HathiTrust Digital Library

Update On November Activities

December 13, 2013

Top News

New Partners

HathiTrust was very pleased to welcome the University of Alabama as a new member in November, and, in late-breaking news, the University of Massachusetts in early December. Please see the press releases for Alabama and UMass Amherst for more information.

Executive Director Search

The Search Committee for the HathiTrust Executive Director position completed its preliminary review of applications and conducted telephone interviews with nearly a dozen top candidates. The Search Committee expects to conduct in-person interviews in January.

HTRC Call for Proposals

The HathiTrust Research Center is seeking proposals for prototyping projects to define and implement a tool or service that will help scholars better identify and select relevant resources at scale from the HathiTrust corpus and/or facilitate the construction of large-scale worksets useful for scholarly analyses. Grants of $40,000 will be offered to each of four successful respondents to be conducted over a nine-month period beginning April 2014. Workset Creation for Scholarly Analysis: Prototyping Project (WCSA) is generously funded by the Andrew W. Mellon Foundation.

A complete copy of the RFP is available online at http://worksets.htrc.illinois.edu/worksets/?page_id=20. Letters of intent are due December 16, 2013 and full proposals are due January 13, 2014.

Government Documents Initiative

HathiTrust has prepared a FAQ to accompany the recent call for US federal government documents records. The records will be used for analysis to gain a greater understanding of the total corpus of US federal government documents in support of HathiTrust's gov docs initiative. More information is available at http://www.hathitrust.org/usgovdocs.

Nominations for User Support Working Group

The User Support Working Group is seeking nominations for 2-4 new members. This is a great opportunity to get involved in the day-to-day work of the partnership. We are seeking both staff who have expertise in providing user support, and those who have expertise in cataloging to do investigation into reports of cataloging errors we receive from users. Nominations should be submitted by January 17, 2014 via the nomination form at http://bit.ly/1f9LrAa. The form includes information about expected workload and learning opportunities.

December Forecast

Continue adding support for full-text indexing of JATS articles, and indexing of volumes in "chunks".

Continue development of ePub and PDF generation from JATS.

Continue to explore relevance ranking solutions.

Papers & Presentations

Sigrid Cordell, Jeremy York, "HathiTrust: The Collection and Its Uses", NEFLIN Webinar, November 7, 2013.

Beth Plale, "Opportunities and Challenges of Text Mining HathiTrust Digital Library", The Hague, November 15, 2013.

Partner-specific:

Craig Willis and Miles Efron, "Finding information in books: characteristics of full-text searches in a collection of 10 million books", ASIS&T 2013.

Kat Hagedorn, Meghan Musolff, Angelina Zaytsev, "Your Road Map to Digital Collections at the Library", UM Digital Collections/HathiTrust Workshop, University of Michigan Library, November 21, 2013.

There's an elephant in the library.™

www.hathitrust.org

Update On November Activities

Ingest

Validation service for locally-digitized materials

HathiTrust created a beta version of a new web-based service to validate single image files, which is available at http://bit.ly/1gvD9q7. Please feel free to try the validator and provide feedback using the form at http://bit.ly/1byxI1t. Planning and development continued on a cloud-storage-based service to validate entire volumes, scheduled for release in January.

General

HathiTrust began ingest of a set of locally-digitized AgriLife Research Bulletins (formerly the Bulletin of the Texas Agricultural Experiment Station) from Texas A&M University, and completed ingest of a set of University Press of Florida backfile publications, in time to mark University Press Week. The UPF collection of materials can be viewed at http://bit.ly/1guQG1h.

HathiTrust corresponded with staff from the University of Chicago and the Universidad Complutense de Madrid about future ingest of locally-digitized volumes, and consulted with staff from the University of Massachusetts Amherst and Boston College about ingest of Internet Archive-digitized volumes.

Working Groups and Committees

Program Steering Committee

The Program Steering Committee is in the process of appointing a Government Documents Initiative Planning and Advisory Group and will be appointing members in the next few weeks. This is in response to one of the ballot initiatives approved at the Constitutional Convention. The PSC is also recommending allocation of funds to complete a framework for certifying the quality of individual volumes in HathiTrust, and, if approved, will appoint an advisory group to oversee this effort. The Committee continues to work on charges for a Shared Print Archive initiative, a reconstituted Collections Committee, and new initiatives regarding rights and access and metadata policy.

Projects

Copyright Review

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

You can follow HathiTrust on Twitter or Facebook

Subscribe to email updates (via Google Groups)

There's an elephant in the library.™

www.hathitrust.org

Update On November Activities

	November		Overall	
	Public Domain	All Determinations	Public Domain	All Determinations
CRMS-US	3,596	9,005	155,674	299,885
CRMS-World	2,000	5,072	41,816	79,420
Total	5,596	14,077	197,490	379,305

Government Documents Registry

The project team has updated initial use cases based on focus group feedback and drafted initial functional requirements for the registry. The team continues to analyze existing metadata records to define possible methods for identifying duplicate documents. A list of known federal agencies has been compiled and is being used to review the comprehensiveness of sources for name authority records such as VIAF and the LC Name Authority Headings.

