HathiTrust Digital Library

Update On August Activities

September 13, 2013

Top News

New HathiTrust Partners

HathiTrust is very pleased to welcome Allegheny College (view the full press release) and the University of Alabama as its newest partners.

Executive Director Search

by Brian Schottlaeder, The Audrey Geisel University Librarian, University of California, San Diego and Chair, HathiTrust Board of Directors

On behalf of the HathiTrust Board of Directors, I am pleased to announce that the search for the successor to John Wilkin is underway. The Executive Director position will offer the right individual a unique opportunity to help us advance our collective mission and strategic objectives.

The full description of the position, responsibilities, and desired qualifications is available at http://www.hathitrust.org/jobs_executive_director .

Applications should be made through the posting on the University of Michigan jobs site. Nominations may be sent to hathitrust-ed-nomination@umich.edu. Review of applications will begin September 30, 2013 and continue until the position is filled.

I encourage you to share this announcement widely and to think expansively about nominating qualified individuals. Your nomination need include no more than a name; we'll do the rest!

Assistant Director

HathiTrust announces the appointment of Jeremy York as Assistant Director for HathiTrust. Jeremy began working for HathiTrust in 2008, a few months prior to its formal launch, and has been responsible for a broad range of coordinating activities among the partnership.

Government Documents Registry Online Focus Groups

HathiTrust is engaged in an initiative to create a metadata registry for the comprehensive corpus of US federal government documents produced from 1789 to the present. The registry will be available to everyone. (For more information on the project, visit the project page.)

As part of our efforts to define the functionality that will be needed for the registry, we will be holding a series of online focus groups. These sessions are open to anyone who is interested in providing feedback on ways that the registry might be used.

The focus groups will be held on:

- Date 1: 9/23 (Monday); 1-3 pm EST
- Date 2: 9/25 (Wednesday); 4-6 pm EST
- Date 3: 9/27 (Friday); 10-12 am EST

September Forecast

Complete the development of ePub and PDF generation from JATS.

Continue to explore improvements to relevancy ranking.

Work on adding support for indexing of JATS articles.

Papers & Presentations

Jeremy York, "HathiTrust: Key Concepts and Issues in Managing the Digital Archive", ICPSR Summer Workshop, August 1, 2013.

Update On August Activities

- Date 4: 10/1 (Tuesday); 2-4 pm EST
- Date 5: 10/2 (Wednesday); 12-2 pm EST

If you are interested in participating, please email valglenn@umich.edu by September 19th with your two preferred dates/times. If you are interested in giving feedback and are unable to attend any of the scheduled sessions, please contact Valerie Glenn at the above email address.

You can follow HathiTrust on Twitter or Facebook

Subscribe to email updates (via Google Groups)

Ingest

Internet Archive

The University of Connecticut and the University of Illinois at Urbana-Champaign submitted bibliographic metadata for volumes digitized by the Internet Archive in preparation for deposit of the volumes into HathiTrust. HathiTrust corresponded with the Library of Congress, Pennsylvania State University, and the University of Maryland about future ingest of Internet Archive-digitized content.

Locally-Digitized

HathiTrust provided support to Texas A&M University and the University of the Illinois at Urbana-Champaign as they prepared to deposit locally-digitized content into HathiTrust. The University Press of Florida began to make arrangements to deposit backfile publications in HathiTrust on an open access basis, and the University of Pittsburgh renewed conversations about deposit of locally-digitized files.

In September, HathiTrust will begin development on two online validation services designed to help partners prepare locally-digitized content to HathiTrust specifications prior to deposit. The first is a web-based service to interactively validate single image files and is planned to be complete in September or early October. The second is conceived as a cloud storage-based service to validate entire volumes and is planned to be complete in October or November.

Projects

Bibliographic Data Management

The California Digital Library (CDL) team continued to work with staff at the University of Michigan to bring the current bibliographic management system at the University of Michigan and CDL's new Zephir system into parity prior to beginning a parallel phase in which Zephir will shadow the current system over a period of several weeks. Institutions depositing content are contributing records both to the University of Michigan and CDL, and CDL will be working with the User Support Working Group to test new workflows for bibliographic record correction. Please see http://www.hathitrust.org/ingest_checklist for information about submitting records to HathiTrust. Any questions about Zephir or content ingest should be directed to feedback@issues.hathitrust.org.

Update On August Activities

Copyright Review

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

	August		Overall	
	Public Domain	All Determinations	Public Domain	All Determinations
CRMS-US	3,160	7,824	145,928	276,920
CRMS-World	2,697	5,131	34,932	65,061
Total	5,857	12,955	180,860	341,981

HathiTrust Research Center - Author Gender Metadata

Stacy Kowalczyk, Assistant Professor at Dominican University worked with Zong Peng of the HTRC technical team over the summer to identify author gender and make this information available through HTRC. They extracted 606,000 unique personal author strings looking at the nearly 3.2 million bibliographic records in the HTRC. Using the VIAF, census bureau data, and lists of names from several web based sources, the HTRC has a preliminary gender identification of approximately 80% of the public domain corpus (19% from VAIF). The initial findings show that 70% of all personal author name strings are male and 10% are female; the remaining 20% are yet to be identified. Dr. Kowalczyk and HTRC continue to improve the identification rate and verify and validate the initial gender identifications. The author gender information shows up as an attribute of a volume in the user's HTRC workset.

mPach

Staff at the University of Michigan prepared presentations on mPach to be delivered at the RCDL'2013 and JATS-Con 2013 conferences. Staff refined workflows for enabling other institutions to use mPach and discussed challenges associated with rendering TeX that occurs within JATS articles. More information about the mPach project can be found at http://www.hathitrust.org/mpach.

