HathiTrust Digital Library

SPECIAL EDITION - 2014 Year in Review

From the Executive Director

February 2, 2015

We're proud to present our annual Year in Review to you. Since I joined HathiTrust in May, I've had a great time visiting some of you personally to discuss some of what is covered here, and to hear your thoughts and ideas for our partnership's growth. As you can see here, we've passed some significant milestones, and expect the coming year to be exceptionally productive. Now in our seventh year, and ten years after the start of the Google-Library project that preceded us, we hold over 13 million volumes from the collections of our members. Thanks in part to the institutions taking part in the Copyright Review Management System project, we are close to having 5 million of these available either as public domain materials or licensed for access by the rightsholder. I want to especially greet and welcome our 14 new members listed below (including one in Lebanon), which brings us to 103 members overall. Having prevailed in the Second Circuit Court of Appeals in our conflict with the Authors Guild, we enter 2015 with the remainder of the dispute resolved. We can now focus on core activities that advance the public good and help our member libraries better serve their users and manage their collections.

Many long planned efforts are beginning to bear fruit. You can expect to see more action in our efforts to expand and enhance access to US federal government documents collections, and we will make the first releases of the Registry of Federal Documents later this year. The Print Monographs Archive Planning Task Force will present their recommendations for implementing this program and those will be shared with you all. The HathiTrust Research Center is poised to expand their services in the coming year, offering advanced researcher support services as well as training and services available to member libraries. 2015 will also mark the first of what will now be an annual election of new members to the Board of Governors, and the first major turnover of membership on the Program Steering Committee. Details on the appointment process to PSC will be announced this spring, and nominations for election to the Board of Governors will open later in the year.

Thanks to everyone who has contributed time, ideas, and energies towards making HathiTrust a stronger organization. We'll continue to rely on member participation to steer and carry out our necessary work. I hope your year has gotten off to as good a start as mine.

-- Mike Furlough

Highlighted Achievements and Activities

Details on each item can be found in the monthly updates from 2014, available at http://www.hathitrust.org/updates.

Rulings in Authors Guild Lawsuit Appeal

The U.S. Second Circuit Court found in favor of HathiTrust in the Authors Guild lawsuit against us. In early January, the remaining plaintiffs resolved their dispute with the HathiTrust members named in the case, and the case was dismissed by the court. View HathiTrust statements on the appeal and resolution of the lawsuit.

New Executive Director

HathiTrust announced the appointment of Mike Furlough as the Executive Director of HathiTrust. Mike began on May 19.

First Annual Member Meeting

HathiTrust held its first annual Member Meeting on October 10, 2014. Meeting Notes, presentations,

and other documentation from the meeting are posted online, as is a blog post containing reflections on the meeting by Executive Director Mike Furlough.

New Partners

13 institutions joined HathiTrust in 2014, bringing the total number of members to 103:

- American University of Beirut
- Case Western Reserve University
- Florida State University System
- Georgetown University
- Georgia Tech*
- Montana State University
- Mount Holyoke College
- Northeastern University
- Oklahoma State University
- Rutgers University
- Texas Tech University
- University of Maine
- University of New Mexico
- University of Texas System

* Georgia Tech joined in early 2015

New Content

HathiTrust members and other institutions contributed 2,121,955 volumes to the repository, surpassing 11 million volumes in February 2014 and 13 million volumes in December 2014. 1,327,126 of the new volumes, and nearly 5 million overall, are in the public domain.

New contributors included Emory University, the Getty Research Institute, Keio University, Knowledge Unlatched, McGill University, The Ohio State University, the Sterling & Francine Clark Art Institute, and the University of Alberta. New locally-digitized content was received from the University of Illinois, Yale University, Boston College, and Columbia University. Contributions of all content are shown in the table at the end of the update.

Governance and Working Groups

Board of Governors

2015 Budget

HathiTrust members voted in December to accept the proposed 2015 total budget and fees.

Board Changes

Indiana University's representative Brad Wheeler stepped down from the HathiTrust Board of Governors in May and was replaced by Brenda Johnson. Later Indiana designated Carolyn Walters to serve, following the departure of Brenda Johnson to University of Chicago Library.

Pat Steele, of the University of Maryland, stepped down from the Board of Governors; the Board will appoint a replacement as specified in the HathiTrust bylaws.

Effective January 1, 2015, the new officers of the Executive Committee are:

- Chair, Board of Governors: Richard Clement, University of New Mexico
- Chair-elect/treasurer: Lizabeth (Betsy) Wilson, University of Washington
- Past Chair: Sarah Michalak, Univerisity of North Carolina, Chapel Hill
- Chair, Program Steering Committee: Bob Wolven, Columbia University
- Ex-officio: Mike Furlough, Executive Director, HathiTrust

Decisions and Activities

Major decisions and activities by the Board included:

- Allocation of nearly $1,000,000 over four years to support the HathiTrust Research Center (HTRC), based on a proposal from the HTRC executive leadership team, and

pending the finalization of schedules for service development and reporting.