HathiTrust Research Center

HTRC co-director Prof. Beth Plale gave a talk at Koninklijke Bibliotheek in Den Haag, Netherlands on Nov 15 on "Opportunities and Challenges of Text Mining HathiTrust Digital Library". Text mining at scale was demoed at IEEE/ACM Supercomputing November 17-22 in Denver, and the HTRC team demonstrated text mining by dynamically deploying a Hadoop cluster onto Big Red II at Indiana University. In early November, the HTRC team held a hands-on session at Indiana University for 25 researchers from a variety of disciplines, hosted by Catalyst Center for digital humanities.

mPach

University of Michigan staff began scoping services to be offered by Michigan Publishing using the mPach platform, and considering arrangements that will need to be made to support use by HathiTrust member institutions or other organizations. To clarify the relationships between various parties potentially using mPach, the list of modules has been revised to separate Submitter from Prepper, and the diagram of the major parts of mPach has been revised to show that Submitter (the mechanism that does final validation before submitting content to HathiTrust) will be operated by Michigan Publishing.

Zephir

Partner submissions of bibliographic metadata have gone smoothly since the transition of bibliographic management from the University of Michigan to the California Digital Library's Zephir system at the beginning of November. CDL has ingested 257,072 new or updated records through Zephir since that time. Information on submitting records to Zephir is available at http://www.hathitrust.org/bib_data_submission for an overview of the bibliographic metadata submission

There's an elephant in the library.™

www.hathitrust.org

process.

Development Updates

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

Development Environment

Staff completed work on web server upgrades for the HathiTrust development environment and began testing of the application codebase in preparation for subsequent upgrades to production servers.

Full-text Search

Staff continued progress to support full-text indexing of JATS content, and indexing of volumes in general into a configurable number of "chunks" to improve relevance ranking of large documents. Staff ran processes to re-index approximately 18,000 volumes affected by the bug reported in last month's update, and approximately 5.5 million volumes whose associated metadata was affected by updates of print holdings received from partner institutions.

Image Server

Staff modified the image server for HathiTrust applications to use Unifont when embedding OCR in PDFs in cases where the language of the volume is not supported by Deja Vu Sans. This will allow more PDFs to be searchable.

Print Holdings

Staff loaded updated holdings information from HathiTrust partners, submitted prior to partner 2014 fee calculations.

Outages

No outages were reported in November.

Total Volumes Added	November	Overall
Boston College	0	2,363
Columbia University	0	65,035
Cornell University	3,607	437,475
Duke University	1	4,525
Harvard University	5	237,435
Indiana University	159	195,579
Library of Congress	0	89,724
North Carolina State University	0	3,196
Northwestern University	176	37,464
New York Public Library	3	288,370
Penn State	1,179	66,491
Princeton University	0	251,709
Purdue University	0	44,695
Texas A&M	26	26
Universidad Complutense	12	112,013
University of California	10,419	3,445,878
University of Chicago	181	35,568
University of Florida	176	9,763
University of Illinois	539	112,965
University of Michigan	2,765	4,665,517
University of Minnesota	690	112,965
UNC - Chapel Hill	0	17,025
University of Wisconsin	43	555,921
University of Virginia	0	50,821
Utah State University	0	117
Yale University	0	23,678
Total	19,981	10,866,212

Public Domain (~32% of total)

Total*	31,247	3,525,560

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

There's an elephant in the library.™

www.hathitrust.org

HathiTrust Digital Library

Update On November Activities

User Support Issues	November	October
Content	**253**	**249**
Quality	244	242
Collections	8	7
Cataloging	**134**	**189**
Access and Use	**105**	**164**
Copyright	49	90
Permissions	6	8
Takedown	0	2
Print on Demand	1	0
Inter-library loan	0	0
Full-PDF or e-copy requests	15	12
Datasets	1	4
Data Availability and APIs	2	2
Reuse of content	7	3
Web applications	**27**	**36**
Functionality problems	4	11
Problems with login specifically	2	1
General questions about login	0	1
Partners setting up login	5	0
Usability issues	0	0
Feature requests	2	2
Partner Ingest	**5**	**3**
General	**66**	**105**
Partnership	5	6
Infrastructure	0	0
Miscellaneous	61	99
Total	**590**	**746**

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

Most-accessed volumes
The psychology of selling and advertising, by Edward K. Strong.
The human figure, by John H. Vanderpoel
The five laws of library science, by S. R. Ranganathan.
Quicksand, by Nella Larsen.
Practical Anatomy of the Rabbit: an Elementary Laboratory Text-book in Mammalian Anatomy, by B.A. Bensley.
The Sunlight Book of Knitting and Crocheting, by Adelaide Gray
Consumption of the Lungs and Kindred Diseases, Treated and Cured by Kerosene, by Charles Oscar Frye.
Why England Slept, John F. Kennedy.
Roster of the Confederate soldiers of Georgia, 1861-1865, v.2.
The Book of a Hundred Hands, by George Brant Bridgman.

There's an elephant in the library.™

www.hathitrust.org