HathiTrust Digital Library

Update On April Activities

May 9, 2014

Top News

HathiTrust Board

The HathiTrust Board of Governors will meet on May 9, 2014 in Columbus, Ohio immediately after the ARL meeting. The very full agenda includes a discussion about copyright in the context of mass digitization, consideration of a draft policy for limiting access to materials with personal sensitive information, and a review of HathiTrust's mission and goals, stemming from a 2011 Constitutional Convention ballot initiative.

Program Steering Committee Nominations

With Mike Furlough assuming the role of Executive Director of HathiTrust, a vacancy has opened up on the Program Steering Committee. The Board of Governors welcomes nominations to fill a two-year term on the PSC, commencing June 1, 2015. Nominations may be submitted by Member Representatives, but self-nominations are also welcome. Nominees should be at the AUL or senior management level to ensure an appropriate level of experience in the issues at hand.

Nominations should include the name, title, and institution of the nominee, and should be sent to Brian Schottlaender at becs@ucsd.edu with the Subject line "HT PSC Nomination." Nominations should be received by May 16, 2014.

The Program Steering Committee "reviews HathiTrust's development agenda, shaping initiatives and strategies for Board discussion and decision-making, and considering the implications of those initiatives for the future." The Committee meets virtually roughly biweekly, and may hold one to two in-person meetings per year. Much of the Committee's work is carried out through working groups or task forces formed to address specific issues and initiatives. For more information, see http://www.hathitrust.org/psc.

User Support Working Group Nominations

The User Support Working Group is seeking nominations for up to 2 new members. We are seeking staff who have expertise in providing general user support and those who have expertise in cataloging in particular. To submit nominations and for further information about the working group, please visit this form.

HathiTrust and DPLA

HathiTrust was the "top Hub" reported in DPLA's April Hupdate. View the full report for more details.

May Forecast

Test Solr's grouping and block-join functionality at scale for work on relevance ranking improvements.

Continue development of application to enhance administration of users with special access to materials.

Continue improvements and bug fixes in the "search in this text" feature.

Integrate the new Image Server capabilities for continuous text (e.g., JATS encoded articles without page breaks) into PageTurner.

HathiTrust Digital Library

Update On April Activities

Ingest

General

HathiTrust ingested new locally-digitized content from the University of Illinois and the University of Delaware, and corresponded with several other institutions that are preparing locally-digitized volumes for deposit.

California Digital Library (CDL) loaded 106,241 new or updated bibliographic records from partners into Zephir.

Working Groups and Committees

Program Steering Committee

The Program Steering Committee (PSC) continued the process of appointing working groups and committees. The roster for the Collections Committee is now complete, and includes Ivy Anderson (chair, and PSC liaison), Sharon Farb, Dan Hazen, Carmelita Pickett, Bryan Skib, Claire Stewart, Tom Teper, and Ann Thornton. Tom Teper will also serve as chair of the Print Monographs Archive Planning Task Force, now in the process of appointing members.

Elaine Westbrooks will chair a new Rights & Access Working Group. The PSC is seeking additional members to serve on this group, and will issue a formal call for volunteers. Those interested in serving on this group may also contact Bob Wolven, PSC chair, at wolven@columbia.edu.

The PSC has also begun working with the Zephir Management Team at California Digital Library to form a Zephir Advisory Group (ZAG). This group will draft and recommend new features and service enhancements for the Zephir metadata management system as well as metadata policies that: have strategic impact on and/or implications for the broader HathiTrust community and/or require resources beyond the current allotment for running Zephir. The ZAG will also serve in a consulting capacity when the CDL Zephir Operations Team drafts policies and procedures to address operational considerations.

Projects

Copyright Review

A summary of the determinations from HathiTrust copyright review activities in April is given below. See CRMS-US and CRMS-World, projects funded by the Institute of Museum and Library Services (IMLS), for further information.

Papers & Presentations

Seth Johnson, Bryan Smith, Kevin Hawkins, "mPach Integrated Publishing and Archiving of Journals in HathiTrust", April 1-2, 2014.

Jeremy York, "Getting the Most Out of HathiTrust: An Overview of Resources, Tools, and Services", Oakland University, April 10, 2014.

J. Stephen Downie, "HathiTrust Research Center: The Workset Creation for Scholarly Analysis (WCSA) Prototyping Project", University of Western Ontario, April 14, 2014.