- Allocation of an additional $115,000 to extend staffing in support of development of the Government Documents Registry.
- Approval of the 2015 annual budget for vote by the membership.
- Approval of the first annual HathiTrust Membership Meeting, held in Washington, DC on October 10.
- Appointment of 2 new members to the Program Steering Committee: Robert McDonald, Associate Dean, Library Technologies, Indiana University, and Chris Freeland, Associate University Librarian, Washington University in St. Louis.

Orphan Works Roundtable

Sarah Michalak (then Chair of the HathiTrust Board Executive Committee), Mike Furlough, and Melissa Levine, Lead Copyright Officer at the University of Michigan Library, participated in a Roundtable discussion organized by the U.S. Copyright Office on Orphan Works and Mass Digitization. Comments on the discussion submitted by HathiTrust are available at http://www.hathitrust.org/comments-orphan-works-mass-digitization.

Program Steering Committee

Major activities of the Program Steering Committee included:

- The charging and appointment of a Government Documents Initiative Planning and Advisory Group, Collections Committee, Rights and Access Working Group, and Print Monographs Archive Task Force, and the charging of a Zephir Advisory Group. Reports or initial deliverables for some of these groups are expected in early 2015.
- Identification of four broad areas of planning

and activity in 2015: Non-Text Formats; Quality Assurance and Validation; Services for Users who have Print Disabilities; and Metadata Strategies and Policies. Planning briefs on these topics were made available for the 2014 Member Meeting and are available at http://www.hathitrust.org/psc_planning_briefs.

User Support Working Group

Statistics on user support issues received in 2014 are available in a table at the end of the update.

Projects

Copyright Review

In January 2014, project staff completed copyright review of all works in HathiTrust to that time that were eligible for review under the Copyright Review Management System-United States project. More than 160,000 of the 300,000 works reviewed in this project were found to be in the public domain and made available through HathiTrust.

The University of Michigan received a third grant award from the Institute of Museum and Library Services for copyright determination work. A portion of the grant will include exploration of sustainability options with HathiTrust.

In September, HathiTrust began focusing exclusively on reviews of works in the CRMS-World project in order to meet that project's goals. During 2015 a new strategy for handling works that fall outside the project's scope, including special requests, as a part of planning for CRMS sustainability and business planning.

A summary of the determinations from HathiTrust copyright review activities from 2014 is given belo-See CRMS-US and CRMS-World for further information. CRMS-US and CRMS-World are projects generously funded by the Institute of Museum and Library Services.

SPECIAL EDITION - 2014 Year in Review

	2014		Overall	
	Public Domain	All	Public Domain	All
CRMS-US	9,526	12,440	167,617	317,852
CRMS-World	44,945	84,032	88,759	168,371
Total	54,471	96,472	256,376	486,223

Government Documents Initiative

General

More than 40 institutions submitted bibliographic records for US federal government documents in response to a call for records from HathiTrust to better understand the scope of the corpus of US government documents, and the portion that have already been digitized. This work is a part of larger HathiTrust initiative to expand and enhance access to US federal government documents.

Registry

An effort to build a registry of US federal government documents is another facet of this larger initiative. Work on the Government Documents Registry focused on the development of functional objectives for the Registry, and the development of strategies and processes to 1) identify duplicate records and understand relationships between different record sets and 2) identify gaps in government documents holdings, with an eye toward being able to determine the comprehensiveness of certain sets of materials in the HathiTrust repository.

HathiTrust hired a new Applications Developer, Josh Steverman, who will be the primary developer of the registry.

HathiTrust Research Center (HTRC)

Major activities included:

- Awarding 4 recipients of project awards for the Workset Creation for Scholarly Analysis (WCSA) project funded by the Andrew W. Mellon Foundation.
- The alpha release of a page features dataset.
- Receipt of a $324,84 grant award from the National Endowment of the Humanities for the project "Exploring the Billions and Billions of Words in the HathiTrust Corpus: HathiTrust+Bookworm".
- Release of a Request for Proposals for Advanced Collaborative Support (ACS), a newly launched service of the HTRC. Proposals were due on January 8th, 2015 and awardees will be announced soon. A second round of requests will be issued in 2015.
- Planning for offering 'non-consumptive' access to in-copyright volumes in the HathiTrust repository.
- Significant progress toward the release of version 3.0 of the HTRC. New features include the HTRC Data Capsule (a secure environment for performing computation on data from HathiTrust), an improved user experience and single sign-on services (except for the Data Capsule). Version 3.0 in in beta testing through January 30, 2015, and is available at https://htrc2.pti.indiana.edu/. Please send feedback to htrc-tech-help-l@list.indiana.edu. You can also sign up to HTRC email lists to receive updates and announcements.

