HathiTrust Digital Library

SPECIAL EDITION - 2012 Year in Review

January 28, 2013

2012 brought to a close the initial 5-year charter period that HathiTrust was granted by its founding institutions. 5 years later, the collaborative is stronger than ever. More than 70 academic and research institutions from around the world participate in HathiTrust, supporting a digital repository of 10.6 million volumes and a host of shared activities, all geared toward the provision of greater access to the scholarly and cultural record, more secure preservation, and greater research opportunities for our constituencies than we have ever had before. As we launch into a new year, and a new stage of HathiTrust, it is worthwhile to reflect on our progress and achievements in 2012. These include:

- A significant legal victory and affirmation of Fair Use in the case of Authors Guild v. HathiTrust
- Many new partners and a new governance structure
- A steady stream of technological improvements and enhancements
- Development of new services and infrastructure
- Continued engagement with our community in the form of presentations, discussions, and the HathiTrust Research Center Uncamp

A recap of activities in these areas and more can be read below.

About HathiTrust

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.

Highlighted Achievements and Activities

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

Special News

In a decisive victory for libraries and Fair Use, a lawsuit brought against HathiTrust and several participating libraries by the Authors Guild et al. was dismissed. Information about the lawsuit, including responses and analysis from around the Web, can be found on the HathiTrust website.

New Partners

HathiTrust grew from 66 to 78 partner institutions in 2012. New institutions include:

- Brandeis University
- Carnegie Mellon University
- Florida State University
- Iowa State University
- Kansas State University
- Syracuse University
- University of Delaware
- University of Kansas
- University of Vermont
- Vanderbilt University
- Virginia Tech
- Washington University

New content

HathiTrust partners contributed 623,613 volumes to the repository in 2012. 566,044 of these are in the public domain. The University of Florida and Boston College were new contributors in 2012. Many others contributed additional content, as shown in the table near the end of the update.

Locally-digitized Content

Over the course of 2012, HathiTrust interacted with nearly a dozen institutions regarding ingest of locally-digitized content. We released a first iteration of ingest tools to aid institutions in validating and packaging locally-digitized content prior to submission to HathiTrust. We revised documentation surrounding the tools based on feedback from institutions, and we also began to explore with institutions what the next iteration of the tools would look like. If you are using the tools now, think you might in the future, or are interested in more information, we encourage you to join our HathiTrust Ingest Google Group to participate in discussions.

Organization, Working Groups, and Committees

Board of Governors

HathiTrust took bold steps in establishing a new governance model, seating a new Board of Governors, establishing an Executive Committee and Executive Committee officers, and drafting a set of bylaws. The bylaws will be put forward to the partnership for voting in early 2013.

Collections Committee

The Collections Committee completed a report on handling of duplicate volumes in HathiTrust, recommending that HathiTrust retain all duplicate copies for the time being, with periodic assessment.

Communications Working Group

The Communications Working Group released announcements related to HathiTrust's achievement of 10 million volumes, the new Board of Governors, and the Authors Guild lawsuit. The group also produced a new Resources page for HathiTrust, launched a Pinterest account, coordinated a survey of partners to receive input on the next iteration of partner training sessions, and, in collaboration with the UX Advisory Group, created a blog post on collections in HathiTrust.

User Experience Advisory Group

The User Experience Advisory Group consulted on improvements to the HathiTrust PageTurner, including the addition of a version date for volumes, updated messages regarding download of PDFs, and a new landing page for volumes that are restricted from reading due to copyright, but are nevertheless full-text searchable. The UX Advisory Group also provided feedback on a new site-wide redesign currently in development.

User Support Working Group

The User Support Working Group submitted recommendations to the Board of Governors on User Support going forward from 2012. A summary of the User Support issues received in 2012 is given at the end of the review.

Special Initiatives

Accessibility

HathiTrust completed the first phase of improvements to enhance the accessibility of HathiTrust Web applications. With a few minor exceptions that will be addressed in the second phase, HathiTrust interfaces are now compliant with Web Content Accessibility Guidelines (WCAG) 2.0, Level A.

Bibliographic Corrections

HathiTrust accepted and released a new policy on bibliographic corrections: http://www.hathitrust.org/bib_metadata_correction.

Government Documents Registry

HathiTrust initiated a project to build a comprehensive registry of U.S. federal government documents.

Out of Print and Brittle

HathiTrust began offering lawful access to digital copies of works that are out of print, when print copies owned by partner institutions are brittle or missing. More information is available at http://www.hathitrust.org/out-of-print-brittle.

Projects

Bibliographic Data Management

HathiTrust made progress toward the migration of bibliographic data management from the University of Michigan to the California Digital Library's Zephir system. Major activities in 2012 involved improving record loading processes in Zephir, syncing information between Zephir and other HathiTrust systems, exporting data from Zephir for use in the HathiTrust catalog and "hathifiles", development of new bibliographic metadata standards, development and testing of bibliographic record submission processes with current HathiTrust depositors, and progress toward a Zephir service-level agreement. Migration to Zephir is expected to occur in 2013.

