HathiTrust Digital Library

Update On May/June Activities

June 24, 2015

Top News

Thank You and Goodbye to Jeremy York

Many of HathiTrust's members and partners have gotten to know our assistant director, Jeremy York, very very well. He has been involved in so many areas of HathiTrust's development over the last seven years that it's not possible to fully document his contributions. Jeremy has recently decided that the time has come for him to leave HathiTrust and pursue new activities. Ultimately he intends to return to graduate school and pursue a PhD in information studies. In the meantime, he's accepted a new position to manage a funded research grant on digital preservation and publicly funded data. His last day with HathiTrust will be June 23. You can read more about Jeremy's work and plans in this post.

Shared Print Report Released for Comment

The HathiTrust Print Monograph Archive Planning Task Force Final Report is now available for public review and comment. The Task Force recommends a series of actions to rapidly develop the program, which include 1) initiating discussions with archives and libraries to secure retention commitments for approximately 50% of the unique titles in HathiTrust, developing infrastructure through partnerships, seeking community comment, and establishing a Shared Print Operating Committee to continue planning, among others.

If you are planning to attend ALA, this program will be discussed during the Print Archive Network (PAN) meeting, sponsored by the Center for Research Libraries, on Friday June 25 at 9am (Check location and times at http://alaac15.ala.org/node/30197).The Board of Governors thanks the members of the Task Force for the tremendous effort they put into developing a thoughtful, coherent, and rich set of recommendations. The membership included: Tom Teper, Chair (University of Illinois); Clem Guthro (Colby College); Robert Kieft (Occidental College); Erik Mitchell (University of California, Berkeley); Jake Nadal (ReCAP); Jo Anne Newyear Ramirez (University of British Columbia); Matthew Revitt, Recorder (University of Maine); Matthew Sheehey (Brandeis, formerly Harvard University); Emily Stambaugh (California Digital Library); and Karla Strieb (Ohio State).

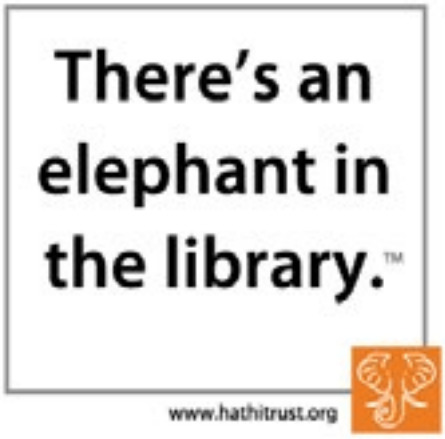
We plan to have other opportunities to discuss this program publicly and to gather feedback. At this time we welcome any comments or questions, specific or general, and you may send these to print-archive-comments@hathitrust.org.

Program Steering Committee Nominations

The Board of Governors welcomes nominations to fill a two-year term on the Program Steering Committee, commencing in August 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.

Editor's Note:

We will adjust the schedule o f Monthly Updates for the summer. This addition covers news from May and June. In July we will send out a mid-year report, and we will publish a July/August Update in Late summer.



Update On May/June Activities

June 24, 2015

Nominations should include the name, title, and institution of the nominee, and a short description of their qualifications for the appointment. Please send nominations to Melissa Stewart (mmstewa@hathitrust.org) with the subject line "HT PSC Nomination." by Friday July 17, 2015. Sarah Michalak, Past Chair of the Board of Governors and Chair of the Nominating committee is coordinating the nominations and appointment process.

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." TheCommittee 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 www.hathitrust.org/psc.

Zephir Advisory Goup Appointed

The Zephir Advisory Group has been formally appointed and begun their work. Membership includes:

Patti Martin, California Digital Library (Chair)
Gary Charbonneau, Indiana University
Timothy Cole, University of Illinois, Urbana-Champaign
Todd Grappone, UCLA
Chew Chiat Naun, Cornell University
John Mark Ockerbloom, University of Pennsylvania
Jonathan Rothman, University of Michigan
Ryan Rotter, University of Michigan
Katheryn Stine, California Digital Library

Their charge can be found at http://www.hathitrust.org/wg_zag_charge.

Getting Locally Digitized Content into HathiTrust

How can your library add locally digitized materials to HathiTrust? Aaron Elkiss of the University of Michigan has written a short post with some background and details of current practices.