Save the date! The third annual HTRC UnCamp will be held at the University of Michigan, March 30-31, 2015. Information on registration and other details will posted soon at http://www.hathitrust.org/htrc_uncamp2015.

mPach

Michigan and HathiTrust staff are currently reviewing expected timelines and deliverables. University of Michigan staff made improvements

to mPach workflow modules designed to normalize and prepare born-digital publications for ingest into HathiTrust. Staff also focused on user interface issues, with specific attention to accessibility.

Repository Updates

Development in 2014 included the following:

New Functionality / Application Changes

Access, Authentication and Authorization

- Modified Web applications to use authenticated members' Shibboleth entityID to establish their institutional affiliation, rather than eduPerson-ScopedAffiliation. This was done in order to facilitate proper identification when a user has multiple affiliations.
- Developed and deployed a system for managing users who have special access to in-copyright materials (e.g., for copyright or quality review).
- Added functionality to automatically expire access keys that are configured to allow special access to content via the HathiTrust Data API.
- Began to add support for "access profiles", which will associate materials with the same access and use restrictions together, facilitating the management of access control parameters.
- Made enhancements to the way authentication and access are handled for institutions that are members of consortia.

Bibliographic Data Management

- The California Digital Library had a successful first year operating Zephir, the bibliographic management system it created and manages for HathiTrust. CDL loaded 2,739,848 new or updated records from HathiTrust members and other contributors into Zephir in 2014.

Collection Builder Application

- Improved Collection Builder performance when sorting lists of items in large personal collections; improved the accuracy of sorting multipart monograph and serial volumes when date information is available.
- Improved end user messaging about the status of items in personal collections, providing separate notifications for items that are in the queue to be indexed, versus those that will never be indexed because they have been deleted from the repository.
- Added functionality to allow collection owners to create multiple collections that have the same name.

Full-text search

- Conducted significant research, development, and testing to improve the relevance ranking of full-text search results. This included research into indexing volumes into a configurable number of "chunks", and investigating the use of the INEX Book Track 2007-2010 test collections to inform choices about relevance ranking algorithms.
- Undertook considerable investigation and development to prepare to use new high performance storage for full-text search services. Issues with storage software have delayed deployment and staff remain in regular communication with the storage vendor to address identified issues.
- Investigated performance issues for HathiTrust full-text search and testing of features under various high load scenarios.
- Performed significant work toward the migration of the Solr index from Solr 3 to Solr 4.
- Added features to support the indexing of JATS XML content.
- Corrected a problem in navigation of full-text search results. The link to the first page of results disappeared if the user navigated beyond a certain number of pages.

- Fixed a bug affecting indexing and full-text searching of an estimated 50% or more of Chinese and Japanese volumes. Searching of these materials is now significantly improved.
- Tested a spelling suggestion feature developed by the California Digital Library for future integration.
- Completed initial work to take advantage of planned changes in the indexing of volume publication dates.
- Tom Burton-West authored 3 blog posts in a series about "Practical Relevance Ranking for 11 Million Books": Part 1, Part 2, Part 3.

Google Analytics

- Updated Google Analytics to track the usage of HathiTust Collections in addition to individual items.
- Modified the configuration for Google Analytics to track uses of volumes (and searches within books) at the volume-level only rather than the page- and volume-level. This better reflects the way the Google Analytics data is being used, and aligns with Analytics' normal processing of heavily parameterized URLs.

ImageServer

- Re-architected the imgsrv application to more efficiently support the generation of derivative formats from a variety of content types (currently digitized books composed of page images and OCR, and in the future, born-digital materials formatted in JATS XML).
- Modified EPUB versions of volumes, delivered only in the HathiTrust mobile interface, to use HTML coordinate OCR when it is available.
- Prototyped new imgsrv capabilities for continuous text (e.g., JATS encoded materials without page breaks) in PageTurner.
- Configured applications (PageTurner, Collection Builder, bibliographic and full-text catalogs) to display thumbnail images in search results from local image files when thumbnails are not returned by the Google Books API.

Ingest

- Released a full-volume validation and packaging service for locally-digitized materials (see http://www.hathitrust.org/ingest_tools).
- Updated the use of quality metrics provided by Google in determining thresholds for content ingest.

PageTurner

- Staff at California Digital Library developed an "Embed this Book" feature that is now available in the "Share" section of the PageTurner sidebar. Users can copy the HTML for embedding either 1up or 2up views into websites and blogs.
- Fixed bugs and made improvements to the "search in this text" widget for navigating from one page of results to another.
- Released a new "skin" for the mobile version of PageTurner, updating the interface to use the common code base shared across the suite of HathiTrust web applications, and be compatible with modern mobile browsers.