HathiTrust Research Center

In the early part of the year, the HTRC completed the agreements necessary to receive public domain data from the HathiTrust Repository. It also began to install systems for discovering, retrieving, correcting, and performing computation on OCR text of digital volumes. HTRC software and tools had their first public demonstration at an enthusiastic and widely successful HTRC "UnCamp" in September, attended by 130 researchers, developers, and librarians from HathiTrust member and non-member institutions. Resources about HTRC software and tools, including presentations, session materials, twitter analysis, and pictures from the UnCamp, are available on the HTRC wiki. A video produced from the event is available at http://www.hathitrust.org/htrc.

IMLS Quality Grant

The grant project team concluded all data gathering activities, including digital review of four 1,000-volume samples of volumes from HathiTrust, physical review of nearly all volumes in one sample and more than half of the volumes in a second sample (to investigate correlation between physical condition and digitization quality). Two of the samples underwent review more than once, as a new methodology was introduced to discover "whole-volume" errors such as missing and duplicate pages. In the coming months, as part of a no-cost extension, members of the team will conduct user studies to evaluate the results of the quality review performed on the sampled volumes. Initial findings from studies undertaken in the grant can be found at the links below. More results will be posted on the project website as analysis concludes and as articles containing the results are published throughout the coming months.

- Inter-rater reliability
- Distribution of error
- Co-occurrence of error
- Physical characteristis of original source

mPach

mPach is a system under development by the University of Michigan Library to publish open access born-digital journal content, along with accompany data and media files, directly into HathiTrust for perpetual access and preservation. Work in 2012 focused on refining the project's design principles and requirements and system architecture, establishing a timeline for the project, and designing and developing mPach modules and associated workflows to a) create archival XML in JATS format from DOCX files and b) deliver the resulting XML and supplementary files through HathiTrust applications.

Repository

Development in 2012 included the following:

New Functionality / Application Changes

- Functionality making it possible to automatically direct users accessing HathiTrust to login.
- Data API
  - New security measures requiring that requests to the API to be signed by an access key provided by HathiTrust. See HathiTrust Data API.

  - Added functionality to deliver PDFs for Expresso Book Machines on the ExpressNet network, and to return watermarked image derivatives in JPEG and PNG formats at a range of resolutions (the API previously only delivered the archival TIFF and JP2 formats).
- The creation of a “tombstone” in cases where volumes are deleted from HathiTrust.
- Full-text Search
  - Introduction of advanced searching options, including support for complex Boolean searches. See Search Tips.
  - Improvements to indexing of Chinese, Japanese, and Korean (CJK) language materials and improved relevance ranking of search results.
  - Significant work was undertaken to improve the relevancy ranking of search results in general, and to introduce a spelling suggestion feature for search queries.
- Imgsrv (a Web application that serves derivatives of master images to Web applications such as the PageTurner)
  - Enhancements to deliver HTML derivatives of born-digital content (in support of mPach and JATS XML).
  - Modifications to PDF construction to optimize the size of PDFs.
- Logging
  - Modifications to track uses of in-copyright works in the special cases where access is permitted.
- PageTurner
  - Movement of the default view from “Classic” (one page at a time) to Scrolling.
  - Addition of volume version (last updated) date. Items are updated when improvements such as higher quality or more complete scans have been made.
  - Functionality to embed HathiTrust volumes in external Web pages.
  - Modifications to make it easier for PageTurner to support display of new formats (such as JATS XML)
  - Improvements to special messages that display in circumstances when lawful access to an in-copyright volume is granted.
  - Modifications to the landing page for Limited (search-only) volumes.
- Tab-delimited “hathifiles”
  - Added new columns for publication date, publication location, language, bibliographic format, and whether or not a volume has been identified as a U.S. federal government document.
- Added support for the Shibboleth “library-walk-in” attribute to allow in-library guests to have certain member privileges when accessing HathiTrust. See HathiTrust Shibboleth.
- Website Redesign
  - The University of Michigan Library User Experience Department completed mockups for a site-wide redesign of the HathiTrust website, including a new home page. Significant steps were taken to implement the design, including work toward the development of a unified framework of Cascading Style Sheets (CSS) for HathiTrust applications.

Infrastructure Changes

- Completion of the first full repository-wide upgrade of metadata for HathiTrust objects, primarily involving PREMIS metadata but also other portions of the METS file. Documentation of PREMIS in HathiTrust is available at http://bit.ly/14eiqOJ. HathiTrust METS profiles and example METS files are available at http://www.hathitrust.org/digital_object_specifications.
- Retiring of two Web servers and installation of

two new Web servers (there are 2 Web servers at each HathiTrust storage instance - 4 Web servers total - that handle user traffic to HathiTrust).

