HathiTrust Digital Library

Update On January Activities

Feburary 13, 2015

Top News

HTRC Operations Manager, 3.0 Beta Release, UnCamp

The HathiTrust Research Center welcomed Dirk Herr-Hoyman as the new HTRC Operations Manager, based at Indiana University. Dirk has many years of experience with large-scale web applications and software development in both the public and private sector. He joins the HTRC from the University of Wisconsin-Madison, where he was involved in research and instructional computing initiatives. This is Dirk's second time on the Indiana University Bloomington campus. His first was as a computer science major from '74-'78.

The HTRC also announces the beta release of HTRC Services v3.0. The 3.0 release features the integration of the HTRC Data Capsule, plus a more welcoming portal, enhanced workset builder functionality and improved security features. The HTRC Data Capsule provides a secure computation and data environment for non-consumptive research. It permits analytical investigation of a corpus, e.g. copyrighted volumes, but prohibits data from leaving the capsule. Try it out at the portal and see the documentation for introduction, user guide, and tutorial.

Other notable enhancements for the 3.0 release include:

- Automatically saving jobs upon completion
- Corrected use of faceted search
- Single sign-on (except for Data Capsule)

Please remember to save the date for the 2015 UnCamp! Registration is now open and information can be found on the event page.

Ingest

Locally-digitized Content

HathiTrust communicated with several institutions about new ingest of locally digitized materials, and ingested a new batch of content from the University of Illinois.

Internet Archive-digitized Content

HathiTrust began ingesting dissertations from the University of Massachusetts, Amherst.

Bibliographic Data Management

The California Digital Library (CDL) loaded 23,635 new and 63,135 updated bibliographic records into Zephir.

February Forecast

Update full-text search services to index and use both bibliographic and item-level date information.

Reassess accessibility features of PageTurner with particular attention to supporting new content types.

Incorporate coordinate OCR into PDFs generated and delivered from HathiTrust.

Continue working on migration to Solr 4.

HathiTrust on the Road

HathiTrust administrative staff will be attending the following upcoming meetings. Please get in touch if you would like to meet with us there:

Jeremy York, RDA and PASIG, San Diego, March 8-13, 2015.

Mike Furlough, Washington Research Library Consortium Annual Meeting, Washington, DC., March 10, 2015.

HathiTrust Digital Library

Update On January Activities

Projects

Copyright Review

A summary of the determinations from HathiTrust copyright review activities in December is given below. See CRMS-US and CRMS-World for further information.

	December		Overall	
	Public Domain	All Determinations	Public Domain	All Determinations
CRMS-US	489	840	168,248	318,887
CRMS-World	3,498	6,141	92,919	175,681
Total	3,987	6,981	261,167	494,568

Government Documents Registry

Project staff continued to develop and refine a process for identifying relationships between US federal government documents based on bibliographic information. Staff ran current relationship detection algorithms on a large set of 2,163,339 government documents records from HathiTrust member institutions (the records describe both volumes that are in HathiTrust and volumes not in HathiTrust but held physically by institutions). The records represent 4,500,379 total, and 2,753,817 distinct, items. As next steps, staff will be reviewing results of the initial pass and making further refinements to the algorithms, before incorporating the records of more than 40 institutions received as part of HathiTrust's call for government documents records in 2013 into the analysis.

Project staff also began conducting an analysis of the contents of bibliographic record MARC 110 fields, and comparison of these values with authority records in VIAF. Preliminary results indicate that 95% of 1,519,368 110 field entries map to a corporate name authority in VIAF. Additionally, staff identified 33,660 VIAF authorities for likely US federal government documents that were not represented in the record set. Work is ongoing, but it is possible that work with VIAF will aid in the detection of gaps in the Registry or identification of government publications in the HathiTrust corpus that are not properly cataloged as such.

Additionally, an FAQ for the Government Documents Initiative was created and is available at http://www.hathitrust.org/help_usgovdocs.

Development Updates

Development updates and activities by HathiTrust institutions included the following:

Access, Authorization, and Authentication:

Papers & Presentations

Sayan Battacharyya, Jeremy York, "Humanistic Inquiry with Large Corpora of Digitized Text and Metadata: Toward New Epistemologies?" Modern Language Association Annual Meeting, Vancouver, British Columbia, January 9, 2015.

Furlough, Farb, Teper, Sandler, Sandore, "HathiTrust Update", ALA Midwinter 2015, Chicago, January 29, 2015.

J. Stephen Downie, "Unlocking the Secrets of 4.5 Billion Pages", University of Victoria, University of Waikato and the National Library in New Zealand.

You can follow HathiTrust on Twitter or Facebook

Subscribe to email updates (via Google Groups)

HathiTrust Digital Library

Update On January Activities

- Improved notification system for unsuccessful attempts of staff to register for special access to in-copyright works.
- Automated warnings of Data API access key expiration for clients that have been granted higher levels of authorization.
- Improved Data API client code examples based on feedback from developers at the HathiTrust Research Center.

