	1
Inter-library loan	0	2
Full-PDF or e-copy requests	15	28
Datasets	1	2
Data Availability and APIs	2	0
Reuse of content	1	1
Web applications	**24**	**18**
Functionality problems	5	9
Problems with login specifically	1	1
General questions about login	2	0
Partners setting up login	3	1
Usability issues	2	2
Feature requests	2	1
Partner Ingest	**3**	**5**
General	**47**	**50**
Partnership	6	7
Infrastructure	0	0
Miscellaneous	41	43

*See User Support Working Group Issue Types for a description of the types of issues included in each category.

Update On December Activities

arrange on-site review with cooperation from HathiTrust member libraries for the approximately 70 volumes that are not available via inter-library loan due to poor condition, non-circulating status, or other reason.

Project staff concluded page-level data collection for the second production sample in December (see the Update on September 2011 Activities for details on the composition of the sample). The full dataset will be sent to the project statistician in early January for analysis. Data collection for the third production run began in the late December. The third production run focuses on Internet Archive-digitized volumes published pre-1923.

Project staff continue to define requirements for a new quality review interface, targeted specifically for review of volume-level errors such as missing, duplicate, and out-of-order pages. Please visit the project website for updates.

Development Updates

Full-text Search

Michigan staff released a new version of the full-text search index in December. The new release corrected an error in the "Original Location" metadata facet and provided additional metadata for advanced search and relevance ranking. It also made it possible for full-text search results and facets to reflect whether or not users from partner institutions are able to view in-copyright items when lawful access is permitted (HathiTrust is currently pursuing providing access to in-copyright works to users who have print disabilities, for preservation uses, and in circumstances where works are copyright-orphaned). Access in these circumstances, which are still pending deployment to partners, is dependent on partner institutions owning or previously owning print copies of works in question and users' location inside or outside the United States.

Michigan staff continued development of an advanced search feature for full-text search, including preliminary testing of the first working prototype in HathiTrust's development environment.

California Digital Library continued work on a spelling suggestion feature for full-

Total Volumes Added

	December	Overall
Columbia University	4	64,176
Cornell University	9,871	383,690
Duke University	21	4,522
Harvard University	434	53,440
Indiana University	324	186,912
Library of Congress	15,769	89,411
North Carolina State University	0	3,196
University of North Carolina - Chapel Hill	0	8,087
Northwestern University	237	5,649
New York Public Library	76	259,453
Penn State University	1,821	42,917
Princeton University	350	249,679
Purdue University	0	887
University of California	114,906	3,287,654
University of Chicago	1,733	10,608
University of Illinois	0	14,503
Universidad Complutense	28	108,668
University of Michigan	22,907	4,504,601
University of Minnesota	916	90,239
University of Wisconsin	15,902	527,334
University of Virginia	12	47,396
Utah State University	0	46
Yale University	0	23,674
Total	185,311	9,966,572

Public Domain (~27% of total)

Total*	50,434	2,712,626

*Includes volumes opened through copyright review or rights holder permissions.

Update On December Activities

text search queries. A CDL developer established an account in the HathiTrust development environment and used a sample index of public domain materials to try different strategies for automatically building a bigram dictionary of words, with the different spellings users might enter.

Tom Burton-West's proposed talk on "HathiTrust Large Scale Search: Scalability meets Usability", was accepted by popular vote for the 2012 Code4Lib Conference in Seattle, WA.

Throttling

Staff at Michigan released a new throttling mechanism for HathiTrust, which allows throttling levels to be set at more granular levels. Users are now less likely to be throttled in the course of normal use since the new throttling policies are applied to specific scenarios such as viewing thumbnail or page images, or downloading PDFs, as opposed to all use generally. Throttling ensures compliance with third-party restrictions on bulk download of materials, and helps to ensure a consistent and reliable experience for all users.

PageTurner

In connection with HTPub, Michigan staff continued work to adapt the HathiTrust PageTurner to display XML content.

Security Risk Assessment and Vulnerability Test

Michigan Library staff continue to work with central IT security analysts to complete the Risk Assessment that was started in November, and have received the final report of the vulnerability penetration test. The report revealed no vulnerabilities that enabled direct or indirect access to the repository, but noted software issues such as cross-site scripting vulnerability and also made recommendations for increased firewalling at the Michigan site. All software issues noted in the report were addressed in December. A broader firewalling project for the data center where the Michigan instance is hosted is already in progress but not yet complete, and so some provisional steps were taken to tighten security while that effort continues.

Outages

HathiTrust services were inaccessible or diminished for several periods in December due to problems related to the release of the new throttling system (all times EST): on Tue, Dec 13 4:25-4:30pm, Wed, Dec 14 11:10am-12:00pm, and Wed 12-21 7:30-10:30am, all page viewing was affected, and on Tue, Dec 13 3:45-5:00pm, full-book PDF download was affected. Additionally, page viewing of volumes classified as "Public Domain in the United States" in HathiTrust was intermittently unavailable on Wed 12-21 from approximately 1-4:30pm EST due to an apparent outage with an externally-hosted proxy detection system.

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

Note on outages:

HathiTrust sends notice upon discovery and resolution of unscheduled outages and in advance of scheduled outages and maintenance work that may result in an outage. We welcome and encourage additional recipients for these notices. If your institution is not receiving outage notifications and would like to, please contact feedback@issues.hathitrust.org.