HathiTrust Digital Library

Update On May Activities

June 13, 2014

Late Breaking News

Ruling in Authors Guild Lawsuit Appeal

HathiTrust released a statement on the decision of the U.S. Second Circuit Court in the lawsuit brought by the Authors Guild et al. against HathiTrust. The decision is a strong affirmation of the work HathiTrust has undertaken to enhance access to and preserve the collections of its member libraries.

Top News

HTRC Page Features Dataset

The HathiTrust Research Center released a new dataset, consisting of page-level features extracted from a quarter of a million books. The dataset is an alpha release, demonstrating the features the HTRC intends to make available across all public domain volumes in HathiTrust and eventually the entire HathiTrust corpus.

Government Documents Registry Applications Developer

HathiTrust is seeking an applications developer to design, implement, and populate a Registry of metadata describing and identifying the comprehensives corpus of U.S. federal government documents. See the full description and apply on the University of Michigan Jobs site.

HathiTrust Board Update

The HathiTrust Board of Governors met on May 9, 2014 in Columbus, OH for one of two in-person meetings held each year (two additional meetings are held by phone each year). The Board heard updates on activities from incoming Executive Director Mike Furlough; a report on the work of the Program Steering Committee by Bob Wolven; and a budget report from Treasurer and chair-elect Rick Clement. Assistant Director Jeremy York gave a presentation entitled “HathiTrust, Copyright Policies and Issues”, covering topics such as access to public domain works in the US and outside the US; lawful uses of in-copyrights works; the Copyright Review Management System (CRMS); and user inquiries.

The Board took the following actions:

- Approved 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.
- Approved allocation of an additional $115,000 to extend staffing in support of development of the Government Documents Registry.
- Approved the process to appoint a replacement to the Program Steering Committee to replace Mike Furlough.
- Approved the first annual HathiTrust Membership Meeting to be held in Washington, DC on October 10. Details about this meeting will be forthcom-

June Forecast

Final testing and production deployment of the automated user access renewal and deletion application.

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

Correct a bug in the large scale search results navigation that makes it difficult to return to the first page of results if the user advances far enough into later result pages.

You can follow HathiTrust on Twitter or Facebook

Subscribe to email updates (via Google Groups) ing in the next several weeks.

Brad Wheeler announced that Brenda Johnson, Ruth Lilly Dean of University Libraries at Indiana University, would be taking his position for Indiana University on the Board of Governors. The Board will next meet by conference call on July 29. The next in-person meeting is scheduled for October 9.

Ingest

General

HathiTrust staff corresponded with staff from the University of Washington, the University of Illinois at Urbana-Champaign, the Getty Research Institute, Boston College, Emory University, the University of California, and the University of Michigan regarding ingest of locally digitized content. HathiTrust ingested additional locally-digitized content from the University of Illinois. HathiTrust also ingested 19 volumes from Knowledge Unlatched.

HathiTrust staff answered questions from the Getty Research Institute, Emory University, and McGill University about ingest of content digitized by the Internet Archive. HathiTrust ingested more than 300 volumes (262 titles) from the Sterling and Francine Clark Art Institute Library.

Working Groups and Committees

Program Steering Committee

Two new members were appointed to the Program Steering Committee to serve 2-year terms, beginning in June. The new members are Robert McDonald, Associate Dean, Library Technologies, Indiana University, and Chris Freeland, Associate University Librarian, Washington University in St. Louis.

Under the aegis of the Program Steering Committee, the Print Monographs Archive Task Force has begun work, with Tom Teper (University of Illinois) serving as chair. The other members are Clem Guthro (Colby), Robert Kieft (Occidental), Erik Mitchell (Berkeley), Jake Nadal (ReCAP), Matthew Sheehey (Harvard) Emily Stambaugh (University of California), and Karla Strieb (Ohio State). The PSC has also been reviewing HathiTrust's use of automated quality metrics provided by Google to reduce the number of poorer quality volumes that are ingested, and will shortly appoint a task force to assess this issue and make recommendations.

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.

Papers & Presentations

Valerie Glenn, "Defining and Identifying the GovDocs Corpus: the HathiTrust Registry", 2014 Depository Library Council Meeting and Federal Depository Library Conference, May 1, 2014.

J. Stephen Downie, "HathiTrust Research Center: The Workset Creation for Scholarly Analysis (WCSA) Prototyping Project", Kungliga Tekniska högskolan (Royal Institute of Technology), Stockholm, Sweden, May 6, 2014.

Jeremy York, Brian E.C. Schottlaender, "The Universal Library is Us: Library Work at Scale in HathiTrust", Educause Review, May 19, 2014.

Tom Burton-West, "Practical Relevance Ranking for 11 Million Books, Part 1". HathiTrust Large-scale Search Blog.

Beth Plale, Keynote: "Bridging Digital Humanities Research and Large Repositories of Digital Text", 2nd Encuentro de Humanistas Digitales, Biblioteca Vasconceles, Mexico City 21 May 2014.

Beth Plale, "HathiTrust and HTRC: the Changing Digital Library", El Colegio de Mexico, Mexico City, 20 May 2014.

	May		Overall	
	Public Domain	All Determinations	Public Domain	All Determinations
CRMS-US	215	315	165,340	314,270
CRMS-World	3,996	7,268	59,652	117,369
Total	4,211	7,583	224,992	431,639

Government Documents Registry

Project staff have been developing normalization rules, focusing on normalization of enumeration and chronology, in order to aid duplicate detection efforts. Staff continue to refine methods for potential identification of related metadata records, as well as methods for the identification of gaps in metadata.

Development Updates

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

Authentication and Authorization

Staff completed development of an application begun in April to simplify and enhance the administration of users requiring access to restricted items. Final testing and production deployment will occur in June.