Harriett Green, "HTRC Workshop" (slides | webinar), THATCamp, Gainesville, FL, April 24, 2014.

HathiTrust Digital Library

Update On April Activities

	April		Overall	
	Public Domain	All Determinations	Public Domain	All Determinations
CRMS-US	1,240	1,418	165,125	313,965
CRMS-World	3,413	7,561	55,632	110,057
Total	4,653	8,979	220,757	424,022

You can follow HathiTrust on Twitter or Facebook

Subscribe to email updates (via Google Groups)

Government Documents Registry

Project staff continued to review and refine automated methods to identify relationships between items, including duplicate volumes. Staff continued to review types of metadata that could be used to identify gaps in holdings, and to investigate strategies to determine the comprehensiveness of certain sets of materials in the repository.

HathiTrust Research Center

Harriet Green of the HTRC team (UIUC Library) gave a Web tutorial on HTRC at The Humanities and Technology Camp (THATCamp), 24-25 April, 2014 in Gainsville, Florida. The session provided an introduction to using the HTRC portal for basic text mining investigations. Attendees learned how to build a workset from the HTRC corpus, apply the textual analysis tools provided in the HTRC portal, and generate visualizations such as word clouds and statistical frequencies. View the slides or a recording of the presentation.

HTRC is making progress on an internal review of its security practices in anticipation of a review by HathiTrust. The process is engaging University security teams at both University of Illinois and Indiana University.

mPach

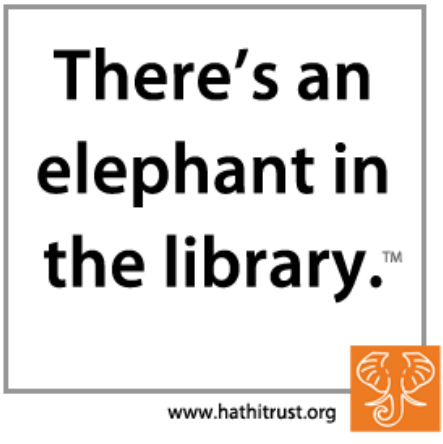
Staff began a major code refactoring of the Norm module, which will include moving much of the functionality into XSLT.

Development Updates

HathiTrust institutions performed the following work related to applications and Web interfaces:

Authentication and Authorization

Staff continued development of an application to simplify and enhance the administration of users (such as staff performing copyright review) who are permitted to have special access to restricted items. The application will support renewal, deletion, and automatic expiration of such users and is expected to be completed in June.



HathiTrust Digital Library

Update On April Activities

Full-text Search

Staff created two test indexes, each containing nearly 3 million volumes in three shards (sections), to test the scalability and performance of several features related to Solr's grouping functionality: field-collapsing, the use of docValues for faceting, and the Collapsing Query Parser. The first index uses individual pages rather than whole documents as the primary unit of indexing and includes approximately 900 million Solr documents. The second index uses 3,000 word "chunks' as the unit of indexing and includes approximately 90 million Solr documents. Performance testing will begin in May. A third index using Solr's block-join indexing will also be created in May. This work is part of an exploration of ways to improve Solr's relevance ranking for HathiTrust volumes.

Staff continued to troubleshoot network performance and stability issues with new high-performance storage for full-text search, and are currently engaged with the highest levels of both the storage and networking system providers.

ImageServer

Staff deployed a new version of the imgsrv application. The new version more effectively supports the generation of derivative copies from a variety of content types (currently digitized books composed of page images and OCR, and in the near future, born-digital articles formatted in JATS XML). EPUB versions of content, delivered only through the mobile interface, are now built using item HTML coordinate OCR when the HTML OCR is available. This provides a better user experience than EPUBs created with plain, unstructured OCR, which lack paragraph breaks.

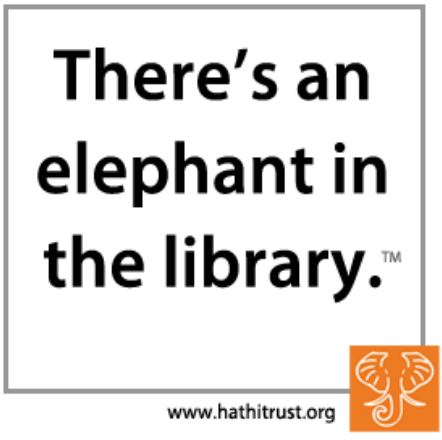
PageTurner

Staff made progress on improvements and bug fixes in the generation of the "search in this text" search results page.

