HathiTrust Digital Library

Update On March Activities

April 10, 2015

Top News

2015 HathiTrust Research Center UnCamp

The 3rd annual HTRC UnCamp was held at the University of Michigan on March 30 and 31. More than 130 registrants attended the event, in addition to HTRC staff from Indiana University and the University of Illinois. The UnCamp included keynote addresses by Michelle Alexopoulos (University of Toronto) and Erez Liberman Aiden (Baylor University) and numerous other presentations, posters, and demonstrations. The meeting agenda and participants are available at http://www.hathitrust.org/htrc_uncamp2015 and more information will be posted soon. The Digital Library Federation published a blog post aptly summarzing the day-and-a-half event. The HTRC would like to thank all attendees, participants, and speakers for a fabulous and successful UnCamp!

5 Million Open Volumes

HathiTrust achieved a major milestone in March, surpassing 5 million "open" volumes, including materials that are both in the public domain and open access. Read more in a blog post by Executive Director Mike Furlough.

Duke University Press Opens Access to Backfile Publications

Duke University Press opened access to more than 140 backfile publications in HathiTrust. Read the full announcement.

Revised Bylaws Posted

A revised version of the Bylaws of HathiTrust, as amended by the members in February 2015, is now available.

US Federal Government Documents Initiative Reports Now Available

In October 2014 the Government Documents Initiative Planning and Advisory Working Group submitted a status report on HathiTrust's Government Documents Initiative and recommendations for further action and investment in the program. The HathiTrust Program Steering Committee (PSC) endorsed the recommendations in principle and proposed to the Board of Governors that they be used as a guide for further action. At its February 2015 meeting, the Board of Governors endorsed the recommendations by the Advisory Working Group and PSC, noting the impact that this initiative would have for member libraries and the public at large, as well as the potential for this initiative to reinforce the actions planned in other initiatives of the membership. The Board directed Mike Furlough and HathiTrust staff to develop a preliminary implementation and staffing plan to be discussed at the May 2015 Board of Governors meeting in San Francisco. While that plan is being developed, we are publishing both the Advisory and Working Group report and Program Steering Committee report to inform the membership and wider commu-

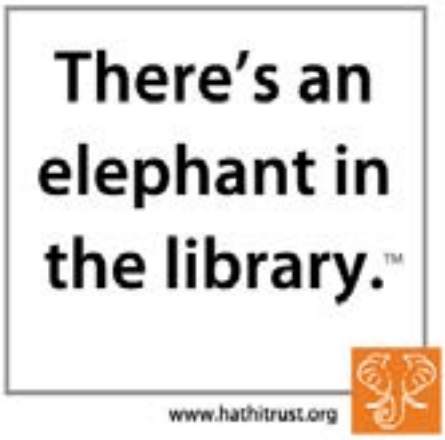
April Forecast

Begin production phase-in of high-performance storage system for full-text search

Release Solr plug-in to reduce memory use into production and begin 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.

Continue Testing mechanism for manual bibliographic record relationship detection.



Update On March Activities

April 10, 2015

nity of discussions and planning to date. Further details on the next steps for the Government Documents Initiative will be coming in the next few months.

User Support Working Group Nominations

The User Support Working Group is seeking nominations for up to 4 new members. We are seeking staff who have expertise in providing general user support and those who have expertise in cataloging in particular. Nominations are due May 15, 2015. To submit nominations and for further information about the working group, please visit http://tinyurl.com/m9qlyyg.

Ingest

Google-digitized Content

HathiTrust began ingest of materials, comprised mainly of US federal government documents, from the University of Iowa.

Locally-digitized Content

HathiTrust corresponded with Northwestern University, Princeton University, and Boston College about submission of new content, and ingested additional content from Emory University and the University of Illinois.

Bibliographic Data Management

The California Digital Library loaded 85,457 new or updated records into Zephir.