Full-text Search

Staff conducted thorough testing to identify the probable source of performance and stability problems encountered with new high-performance storage purchased for full-text search. Staff are in close communication with the supplier and an update to address the problems is expected to be available in June for testing.

Staff conducted performance tests on the page-level and 3,000-word chunk indexes described in the Update on April 2014 Activities. Tests using page-level indexing indicated that query performance, faceting performance and grouping performance would be unacceptably slow given current hardware. Preliminary results using 3,000-word chunks showed that memory for faceting search results would need to be between 1.2 and 1.5 times greater than the memory currently in use in order for faceting to be functional. Even at that level, query response times were slower than desired. Tests using both indexes will be repeated when new high-performance storage is in place.

Staff obtained the INEX Book Track 2007-2010 test collections, which include MARC metadata and the full text of between 40,000 and 50,000 books, and are investigating the use of the collections to help inform choices about relevance ranking of full-text search results. Staff conducted tests using the default Solr/Lucene

Papers & Presentations

Miao Chen, "HathiTrust Research Center: Technical Challenges", Data to Insight Center Office of Sponsored Programs Workshop, June 4, 2014.

Partner Presentations

Kirsten Clark (University of Minnesota), Amy Springer (University of Minnesota), Catherine Morse (University of Michigan), "HathiTrust 101", 2014 Depository Library Council Meeting and Federal Depository Library Conference, May 1, 2014.

ranking algorithm as well as three new algorithms available in Solr/Lucene 4.0 (BM25, Language Model with Dirichlet smoothing, and DFR). Testing will continue in June.

Tom Burton-West authored the first of a series of blog posts about "Practical Relevance Ranking for 11 Million Books."

Google Analytics

Staff updated Google Analytics to be able to track the usage of HathiTust Collections in addition to individual items.

mPach

Michigan developers began a full review of the accessibility of the PageTurner application. This work is expected to have implications for the display of XML content, including mPach JATS articles. Staff continued to work on an XSLT implementation of Norm for DocX to JATS conversion. More information about the mPach project can be found at http://www.lib.umich.edu/mpach and http://www.hathitrust.org/mpach.

PageTurner

Staff fixed bugs and made improvements to the "search in this text" widget for navigating from one page of results to another. The modifications will be released into production in June.

Server replacement cycle

Staff ordered new full-text search servers to replace servers scheduled for retirement, but delivery was delayed by the supplier. Installation is still expected to begin in June, but will likely not be complete until July.

Availability

Repository

Cumulative 12-month availability of repository access: 99.867%

User Support Issues	May	April
Content	**131**	**154**
Quality	124	143
Collections	7	11
Cataloging	**285**	**187**
Access and Use	**142**	**142**
Copyright	88	101
Permissions	6	10
Takedown	0	1
Print on Demand	0	1
Inter-library loan	0	0
Full-PDF or e-copy requests	17	14
Datasets	3	3
Data Availability and APIs	3	2
Reuse of content	4	4
Web applications	**18**	**20**
Functionality problems	8	10
Problems with login specifically	1	3
General questions about login	0	1
Partners setting up login	0	0
Usability issues	0	0
Feature requests	1	1
Partner Ingest	**7**	**11**
General	**93**	**110**
Partnership	7	18
Infrastructure	0	0
Miscellaneous	86	92
Total	**676**	**624**

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

HathiTrust Digital Library

Update On May Activities

Total Volumes Added	May	Overall
Boston College	86	3,197
Columbia University	0	65,037
Cornell University	4,787	453,905
Duke University	516	7,774
Harvard University	0	237,435
Indiana University	15	195,666
Keio University	0	88,956
Knowledge Unlatched	19	19
Library of Congress	953	108,882
New York Public Library	129	291,790
North Carolina State University	0	3,196
Northwestern University	1	37,644
Ohio State University	2,744	23,852
Penn State	2,008	81,207
Princeton University	3	251,713
Purdue University	0	44,698
Sterling & Francine Clark Art Institute	326	326
Texas A&M	0	1,201
Universidad Complutense	3	112,151
University of California	5,876	3,500,120
University of Chicago	2	39,171
University of Florida	101	9,866
University of Illinois	608	136,299
University of Massachusetts, Amherst	1,704	11,115
University of Michigan	1,569	4,672,249
University of Minnesota	15	119,877
UNC - Chapel Hill	0	17,025
University of Virginia	4	50,825
University of Wisconsin	128	556,101
Utah State University	0	117
Yale University	0	23,678
Total	21,597	11,145,111

Public Domain (~33% of total)

Total*	17,130	3,756,406

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

Most-accessed volumes

Consumption of the Lungs and Kindred Diseases, Treated and Cured by Kerosene, by Charles Oscar Frye.
Journal of the...annual convention of the Episcopal Church in the Diocese of Connecticut, 1867-71.
The Human Figure, by John H. Vanderpoel.
Quintus Curtius [History of Alexander], Vol. 2, with an English translation by John C. Rolfe.
Memoir of Colonel Benjamin Tallmadge.
West Side story, a novelization, by Irving Shulman.
History of wages in the United States from Colonial times to 1928, United States Department of Labor.
Coffee processing technology, v. 1, by Michael Sivetz and H. Elliott Foote
The fool; his social and literary history, by Enid Welsford
Roster of the Confederate soldiers of Georgia, 1861-1865, v.1.

HathiTrust's mobile site was incorrectly directing users to the full site from Tuesday, May 6 at 1:30pm to Thursday, May 8 at 5:45pm.

HathiTrust was unavailable on Thursday, May 8 for 6 brief periods between 1:44pm and 1:55pm approximately 40 seconds in length due to a software stability problem that occurred at one instance while the other site was down for routine maintenance.

There's an elephant in the library.™

www.hathitrust.org