Development Updates

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

Analytics

Staff added event-tracking features to links in HathiTrust that make it possible to filter results in HathiTrust Analytics based on whether a user is logged in from a HathiTrust partner institution or a University of Michigan Friend Account. In-

HathiTrust Digital Library

Update On August Activities

formation about HathiTrust's policies on privacy and logging of user activity is available at http://www.hathitrust.org/privacy.

Data API

Staff released version 2 of the HathiTrust Data API. Documentation of the new version is available at http://www.hathitrust.org/data_api. Some of the differences from version 1 include the abilities to specify the formats of the resources returned and parameters such as the height, width, and size of images. Version 2 of the Data API has been configured to support retrieval of articles and article supplementary material in conjunction with mPach. Version 2 has new URL syntax and a version parameter is required. Version 1 of the Data API is scheduled to be taken out of service on November 1, 2013.

Full-text Search

Staff continued testing new high-performance storage for full-text search and developed a proof-of-concept process for integrating the new storage into daily indexing routines. Pricing for networking equipment to connect the storage to search indexing servers has been received and purchase is underway. Staff expect to install the new equipment and begin performance testing with live data in late September or early October.

Staff continued to test the Solr index's grouping functionality as part of efforts to improve relevancy ranking of full-text search results. Staff also contributed an initial patch to Lucene to correct an issue with the ranking of long documents in the BM25 ranking algorithm.

PageTurner

Staff deployed a new robots.txt allowing search engines to crawl PageTurner and Collection Builder pages with a "noarchive" meta tag.

Outages

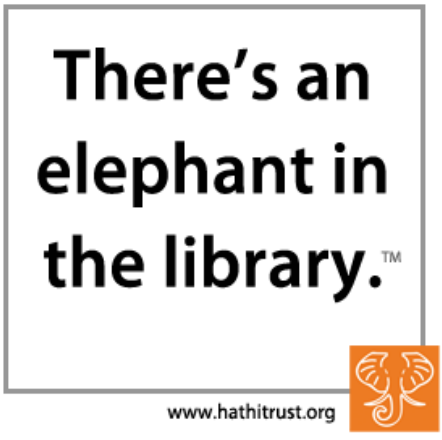
No outages were reported in August.

Total Volumes Added	August	Overall
Boston College	2	2,363
Columbia University	0	65,033
Cornell University	2,738	429,752
Duke University	0	4,523
Harvard University	0	236,069
Indiana University	13	195,349
Library of Congress	0	89,724
North Carolina State University	0	3,196
Northwestern University	939	36,420
New York Public Library	1	288,357
Penn State	710	64,774
Princeton University	0	251,705
Purdue University	0	44,692
Universidad Complutense	1	111,984
University of California	12,084	3,407,326
University of Chicago	468	33,542
University of Florida	5,518	7,586
University of Illinois	5	111,134
University of Michigan	3,310	4,653,823
University of Minnesota	1,835	109,727
University of North Carolina - Chapel Hill	434	17,022
University of Wisconsin	5	555,815
University of Virginia	0	50,817
Utah State University	0	117
Yale University	0	23,678
Total	28,063	10,794,528

Public Domain (~32% of total)

Total*	21,915	3,452,123

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



HathiTrust Digital Library

Update On August Activities

User Support Issues	August	July
Content	**344**	**322**
Quality	335	313
Collections	6	8
Cataloging	**111**	**140**
Access and Use	**183**	**190**
Copyright	120	125
Permissions	4	8
Takedown	1	2
Print on Demand	0	0
Inter-library loan	0	2
Full-PDF or e-copy requests	21	16
Datasets	4	1
Data Availability and APIs	1	0
Reuse of content	4	5
Web applications	**26**	**27**
Functionality problems	9	8
Problems with login specifically	0	3
General questions about login	3	2
Partners setting up login	0	2
Usability issues	1	2
Feature requests	1	2
Partner Ingest	**8**	**5**
General	**64**	**39**
Partnership	7	7
Infrastructure	0	0
Miscellaneous	57	32
Total	**736**	**723**

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

Most-accessed volumes

Title
Roster of the Confederate soldiers of Georgia, 1861-1865, v.1.
Godey's Magazine, v.40-41 1850.
The Magistrates of the Roman Republic, Vol. 1, by T. Robert S. Broughton.
Annual Report and Statements of the Chief of the Bureau of Statistics on the Commerce and Navigation of the United States, 1881/82, by the Treasury Department.
The Mummy! A Tale of the Twenty-Second Century, by Mrs. Loudon.
De Norske Settlementers Historie, by Hjalmar Rued Holand.
Interstate Commerce. Debate in Forty-Eighth Congress, Second Session [-Fiftieth Congress], on the Bill (H.R. 5461) to Establish a Board of Commissioners of Interstate Commerce and to Regulate Such Commerce, etc., Vol. 1.
Il Canzoniere. Riordinato da Luigi Domenico Spadi con le Interpretazioni di Giacomo Leopardi, by Francesco Petrarcha.
A Standard History of Ross County, Ohio, Vol. 2, Ed. Lyle S. Evans.
Radio for the Millions, Prepared by the Editorial Staff of Popular Science Monthly.