Projects

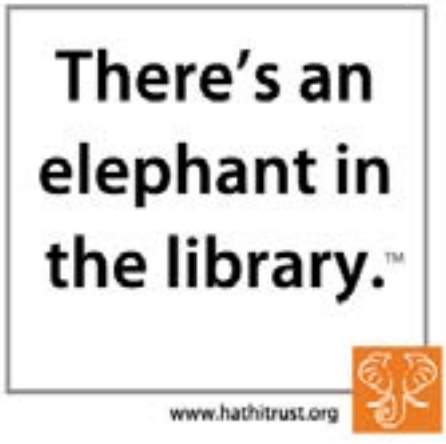
Copyright Review

A summary of the determinations from HathiTrust copyright review activities in March 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.

	March		Overall	
	Public Domain	All Determinations	Public Domain	All Determinations
CRMS-US	865	1,352	170,239	321,945
CRMS-World	4,006	6,942	100,740	189,575
Total	4,871	8,294	270,979	511,520

You can follow HathiTrust on Twitter or Facebook

Subscribe to email updates (via Google Groups)



HathiTrust Digital Library

Update On March Activities

April 10, 2015

Government Documents Registry

Project staff made progress on an initial mechanism for conducting manual review of bibliographic records to determine whether and how the works described by the records are related (e.g. duplicates). The mechanism will be tested throughout April and May, and would be used especially in cases where the determination made during the automated relationship detection was of lower confidence.

HathiTrust Research Center Updates

The HTRC renamed its services environment to SHARC (Secure HathiTrust Analytic Research Commons) and released SHARC v3.1 on March 26, 2015. The release was mainly aimed at bug fixes, but has a small set of new features. Changes to the main production service include:

- Workset Listing by My Worksets and All Worksets.
- Algorithm results have sortable Algorithm column and default order is by Time.
- Email for user account registration now comes from sharc@indiana.edu
- Data Capsule results are only available after a human review
- API for Public Worksets, used by the beta version Bookworm (in the Sandbox)

Changes to the Sandbox service include:

- Initial alpha version of Bookworm
- Updates to the alpha version Feature Extraction

We welcome community scholars to try the 3.1 release and provide feedback. Please send comments to htrc-tech-help-l@list.indiana.edu. Notes on the previous 3.0 release are available on the HTRC wiki.

HTRC staff created a LibGuide for the HT+Bookworm tool prototype.

For those interested in a more advanced deep dive on topic exploration within the HTRC Data Capsule, we will be offering a tutorial at JCDL 2015 on June 21 in Knoxville, TN. For more information, see http://www.jcdl2015.org/tutorials-workshops (Registration is now open!).

Development Updates

Development updates and activities by HathiTrust institutions included the following:

Full-text Search

- Released code to take advantage of item-level date information for serials. The new date information will be available in HathiTrust search interfaces

HathiTrust on the Road

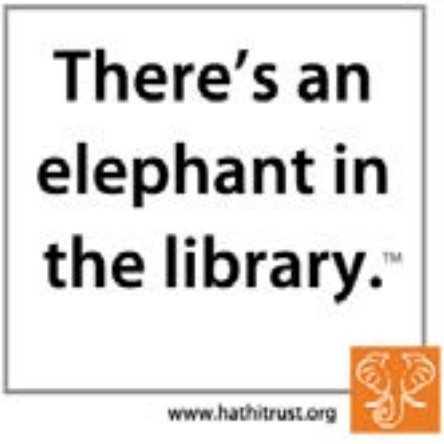
Staff from HathiTrust will be attending the following meetings and conferences in the next few weeks. Please be in touch if you would like to meet with us during our travels:

DPLAFest, Indianapolis, IN, April 17-18 2015:

- Mike Furlough, Executive Director
- Jeremy York, Assistant Director
- Angelina Zaytsev, Project Librarian
- Melissa Levine, Director, CRMS Project, Copyright Officer, Univ of Michigan

IMLS Focus: the National Digital Platform, Washington, DC, April 28, 2015: Mike Furlough

Association of Research Libraries Member Meeting, Berkeley, CA, April 29-30, 2015: Mike Furlough



Update On March Activities

April 10, 2015

when the repository is re-indexed in April.