Register for Webcasts for New Members

HathiTrust will host two webcasts later this summer to provide an overview for members who have recently joined. All members are welcome to attend these sessions, which will be held on July 29 at 4:00pm EDST and August 5 at 11:00am EDST.

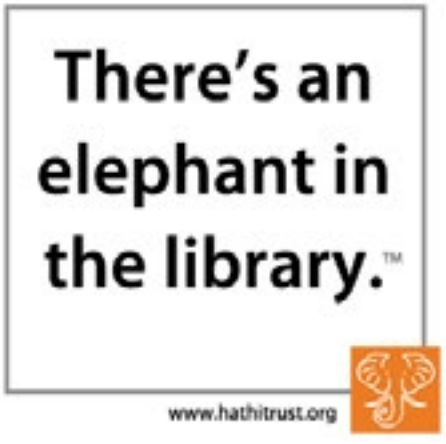
We ask that all attendees register, and urge you to organize group viewing sessions at your library. You may register here: http://goo.gl/forms/BZx1sWbRSW
Access information will be provided to registrants before these events.

HathiTrust on the Road

Staff from HathiTrust will be attending the following meetings and conferences in June. Please be in touch if you would like to meet with us during our travels:

Mike Furlough:

- American Library Association, San Francisco, CA. June 25-28.
- American Association on Higher Education and Disability, Minneapolis, MN. July 15-17.



Update On May/June Activities

June 24, 2015

Ingest

Locally-digitized Content

Staff working on HathiTrust processes ingested locally-digitized content from University of Illinois, Urbana-Champaign, University of Missouri, and University of Delaware. They also communicated with the Frick Collection, Universidad Complutense de Madrid, Princeton University, Northwestern University, Univerity of North Carolina, University of Florida, University of Alabama, Boston College, and University of Maryland.

Bibliographic Data Management

The California Digital Library (CDL) loaded 66,463 new, and 366,197 update records.

Projects

Copyright Review

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

	May		Overall	
	Public Domain Determinations	All Determinations	Public Domain Determinations	All Determinations
CRMS-US	966	1,636	171,997	324,735
CRMS-World	3,803	7,595	108,731	204,764
Total	4,769	9,231	280,728	529,499

US Federal Documents Registry

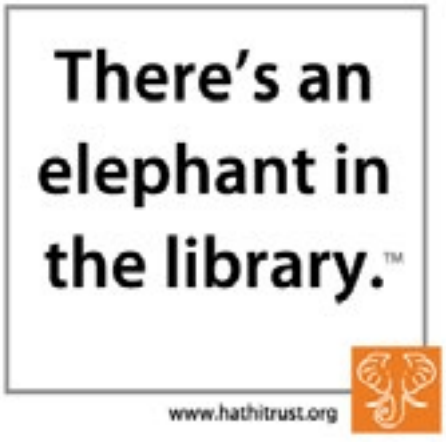
As of June 2, there are 631,197 US federal documents in HathiTrust.

Staff have been prepared the alpha release of the US Federal Government Documents Registry. Initial access to the data will be via a graphical interface, though there are plans to develop an API in the future. More than 15 million records have gone through the relationship detection process, yielding 3,661,389 clusters and 845,342 distinct items.

Work has also progressed on the manual review of records that aren't obviously duplicates, but aren't necessarily distinct. A rudimentary system is currently in place, allowing for some reviews to be made. Once enough decisions have been made, we'll use that information to refine the relationship detection process.

You can follow HathiTrust on Twitter or Facebook

Subscribe to email updates (via Google Groups)



Update On May/June Activities

June 24, 2015

HathiTrust Research Center

HTRC's Secure HathiTrust Analytic Commons (SHARC) can now be accessed by this URL: https://sharc.hathitrust.org.

Users should be aware that the SHARC team has set a monthly maintenance window for the 1st Tuesday of each month, starting in May. This gives the SHARC team a chance to apply patches, fixes and small changes, as well as perform clean-up, that requires the service to be down.

