HathiTrust Digital Library

SPECIAL EDITION - 2012 MID-YEAR REVIEW

June 25, 2012

In the first half of 2012, HathiTrust continued to expand our partnership, to further develop and refine our services, and to benefit from grant funded evaluations and explorations. In this period we have also seen momentum building for our HathiTrust Research Center, and the important milestone of the establishment of a new Board of Governors. The following provides detail on the richness of our many activities and accomplishments.

Highlighted Achievements and Activities

Details on each item can be found in the monthly updates from 2012, available at http://www.hathitrust.org/updates.

New Partners

Two new partners joined HathiTrust in the first half of 2012:

- Washington University
- University of Delaware

New Content

HathiTrust Partners contributed nearly 400,000 volumes to HathiTrust from January – June 2012, raising the total number of total volumes to 10.4 million (view our Ten Million and Counting blog post and timeline). More than 3 million of this total (about 30%) are in the public domain. Deposits from all institutions are shown in the table to the right.

Locally-digitized content

HathiTrust began or continued conversations with several institutions regarding direct ingest of locally-digitized content:

- Columbia University
- Emory University
- Getty Research Center
- Northwestern University
- Princeton University
- Yale University
- University of Florida
- University of Illinois
- University of Iowa
- University of Michigan Press
- University of Utah
- Utah State University Press

The University of Michigan created a first iteration of tools that partners can use to package their content to HathiTrust specifications prior to submission.

Internet Archive-digitized content

HathiTrust began conversations with the Getty Research Center, Penn State University, and the University of Florida regarding ingest of volumes from the Internet Archive.

Total Volumes Added	Jan-June 2012	Total
Columbia University	8	64,184
Cornell University	16,181	399,871
Duke University	1	4,523
Harvard University	146,299	199,739
Indiana University	752	187,664
Library of Congress	5	89,416
North Carolina State University	0	3,196
University of North Carolina - Chapel Hill	1	8,088
Northwestern University	1,554	7,203
New York Public Library	106	259,559
Penn State University	405	43,322
Princeton University	1,170	250,849
Purdue University	23,894	24,781
University of California	49,128	3,336,782
University of Chicago	10,213	20,821
University of Illinois	81,648	96,151
Universidad Complutense	3,159	111,827
University of Michigan	34,767	4,539,368
University of Minnesota	9,231	99,470
University of Wisconsin	11,874	539,208
University of Virginia	1,526	48,922
Utah State University	44	90
Yale University	4	23,678
Total	392,140	10,358,712

Public Domain (~30% of total)

Total*	320,629	3,033,255

SPECIAL EDITION - 2012 MID-YEAR REVIEW

Google-digitized content

HathiTrust ingested large numbers of volumes from the University of Illinois (~80,000 volumes) and Harvard University Library (~150,000 volumes).

Governance, Working Groups, and Committees

Board of Governors

HathiTrust conducted elections for a new Board of Governors in March and established the Board, composed of both elected and appointed members, in April. The proposal to create a Board of Governors was one of the proposals accepted by partners at the HathiTrust Constitutional Convention in October 2011 (view all proposals). The Board took the reins from an Executive Committee, which was established by the founding HathiTrust partners. A report on the Board's first meeting is posted in the Update on May 2012 Activities.

Collections Committee

The Collections Committee released its report on duplicate volumes in HathiTrust, recommending that HathiTrust retain all duplicate copies ingested into the repository for the time being, with periodic reassessment. The Committee also made progress on a process for responding to requests and offers to include additional materials in HathiTrust.

Communications Working Group

The Communications Working Group produced a Resources page for HathiTrust, containing overview documents, handouts, and guides created by HathiTrust partner libraries, the Communications Working Group, and non-partner sources. The working group released blog posts on HathiTrust's achievement of 10 million volumes, full-text search enhancements, and, in collaboration with University of Michigan staff and the UX Advisory group, creating collections in HathiTrust. The Communications group launched a Pinterest account for HathiTrust, and submitted a briefing for the new Board of Governors.

User Experience Advisory Group

The UX Advisory Group made recommendations on improvements to the PageTurner application, including the addition of the version date and new labeling to clarify when full-PDF download is available. The group collaborated with staff at Michigan and the Communications Working Group on a blog post about creating collections in HathiTrust, and began to focus attention on a project to redesign the HathiTrust home page.