- Tested a Solr plug-in to reduce memory use in Solr 4. Testing found an order of magnitude reduction in memory use in the in-memory version of the "index to the index" for full-text search. The plug-in will be put into production during the next re-indexing job as well. This will allow an increase in the number of shards serving the index from 12 to 18 without additional hardware, an increase in the amount of memory available for OS I/O caching on the search servers, and facilitate future testing of relevance ranking.
- Began work on a test framework for relevance ranking.
- Created a prototype interface that displays search results from two different ranking algorithms side-by-side.
- Fixed a bug in indexing code that introduced errors in the item information for some serials. The affected volumes were re-indexed to restore the correct information.

Handle Service

- Transferred the Handle (persistent identifier) service to new servers.

PageTurner

- Added language to identify works that are in the public domain but for which access is limited due to privacy concerns.
- Deployed improvements to accessibility features with particular attention to support for new content types.
- Updated the PDF generation process to use coordinate OCR information where it is available, allowing highlighting of search results in downloaded PDFs.

Recording Sources of Digital Objects

- Implemented a new strategy for recording administrative information about digital objects, such as who digitized and deposited the content.
- Implemented a single scheme for identifying institutions across repository systems in conjunction with this change.
- Staff at the University of Michigan and the California Digital Library coordinated to harmonize the administrative information between the repository and bibliographic data management systems.

Zephir

- Released an API to facilitate use of bibliographic records managed in Zephir in workflows for content ingest.
- Discussed strategies for improving tracking and reporting about records and digital items submitted for deposit in HathiTrust, and possible modi-

Papers and Presentations

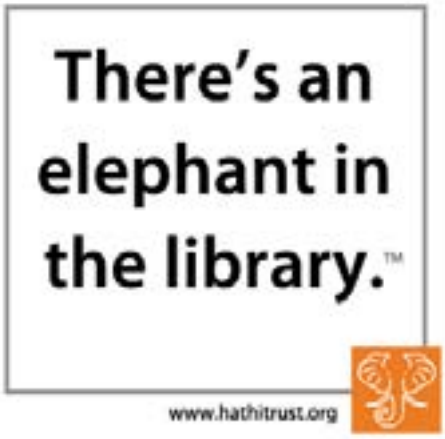
Beth Plale, "HathiTrust: Large-Scale Repository in the Humanities Unlocking the Secrets of 4.6 Billion Pages", Cyberinfrastructure Day, University of Missouri, March 3, 2015.

Mike Furlough, "HathiTrust and Collective Stewardship at Scale," Washington Research Libraries Consortium 2015 Annual Meeting, George Washington University, Washington, DC, March 10, 2015.

Robert McDonald, "Elephant in the Room: Scaling Storage for the HathiTrust Research Center," PASIG 2015, San Diego Super Computer Center, March 13, 2015.

Jeremy York, "Digital Humanities at Any Scale", Digital Humanities and the futures of Japanese Studies, University of Michigan Library, March 13, 2015.

Sayan Bhattacharyya, seminar presentation on "Rethinking Text as Process in the Humanities, Digital and non-Digital," ACLA Annual Meeting, Seattle, WA, March 26-29, 2015.



fications to the workflow for processing record corrections.
- Discussed a preliminary draft of policies for handling requests for reporting from Zephir and Zephir system enhancements.

Availability

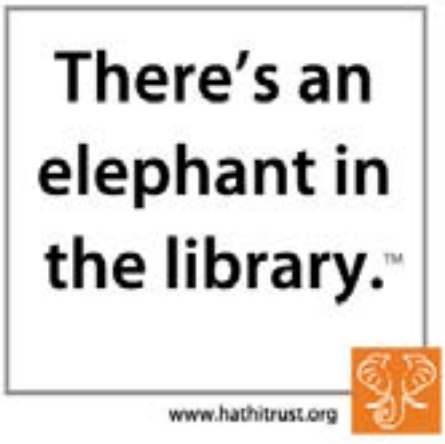
Repository

Cumulative 12-month availability of repository access*: 99.971% (-0.001%)

HathiTrust objects may have briefly been unavailable on Tuesday, March 24 from 15:45-15:50 ET through their handle.net persistent URLs due to a server misconfiguration. Staff immediately noticed and corrected the problem.