Repository and Infrastructure Changes

Server Replacement

- Completed the replacement cycle for production web servers at the Michigan and Indiana repository instances.
- Ordered and installed replacement servers for HathiTrust full-text search infrastructure.

Storage Replacement Infrastructure

- Completed installation of new and replacement storage for 2014.
- Purchased and completed an early installation

SPECIAL EDITION - 2014 Year in Review

of approximately half of the new storage for the 2015 cycle. The storage was purchased to accommodate substantial repository growth this fall, which exceed earlier projections.

- Purchased and received remaining new and replacement storage for 2015.

Security

- Released statements on the "Heartbleed bug" and "Shellshock" bash vulnerability.

Updated Volume Identifiers

- Performed a one-time batch change to a set of approximately 320,000 volume identifiers. The affected volumes were ingested with an incorrect identifier due to a vendor issue. A full list of the updated identifiers is available at http://www.hathitrust.org/hathifiles. Any institutions or individuals that save links to HathiTrust volumes locally should update these identifiers to ensure working links. Please contact feedback@issues.hathitrust.org with any issues or questions.

Availability

Cumulative 12-month availability of repository access*: 99.964% (+0.015%)

Papers and Presentations

All papers and presentations from 2014 are listed at http://www.hathitrust.org/papers.

SPECIAL EDITION - 2014 Year in Review

Volumes added	2014	Overall
Boston College	900	3,263
Columbia University	8,359	73,395
Cornell University	72,574	510,065
Duke University	3,681	8,206
Emory University	52	52
Getty Research Institute	18,979	18,979
Harvard University	600,675	838,110
Indiana University	333,231	528,811
Keio University	90,094	90,094
Knowledge Unlatched	28	28
Library of Congress	19,168	108,892
McGill University	893	893
New York Public Library	6,465	294,835
North Carolina State University	0	3,196
Northwestern University	19,175	56,677
Ohio State University	61,129	61,129
Penn State University	319,513	387,717
Princeton University	1,098	252,808
Purdue University	2,793	47,488
Sterling & Francine Clark Art Institute	358	358
Texas A&M University	1,245	2,446
Universidad Complutense	5,221	117,235
University of Alberta	76,106	76,106
University of California	164,426	3,612,596
University of Chicago	13,341	51,976
University of Connecticut	4,637	4,637
University of Delaware	48	48
University of Florida	103	9,866
University of Illinois	205,156	318,131
University of Massachusetts	11,614	11,614
University of Michigan	46,720	4,712,752
University of Minnesota	28,782	144,717
University of North Carolina - Chapel Hill	0	17,025
University of Virginia	386	51,207
University of Wisconsin	4,851	560,775
Utah State University	0	117
Yale University	154	23,832
Total	2,121,955	13,000,076

Public Domain (~37% of total)

Total*	1,327,126	4,869,281

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

SPECIAL EDITION - 2014 Year in Review

User Support Issues	2014	2013
Content	**1,102**	**1,106**
Quality	966	987
Collections	136	119
Cataloging	**894**	**980**
Access and Use	**1,330**	**1,350**
Copyright	986	997
Permissions	105	107
Takedown	8	7
Print on Demand	3	4
Inter-library loan	22	16
Full-PDF or e-copy requests	203	216
Datasets	36	48
Data Availability and APIs	15	14
Reuse of content	41	48
Web applications	**270**	**299**
Functionality problems	107	89
Problems with login specifically	18	16
General questions about login	16	24
Partners setting up login	13	20
Usability issues	2	16
Feature requests	19	21
Partner Ingest	**144**	**66**
General	**853**	**713**
Partnership	100	100
Miscellaneous	753	613
Total	**4,252**	**4,114**

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

Most-accessed volumes

- The Human Figure, by John H. Vanderpoel.
- The Lion Monument at Amphipolis, by Oscar Broneer
- Quicksand, by Nella Larsen.
- Godey's Magazine, v.40-41, 1850.
- Consumption of the Lungs and Kindred Diseases, Treated and Cured by Kerosene, by Charles Oscar Frye.
- Quintus Curtius [History of Alexander], Vol. 1, with an English translation by John C. Rolfe.
- Modern California Houses: Case Study Houses, 1945-1962, by Esther McCoy.
- The Book of a Hundred Hands, by George Brant Bridgman.
- Quintus Curtius [History of Alexander], Vol. 2, with an English translation by John C. Rolfe.
- The Five Laws of Library Science, by S. R. Ranganathan.

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. You can follow HathiTrust on Facebook and Twitter.