User Support Issues Types	Total
Content	**831**
Quality	778
Non-partner Digital Deposit	5
Collections	28
Cataloging	**194**
Access and Use	**639**
Copyright	377
Permissions	75
Takedown	6
Print on Demand	2
Inter-library loan	2
Full-PDF or e-copy requests	83
Datasets	11
Data Availability and APIs	7
Reuse of content	10
Web applications	**89**
Functionality problems	27
Problems with login specifically	3
General questions about login	15
Partners setting up login	13
Usability issues	6
Feature requests	6
Partner Ingest	**15**
General	**578**
Partnership	40
Infrastructure	4
Miscellaneous	534
Total	**2,346**

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

User Support Working Group

A summary of the issues received by the User Support Working group is shown in the table above. The working group made several improvements to its workflow for handling user inquiries - those related to content quality especially, but in other areas as well. The group worked on recommendations for a future structure and process for responding

to user inquiries, which is one of the responsibilities specified in its charge.

Special Initiatives

Bibliographic Data Management

California Digital Library (CDL) staff loaded all records that are present in the current bibliographic management system at the University of Michigan into Zephir, the new HathiTrust bibliographic management system, which is now in final stages of development. The CDL team performed load testing during the ingest of records and worked to address discrepancies between records in the two systems. Staff created prototype exports of data that will be used to support the HathiTrust bibliographic catalog and "Hathifiles" inventory files. CDL worked with Michigan to finalize a record submission standard, and began to develop documentation and guidelines for submitting bibliographic records to Zephir, and documentation of the reports to be provided to institutions when records are loaded. Details about the submission standard, and additional information to be requested when records are submitted to HathiTrust, will be forthcoming.

HathiTrust Research Center (HTRC)

The HTRC completed all the agreements necessary to receive Google-digitized materials from the HathiTrust repository. Staff from Indiana University worked with staff at the University of Michigan to begin transferring OCR text files for the more than 3 million public domain volumes in HathiTrust to the HTRC.

The HTRC released a report on its activities from October 2011 to March 2012, detailing a variety of significant technical accomplishments, outreach activities, and strategic initiatives. The HTRC will be holding an "Uncamp" at Indiana University this September. Please visit the HTRC webpage and view the report above for further information on HTRC activities.

IMLS Quality

The IMLS Quality grant team completed page-level review (sampling within each volume) of three 1,000-volume samples from HathiTrust and reported initial findings (see the links under Quality Review on the results page of the project website). The team developed a new whole-volume review interface to facilitate detection of errors that affect the entire volume (such as missing, duplicate, and out-of-order pages) as well as the severity of page-level errors. Project staff reviewed the first two 1,000-volume samples in this new interface in order to be able to compare results with page-level review.

Project staff completed physical review of ~90% of volumes in the first 1,000-volume sample and 60% of the second 1,000-volume sample to investigate correlation of physical book characteristics with errors in digitized volumes.

The grant team is beginning a sub-study to better describe errors in illustrative content in digitized volumes and has begun to shift focus to the final, user research portion of the grant.

mPach

Staff at the University of Michigan worked on modifications to the HathiTrust PageTurner to display JATS XML and developed the first iteration of a tool that creates valid JATS XML from simple DOCX files. Staff also worked on specifications for a Submission Information Package for mPach content, began development of wireframes for the mPach Dashboard module (see a description of all mPach modules), and composed design principles and requirements, as well as a project timeline.

Repository

Advanced Full-text Search

Staff at the University of Michigan released several new advanced search features, including operations to search bibliographic metadata in combination with full-text search, limit results to specific publication years, languages, and original formats, revise advanced searches, and search with greater Boolean complexity. These features are described in the Update on April 2012 Activities and a Perspectives from HathiTrust blog post.

Michigan staff undertook work to improve indexing of volumes in Chinese, Japanese and Korean, and improve relevance-ranking of results.

Staff at California Digital Library made significant progress on the development of a spelling-suggester feature for full-text search.