- Installation of new Web load balancers.
- Purchasing and installation of new storage (based on partner 2012 content projections); replacement of storage per regular storage retirement practice.
- Migration of HathiTrust's Drupal-based informational website and VuFind-based catalog from University of Michigan Library infrastructure to HathiTrust infrastructure.
- Gathering of requirements for, and arrangements to purchase new high-performance storage for full-text search, to supplement efforts to improve performance and relevance-ranking of search results.
- Removal of sensitive information from application code for increased security.
- Modification of PageTurner to retrieve bibliographic data from the HathiTrust catalog's VuFind Solr index, rather than Michigan's bibliographic database.
- Migration of mapping information (HathiTrust namespaces to depositing institutions) to a database table for easier maintenance.
- Migration of access control list for special uses of in-copyright materials (e.g., for copyright or quality review purposes) to a database table to streamline maintenance.

Papers and Presentations

All HathiTrust papers and presentation can be accessed at http://www.hathitrust.org/papers.

Manual Copyright Review Update

Copyright reviews and determinations conducted as part of CRMS-US and CRMS-World.

Copyright Determinations

	2012		**Overall**	
	Public Domain	All Determinations	Public Domain	All Determinations
CRMS-US	41,268	79,817	119,822	219,874
CRMS-World	13,445	23,519	14,202	28,795
Total	54,713	103,336	135,777	248,669

HathiTrust Digital Library

SPECIAL EDITION - 2012 Year in Review

New Content

Total Volumes Added	2012	Overall
Boston College	1,842	1,842
Columbia University	214	64,390
Cornell University	31,745	415,435
Duke University	1	4,523
Harvard University	182,545	235,985
Indiana University	8,161	195,073
Library of Congress	311	89,722
North Carolina State University	0	3,196
Northwestern University	7,073	12,722
New York Public Library	121	259,574
Penn State University	1,815	44,732
Princeton University	1,972	251,651
Purdue University	43,741	44,629
Universidad Complutense	3,233	111,901
University of California	95,601	3,383,255
University of Chicago	16,112	26,720
University of Florida	2,008	2,008
University of Illinois	90,384	104,887
University of Michigan	105,235	4,609,836
University of Minnesota	13,973	104,212
University of North Carolina - Chapel Hill	1	8,088
University of Wisconsin	23,046	550,380
University of Virginia	3,403	50,799
Utah State University	71	117
Yale University	4	23,678
Total	632,613	10,599,355

Public Domain (~31% of total)

Total*	566,044	3,278,630

*Includes works opened via copyright review and rights holder permissions.

HathiTrust Digital Library

SPECIAL EDITION - 2012 Year in Review

User Support Working Group

User Support Issues	2012	2011*
Content	**1,038**	**962**
Quality	971	905
Non-partner Digital Deposit	10	6
Collections	57	45
Cataloging	**806**	**238**
Access and Use	**969**	**898**
Copyright	811	500
Permissions	158	51
Takedown	11	11
Print on Demand	8	12
Inter-library loan	24	12
Full-PDF or e-copy requests	198	175
Datasets	38	25
Data Availability and APIs	9	14
Reuse of content	25	27
Web applications	**220**	**229**
Functionality problems	61	66
Problems with login specifically	9	19
General questions about login	21	21
Partners setting up login	21	23
Usability issues	20	30
Feature requests	24	37
Partner Ingest	**40**	**25**
General	**832**	**316**
Partnership	126	83
Infrastructure	4	4
Miscellaneous	702	229
Total	**3,830**	**2,668**

*Statistics in 2011 are from March thru December; 2012 is January thru December

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

2012 Most-accessed volumes

Title	Count
Bradshaw's handbook for tourists in Great Britain & Ireland. Sec 1, 1866	43,175
Investigation of Korean-American Relations, 1978	40,543
Quicksand, by Nella Larsen	25,520
The Congressional cook book; favorite national and international recipes with special articles by eminent government authorities.	14,164
Bradshaw's handbook for tourists in Great Britain & Ireland, Sec 4, 1866.	11,864
Strong medicine, by Blake Donaldson.	10,961
Bradshaw's handbook for tourists in Great Britain & Ireland, Sec 2, 1866	10,424
Perfume and flavor materials of natural origin, by Steffen Arctander	9,832
The Continuations of the Old French Perceval of Chrétien de Troyes, Vol. 3, Pt. 2. Ed. William Roach	9,455
The age of revolution, 1789-1848, by E. J. Hobsbawm.	8,537
The Tosa diary, tr. from the Japanese by William N. Porter, 1912	8,405
Darkwater; voices from within the veil, by W. E. B. Dubois	7,580