Server Replacement Cycle

User Support Issues	April	March
Content	**154**	**181**
Quality	143	168
Collections	11	13
Cataloging	**187**	**203**
Access and Use	**142**	**212**
Copyright	101	144
Permissions	10	13
Takedown	1	2
Print on Demand	1	0
Inter-library loan	0	4
Full-PDF or e-copy requests	14	20
Datasets	3	2
Data Availability and APIs	2	1
Reuse of content	4	4
Web applications	**20**	**18**
Functionality problems	10	7
Problems with login specifically	3	2
General questions about login	1	2
Partners setting up login	0	0
Usability issues	0	0
Feature requests	1	2
Partner Ingest	**11**	**16**
General	**110**	**101**
Partnership	18	5
Infrastructure	0	0
Miscellaneous	92	96
Total	**624**	**731**

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



HathiTrust Digital Library

Update On April Activities

Total Volumes Added	March	Overall
Boston College	0	3,111
Columbia University	0	65,037
Cornell University	4,744	449,118
Duke University	0	7,258
Harvard University	0	237,435
Indiana University	4	195,651
Keio University	0	88,956
Library of Congress	0	107,929
New York Public Library	21	291,661
North Carolina State University	0	3,196
Northwestern University	4	37,643
Ohio State University	22	21,108
Penn State	1,271	79,199
Princeton University	0	251,710
Purdue University	0	44,698
Texas A&M	0	1,201
Universidad Complutense	0	112,148
University of California	17,646	3,494,244
University of Chicago	13	39,169
University of Florida	0	9,765
University of Illinois	1,225	135,691
University of Massachusetts, Amherst	0	9,411
University of Michigan	121	4,670,680
University of Minnesota	3	119,862
UNC - Chapel Hill	0	17,025
University of Virginia	0	50,821
University of Wisconsin	0	555,973
Utah State University	0	117
Yale University	0	23,678
Total	25,084	11,123,514

Public Domain (~33% of total)

Total*	57,185	3,739,276

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

Most-accessed volumes

- A family tour from ocean to ocean: being an account of the first amateur motor car journey from the Pacific to the Atlantic, whereby J.M. Murdock and family, in their 1908 Packard "Thirty" touring car, incidentally broke the transcontinental record, by J.M. Murdock.
- Quintus Curtius [History of Alexander], Vol. 2, with an English translation by John C. Rolfe.
- A program for the industrial and regional development of Peru : a report to the Government of Peru, 1960.
- Godey's magazine. v.40-41 1850
- The Tosa diary, by Tsurayuki Ki, translated by William N. Porter.
- The Human Figure, by John H. Vanderpoel.
- Quicksand, by Nella Larsen.
- Annotations upon the Holy Bible, Vol. 2, by Matthew Pool.
- History of wages in the United States from Colonial times to 1928, United States Department of Labor.
- Memoir of Colonel Benjamin Tallmadge, ed. by Henry Phelps Johnston.

Update On April Activities

Staff prepared system configurations and requested pricing for the first replacement cycle of HathiTrust's full-text search servers. The deployed infrastructure will be expanded by approximately 50% to accommodate increases in usage and a doubling of the number of volumes in the repository since the full-text search capability was initially launched in 2009. Installation of the new servers is expected to be complete by the end of June.

Spelling-suggester

Staff tested two language identification programs designed to be used on short documents such as search queries in order to assess whether information about language would be helpful as an additional clue for the spelling suggester. Preliminary results showed that the single best guess by the language identification program is not accurate enough to be useful.

Availability

Repository

Cumulative 12-month availability of repository access: 99.827%*

HathiTrust searching and book-viewing was unavailable for some users on Tuesday, April 1 from 3:22-3:33pm due to a manual error made in database configuration.

HathiTrust book-viewing was unavailable on Wednesday, April 30 from 1:35-1:45pm due to a manual error made in an application configuration file.

* Repository access refers to page viewing and full-text search functionality, i.e., user-facing applications. It does not refer to preservation or storage infrastructure, which is under continual operation.