Full-text Search

- Tested memory needs for Solr 4. Testing revealed that Solr 4 is significantly more efficient than Solr 3. However, staff will need to create a plugin for Solr to take full advantage of Solr 4's memory efficiency improvements.
- Began a process to migrate the index from Solr 3 to Solr 4. Efforts to migrate revealed a bug in the Solr 4.x (Lucene 4.x) indexing code that, in the presence of very frequent words in very large indexes, produces a corrupted index. Michigan staff worked with Lucene committers to determine the problem and create and apply a patch (see https://issues.apache.org/jira/browse/LUCENE-6192). Re-indexing with the patch was completed in January and the new index will go into production in early February.
- Changed a MySQL table involved in page-level indexing from MyISAM to InnoDB to improve indexing throughput.
- Implemented processes to automatically synchronize full-text indexing with HathiTrust Print Holdings database updates and HathiTrust catalog indexing, in order to ensure the correct representation of holdings for items in the full-text index.
- Improved the efficiency of incorporating updates to print holdings information from members in full-text indexing.
- Staff are due to receive, in early February, the long-awaited production-quality soft-

User Support Issues	January	December
Content	**158**	**121**
Quality	143	109
Collections	15	11
Cataloging	**142**	**115**
Access and Use	**121**	**109**
Copyright	76	43
Permissions	8	16
Takedown	0	1
Print on Demand	0	0
Inter-library loan	0	0
Full-PDF or e-copy requests	11	14
Datasets	2	2
Data Availability and APIs	1	0
Reuse of content	1	0
Web applications	**28**	**20**
Functionality problems	12	6
Problems with login specifically	0	1
General questions about login	0	2
Partners setting up login	1	0
Usability issues	0	0
Feature requests	3	1
Partner Ingest	**6**	**13**
General	**103**	**109**
Partnership	9	8
Miscellaneous	94	101
Total	**558**	**487**

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

ware fix for the high-performance storage to address performance and stability problems. The upgrade will be installed and tested promptly, and when confirmed to be stable, the storage will be phased into production.

Storage Replacement Cycle

- Completed installation of new storage equipment at both sites (Michigan and Indiana). The removal of equipment due for retirement is scheduled to begin in mid-February.

Availability

Repository

Cumulative 12-month availability of repository access (user-facing applications): 99.964% (+0.000%). No outages were reported in January.

Zephir

There was a planned outage of the Zephir FTPS server on Wednesday, January 14 from 10-11 AM PST. Members were not able to drop off bibliographic records to Zephir's FTPS server during the outage.

HathiTrust Digital Library

Update On January Activities

Total Volumes Added	January	Overall
Boston College	0	3,263
Columbia University	1	73,396
Cornell University	221	510,286
Duke University	0	8,206
Emory University	0	52
Getty Research Institute	583	19,562
Harvard University	5	838,115
Indiana University	790	529,601
Keio University	18	90,112
Knowledge Unlatched	0	28
Library of Congress	0	108,892
McGill University	0	893
New York Public Library	48	294,883
North Carolina State University	0	3,196
Northwestern University	278	56,955
Ohio State University	7,288	68,417
Penn State University	996	388,713
Princeton University	29	252,837
Purdue University	0	47,488
Sterling & Francine Clark Art Institute	0	358
Texas A&M	0	2,446
Universidad Complutense	56	117,291
University of Alberta	0	76,106
University of California	2,310	3,614,906
University of Chicago	162	52,138
University of Connecticut	0	4,637
University of Delaware	0	48
University of Florida	0	9,866
University of Illinois	11,005	329,136
University of Massachusetts, Amherst	390	12,004
University of Michigan	3,607	4,716,359
University of Minnesota	48,407	193,124
UNC - Chapel Hill	0	17,025
University of Virginia	0	51,207
University of Wisconsin	319	561,094
Utah State University	0	117
Yale University	0	23,832
Total	76,513	13,076,589

Public Domain (~37% of total)

Total*	29,001	4,898,282

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

Most-accessed volumes

- The Human Figure, by John H. Vanderpoel.
- Quicksand, by Nella Larsen.
- Godey's Magazine, v.40-41, 1850.
- Pennsylvania German pioneers; a publication of ... v.42.
- The Book of a Hundred Hands, by George Brant Bridgman.
- Indian boyhood, by Charles A. Eastman.
- Descendants of Governor William Bradford (through the first seven generations), compiled by Ruth Gardiner Hall.
- Roster of the Confederate soldiers of Georgia, 1861-1865, v.2.
- Solid mensuration, by Willis F. Kern and James R. Bland.
- The Five Laws of Library Science, by S. R. Ranganathan.
- Roster of the Confederate soldiers of Georgia, 1861-1865, v.1.