Eric Lease Morgan has put together the "HathiTrust Research Center Workset Browser," which he describes as a "fledgling tool for doing distant text mining against the corpora from the HathiTrust." The Research Center's monthly user group meeting on June 11 featured Eric discussing this proof-of-concept tool. See more on Eric's blog post http://blogs.nd.edu/emorgan/2015/05/htrc-workset-browser/

Eleanor Dickson joined the HTRC this month as Digital Humanities Specialist in the University of Illinois Library. Eleanor, a native of California, comes from Emory University where she was a Research Library Fellow in the Emory Center for Digital Scholarship and Emory's Manuscript, Archives, and Rare Book Library. She gained her master's degree in Information Studies from the University of Texas at Austin and a bachelor's degree in History and English from the University of California, Santa Barbara. Passionate about instruction, open data, and digital scholarship, Eleanor looks forward to exploring these areas and more in her new position.

Development Updates

Recent development updates and activities by HathiTrust institutions have included the following:

Full-text Search

- The development Solr environment was migrated to a new test server and the test scripts, debugging tools, and Solr configuration were tested and adjusted to the new environment. The new test server will allow more realistic performance testing of new Solr features and the re-testing of older features against the Solid State Drives.

Infrastructure

- Core services worked with the vendor to resolve hardware issues preventing system softare upgrade and successfully performed the upgrade for the cluster at the University of Michigan. The storage cluster at Indiana University-Purdue University Indianapolis will be upgraded by the end of June.

Summer Development Forecast

Put Solr plug-in to reduce memory use into production and complete the process of full-text reindexing.

Continue work on a test framework for relevance ranking, including interleaving of search results for the comparison of ranking algorithms.

Remove uses of and dependencies on CoSign for HathiTrust member authentication into HathiTrust (making authentication reliant only on Shibboleth).

Update On May/June Activities

June 24, 2015

PageTurner

- *Social Toolbar:* Added a "social toolbar" to PageTurner and Collection Builder listings for users to readily share links to HathiTrust content/collections. PageTurner provides metadata for Facebook and Twitter to enhance those links (e.g. https://twitter.com/EdenKristina/status/598832485666070528).
- *imgsrv:* Deployed enhanced logging of PDF downloads to generate more accurate reports for content owners. Full-book downloads are now being logged directly in Google Analytics; results will be reviewed in June to confirm accuracy.

Zephir - Outages

- Planned outage: Friday May 8th, 8pm through May 10th, 8pm. Zephir loading, updating, and exporting of records was paused during the outage window. FTPS was available for submissions, and any records submitted during that weekend were processed starting on Monday, May 11th.

Volumes Added

Updated ingest numbers can be found on our website here: http://www.hathitrust.org/statistics_deposited_volumes_monthly

Collection statistics are updated daily here: http://www.hathitrust.org/visualizations_deposited_volumes_current

Papers and Presentations

"Text mining with the HathiTrust Research Center: An introduction to working with digitized text corpora and metadata." Workshop; at the Annual Conference of the Humanities, Arts, Science, and Technology Alliance and Collaboratory (HASTAC), Michigan State University, East Lansing, MI. 30 May 2015.

B. Plale, represented HTRC at RDA Digital Humanities Workshop, Baltimore, MD. 28 May 2015. Two white papers were part of this presentation, "Secure Text Analysis at Scale over Sensitive Text: HTRC Data Capsules" and "Predicting Publication Date: A Text Analysis Exercise over 250,000 Volumes in the HTRC Secure Hathitrust Analytics Research Commons"

Angelina Zaytsev, "HathiTrust and a Mission for Accessibility." Journal of Electronic Publishing, 18.3 (Summer 2015)

Bhattacharyya, Sayan and J. Stephen Downie. "Approaching textuality with the metaphor of the digitized workset." Digital Humanities 2015 (DH 2015) Conference, Sydney, Australia. 29 June - 3 July 2015. (Accepted) Abstract Slides

Auvil, Loretta, Erez Lieberman Aiden, J. Stephen Downie, Benjamin Schmidt, Sayan Bhattacharyya and Peter Organisciak. "Exploration of Billions of Words of the HathiTrust Corpus with Bookworm: HathiTrust + Bookworm Project." Digital Humanities 2015 (DH 2015) Conference, Sydney, Australia. 29 June - 3 July 2015. Poster

Update On May/June Activities

June 24, 2015

Downie, J. Stephen. "HathiTrust: Large-Scale Data Repository in the Humanities." Invited keynote at 2015 International Workshop on Data Management, Beijing Institute of Technology Library, Beijing, China, 12 June 2015.