From February 24, 2015 through March 24, 2015, some users were denied access to volumes with a rights determination of "public domain only within the United States" due a bug in coding changes designed to test for accesses to HathiTrust from proxy servers.

* 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.



HathiTrust Digital Library

Update On March Activities

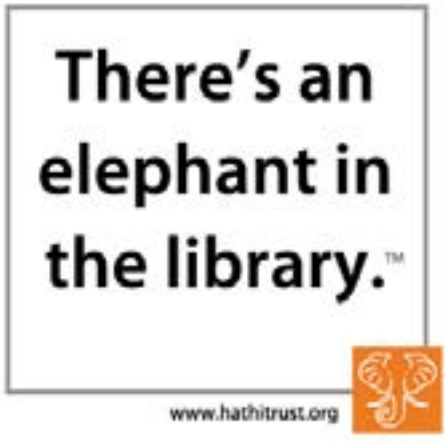
April 10, 2015

User Support Issues	March	February
Content	**148**	**227**
Quality	144	211
Collections	13	12
Cataloging	**154**	**164**
Access and Use	**130**	**157**
Copyright	65	105
Permissions	9	18
Takedown	0	0
Print on Demand	0	0
Inter-library loan	2	2
Full-PDF or e-copy requests	29	12
Datasets	2	5
Data Availability and APIs	0	3
Reuse of content	3	6
Web applications	**47**	**41**
Functionality problems	27	25
Problems with login specifically	1	1
General questions about login	1	3
Partners setting up login	1	0
Usability issues	0	0
Feature requests	0	1
Partner Ingest	**10**	**7**
General	**138**	**134**
Partnership	6	8
Miscellaneous	132	126
Total	**637**	**730**

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

Most-accessed volumes

Quicksand, by Nella Larsen.
Solid mensuration, by Willis F. Kern and James R. Bland
The Human Figure, by John H. Vander-poel
Roster of the Confederate soldiers of Georgia, 1861-1865, v. 1
History of wages in the United States from Colonial times to 1928, United States Department of Labor.
Roster of the Confederate soldiers of georgia, 1861-1865, v.2.
Abstracts of old Ninety-six and Abbeville District wills and bonds, as on file in the Abbeville, South Carolina Courthouse
The Five Laws of Library Science, by S. R. Ranganathan.
Godey's Magazine, v. 40-41, 1850
Roster of the Confederate soldiers of Georgia, 1861-1865, v.3.



HathiTrust Digital Library

Update On March Activities

April 10, 2015

Total Volumes Added	March	Overall
Boston College	294	3,557
Columbia University	0	73,396
Cornell University	9	515,753
Duke University	409	8,615
Emory University	128	180
Getty Research Institute	458	20,588
Harvard University	0	838,122
Indiana University	32	529,798
Keio University	2	90,122
Knowledge Unlatched	0	28
Library of Congress	0	108,892
McGill University	0	893
New York Public Library	6	304,610
North Carolina State University	0	3,196
Northwestern University	8	57,000
Ohio State University	5,431	74,525
Penn State University	27	389,247
Princeton University	0	252,841
Purdue University	0	47,488
Sterling & Francine Clark Art Institute	0	358

Total Volumes Added	March	Overall
Texas A&M	0	2,446
Universidad Complutense	12	117,322
University of Alberta	0	76,106
University of California	14,888	3,639,937
University of Chicago	579	56,981
University of Connecticut	0	4,637
University of Delaware	0	48
University of Florida	0	9,866
University of Illinois	9,818	348,946
University of Iowa	7,551	7,551
University of Massachusetts, Amherst	0	12,007
University of Michigan	757	4,722,050
University of Minnesota	586	334,249
University of Missouri	0	1
UNC - Chapel Hill	0	17,025
University of Virginia	408	51,207
University of Wisconsin	0	561,534
Utah State University	0	117
Yale University	0	23,832

Total Volumes Added	41,403	13,305,071
Total Public Domain (~37% of total)*	37,361	5,004,951

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