Automatic Partner Login

Staff at the University of Michigan developed functionality to allow users from partner institutions to be "automatically" logged in to HathiTrust when fol-

lowing links from local institutional catalogs or other resources.

Changes to Tab-delimited "Hathifiles"

Michigan staff added 5 new fields to HathiTrust's tab-delimited inventory files (view the files or a description). The new fields include publication date, publication location, language, bibliographic format, and an indication of whether or not a volume has been identified as a U.S. federal government document.

Data API

Staff at Michigan developed security enhancements that, beginning October 1, will require developers to use OAuth 1.0 access keys to access the Data API and sign URLs passed to the API with a secret key. Staff also developed a Web client that employs a user's login credentials as proxy for the keys (users can sign up for a University of Michigan "Friend Account" to login). Users can register for keys or use the Web client by visiting http://babel.hathitrust.org/cgi/htdc. It is currently *possible* to use the keys and Web client; use will be *required* beginning October 1, 2012.

Also beginning October 1, 2012, the host "services.hathitrust.org" will be taken out of service. Calls to the Data API will need to use URLs such as the following (note the additional "cgi" in the path):

```
http://babel.hathitrust.org/cgi/htd/meta/mdp.39015019203879
```

rather than

```
http://services.hathitrust.org/htd/meta/mdp.39015019203879
```

On May 1, support for legacy Data API URLs in the following form was removed:

```
http://services.hathitrust.org/api/htd/pathinfo-arguments
```

URLs should be submitted to the API according to the current Data API schema without the "api" path element

```
http://services.hathitrust.org/htd/pathinfo-arguments
```

Michigan staff deployed a Data API security monitoring and reporting script that runs on a daily basis.

Logging usage of in-copyright materials

University of Michigan staff implemented processes to track accesses to in-copyright works in cases where access is permitted. The new processes provide a means for HathiTrust to detect problematic activity such as bulk downloading operations, which may, for example, indicate a compromised user account.

PageTurner

Michigan staff made a number of adjustments and improvements to the PageTurner application and interface. These included:

- A volume version date. A version for each volume is now displayed in the interface, indicating the date the volume was last updated. Volumes are updated when improvements such as higher quality or more complete scans are available.
- Better labeling to indicate when volumes are available for full-PDF download.
- Fixing a bug in the bibliographic metadata emitted in the PageTurner that had prevented license information from being included appropriately.
- Enhancing the access control mechanism for items that are public domain in the United States to better detect whether a user is on U.S. soil when access is proxied.
- Updates to the back-end process by which user feedback is submitted from HathiTrust applications to the HathiTrust ticketing system, including functionality to detect multiple tickets that are submitted for the same volume or record.
- Completing development necessary to allow users to embed works from HathiTrust in local Web pages (see http://www.hathitrust.org/embed for instructions).

New Web servers and load balancers

Michigan staff replaced two Web servers in the Michigan repository instance and moved to a new system of load balancing between the Indiana and Michigan repository instances. Load balancing is used routinely to mask maintenance or upgrade processes that require individual servers or an entire site to be taken offline.

New and Replacement Storage

Michigan staff installed new storage at the Indiana and Michigan sites. The storage was purchased

to accommodate partner projections for content in 2012 and replace storage scheduled for retirement.

Print on Demand Reports

Reports of volumes in HathiTrust that are available for print on demand are available at http://www.hathitrust.org/pod_reports. A new report will be posted on the first of each month.

Web Hosting Infrastructure Changes

Michigan staff moved HathiTrust's Drupal-based informational website and VuFind-based catalog from their initial hosting environments on Michigan library infrastructure to dedicated HathiTrust infrastructure. This move consolidates, and will greatly simplify HathiTrust Web developemnt

Papers and Presentations

All papers and presentations are listed at http://www.hathitrust.org/papers.

HathiTrust is an international partnership of academic and research institutions dedicated to ensuring the preservation and accessibility of the vast record of human knowledge. The partnership owns and operates a digital repository containing millions of public domain and in-copyright volumes, digitized from partnering institution libraries and other sources. The preserved volumes are made available in accordance with copyright law as a shared scholarly resource for students, faculty, and researchers at the partnering institutions and as a public good to the world community. For more information, visit HathiTrust.org.

You can follow HathiTrust on Facebook and Twitter