Downie, J. Stephen. "Metadata in the HathiTrust." Invited talk at 2015 International Workshop on Data Management, Beijing Institute of Technology Library, Beijing, China, 11 June 2015.

Organisciak, Peter, Loretta Auvil, Sayan Bhattacharyya and J. Stephen Downie. "The HTRC Extracted Features Dataset." Joint Conference of the Canadian Society of Digital Humanities/Société canadienne des humanités numériques (CSDH-SCHN) and the Association for Computers and the Humanities (ACH), 1-3 June, 2015, Ottawa.

Upcoming Presentations

Downie, J. Stephen. "The HathiTrust Research Center: Providing analytic access to the HathiTrust Digital Library's 4.7 billion pages." Invited keynote at ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL '15), Knoxville, TN, 24 June 2015.

Bhattacharyya, Sayan and Eleanor Dickson. "Introduction to the HathiTrust Research Center (HTRC): Teaching and research using the power of data and metadata in large text corpora." Workshop, Humanities Intensive Learning and Teaching (HILT) 2015, 28 July 2015.

Bhattacharyya, Sayan and Eleanor Dickson. "Workshop: Advanced Topics in Text Analysis with the HathiTrust Research Center (HTRC)." Humanities Intensive Learning and Teaching (HILT) 2015, Indiana University-Purdue University Indianapolis, IN, 29 July 2015.

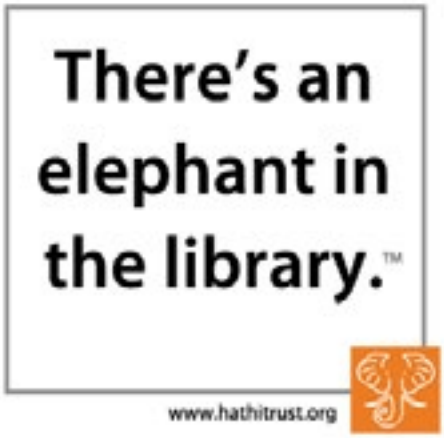
Bhattacharyya, Sayan. "HathiTrust Research Center: Textual Analytics on a Large Scale on the Corpus of the World's First Massive Digital Library." Linguistic Society of America (LSA)'s Biennial Linguistic Institute, The University of Chicago, IL, 13 July, 2015.

Repository Availability

Cumulative 12-month availability of repository access: 99.975% (+0.000%).

HathiTrust was briefly unavailable on Wednesday, May 6, from 11:28 - 11:34 ET after rebooting the wrong server due to an error in the inventory. The error in the inventory has been corrected.

A bug was introduced preventing the downloading of full book PDFs on Wednesday, May 27, from 09:36 - 14:33 ET. PDFs would appear to be built (in PageTurner), but the actual download would fail with a 500 exception. The bug was fixed.



Update On May/June Activities

June 24, 2015

User Support Issues	May	April
Content	**154**	**156**
Quality	139	140
Collections	15	16
Cataloging	**161**	**158**
Access and Use	**140**	**133**
Copyright	65	73
Permissions	12	14
Takedown	0	2
Print on Demand	0	0
Inter-library loan	0	4
Full-PDF or e-copy requests	21	29
Datasets	4	0
Data Availability and APIs	1	4
Reuse of content	1	4
Web applications	**29**	**47**
Functionality problems	10	15
Problems with login specifically	0	1
General questions about login	0	3
Partners setting up login	0	0
Usability issues	0	0
Feature requests	2	2
Partner Ingest	**13**	**15**
General	**97**	**115**
Partnership	6	17
Miscellaneous	91	98
Total	**594**	**607**

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

Most-accessed volumes

Crosses in the Wind, by Joseph James Shomon
War in New Guinea, Official War Photographs of the Battle for Australia.
The History of Royal College (formerly called the Colombo Academy), Written by Boys in the School, 1931.
Solid mensuration, by Willis F. Kern and James R. Bland.
Roster of the Confederate soldiers of Georgia, 1861-1865, v.2.
Quicksand, by Nella Larsen.
The Metco Metallizing Handbook, by H. S. Ingham & A. P. Shepard.
Roster of the Confederate soldiers of Georgia, 1861-1865, v.1.
Comparative Anatomy of the Vertebrates, by George C. Kent.