HathiTrust Digital Library

Update On March Activities

April 11, 2014

Top News

Orphan Works Roundtable

Sarah Michalak, chair of the HathiTrust Board, and Mike Furlough, incoming Executive Director of HathiTrust, participated in a Roundtable discussion organized by the U.S. Copyright Office on March 10 and 11 on Orphan Works and Mass Digitization. Melissa Levine, Lead Copyright Officer at the University of Michigan Library also participated. HathiTrust will be submitting and posting written comments on the Roundtable issues in April.

HTRC Grant Awards

The HathiTrust Research Center is pleased to announce the recipients of 4 prototyping project awards, granted as part of the Workset Creation for Scholarly Analysis (WCSA) project funded by the Andrew W. Mellon Foundation and directed by J. Stephen Downie (PI), Tim Cole (co-PI), and Beth Plale (co-PI). Each project will receive $40,000 to develop a prototype over a nine-month period beginning in late April. HTRC received 15 proposals in response to an RFP released in November, and eight finalists were invited to present projects at a shortlist meeting in February.

The following prototyping projects have been selected:

- "Workset Creation through Image Analysis of Document Pages", Texas A&M University (PI: Keith Biggers)
- "Semantic Analysis of Documents from the HathiTrust Corpus", Waikato University (PI: Annike Hinze)
- "Distributed Metadata Correction and Annotation", Maryland Institute for Technology in the Humanities, University of Maryland (PI: Trevor Muñoz)
- "ElEPHãT: Early English Print in HathiTrust, a Linked Semantic Workset Prototype", Oxford University (PI: Kevin Page)

These projects represent a range of approaches to developing new tools and techniques designed to assist researchers and scholars in 1) identifying and selecting resources from within the HathiTrust and 2) creating worksets of these resources for scholarly analysis. Approaches range from page image analysis, linked data solutions for developing worksets drawn from multiple sources, semantic analysis to support topical clustering, and application development for metadata correction and annotation. A full press release is forthcoming, and additional project information is available at http://worksets.htrc.illinois.edu/worksets/.

Ingest

General

HathiTrust coordinated with the University of Illinois, Emory University, the University of Washington and the Getty Research Institution on the submission of new

Late Breaking News

HathiTrust released a statement regarding the "Heartbleed bug" reported on April 8.

April Forecast

Begin development of a consolidated application to administer staff who are authorized to access restricted items.

Develop and test new spelling suggestion features.

Build test indexes to experiment with Solr's grouping and block-join functionality at scale (part of work towards relevance ranking improvements).

Papers & Presentations

Downie, J. Stephen. 2014. "Two Projects, One Challenge: Common research data issues in MIREX and HTRC." Dublin, Ireland, March 24, 2014.

Update On March Activities

content, and prepared to receive content from the Sterling and Francine Clark Art Institute Library and Knowledge Unlatched.

Zephir

California Digital Library (CDL) loaded 51,669 new or updated bibliographic records into Zephir.

Working Groups and Committees

Program Steering Committee

The Program Steering Committee (PSC) continued to form working groups and committees to carry out its agenda, including action on ballot initiatives from the Constitutional Convention. The Government Documents Initiative Planning and Advisory Group held an initial meeting in Gainesville, Florida on March 17th-18th; the core group is now being expanded with additional members. The former Collections Committee has been reconstituted with a new charge; confirmed members who are continuing on the committee are Ivy Anderson (chair, and PSC liaison), Sharon Farb, Bryan Skib, Claire Stewart, Tom Teper, and Ann Thornton; two additional members are in the process of being appointed. In April the PSC will issue a call through the HathiTrust member representatives for volunteers to serve on two new groups: a Print Monographs Archive Task Force, and a Rights & Access Working Group.

With Mike Furlough's appointment as Executive Director, the Program Steering Committee has an opening for a new member. A call for nominations will be issued shortly.

Projects

Copyright Review

A summary of the determinations from HathiTrust copyright review activities in March is given below. See CRMS-US and CRMS-World, projects funded by the Institute of Museum and Library Services (IMLS), for further information.

	March		Overall	
	Public Domain	All Determinations	Public Domain	All Determinations
CRMS-US	2,956	3,591	163,968	312,667
CRMS-World	2,766	6,620	52,164	102,366
Total	5,722	10,211	216,132	415,033

You can follow HathiTrust on Twitter or Facebook

Subscribe to email updates (via Google Groups)

Update On March Activities

Government Documents Registry

Project staff conducted manual investigations to determine the comprehensiveness of select titles and agencies in the HathiTrust repository, and considered possibilities for identifying related bibliographic records using automated means. Staff also began to explore methods for determining gaps in holdings based on bibliographic metadata.

HathiTrust Research Center

Staff from Indiana and the University of Michigan discussed options for secure transfer of in-copyright materials from the Michigan repository instance to the HTRC. The HTRC-Usergroup continued its monthly meeting series, supplemented by discussion on the HTRC-Usergroup listserv. To subscribe to the list, visit http://bit.ly/1e0e0Zn.

Matthew Wilkins, assistant professor of English at University of Notre Dame was awarded an ACRL (American Council of Learned Societies) fellowship, for a project using works in HathiTrust to study “Literary Geography at Scale”.

mPach

University of Michigan staff evaluated MeTypseset for possible integration into the Prepper/Norm workflow. With indexing of JATS content implemented, work began on evaluating the UI implications for search results within an item that contains no page breaks.

Development Updates

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

Authentication and Authorization

Staff modified Web applications to use an end user’s Shibboleth entityID to establish their institutional affiliation (eduPersonScopedAffliation was used previously). This was done to facilitate proper identification when a user has multiple affiliations.

Staff began to gather requirements for and design an application to administer staff who are granted special access to restricted materials (e.g., staff authorized to review the copyright status or quality of volumes, or access volumes as a proxy for users who have print disabilities).

Full-text Search

Staff continued to work with storage and network equipment suppliers to troubleshoot and optimize performance issues with new high-performance storage for full-text search.

HathiTrust Digital Library

Update On March Activities

Staff deployed new features in full-text searching indexing, including support for indexing of JATS XML for born-digital article content, and indexing of volumes into a configurable number of "chunks". Chunking is being explored as a part of strategies to improve the relevance ranking of full-text search results.

ImageServer

Staff re-architected the imgsrv application to more efficiently generate derivative formats. The changes impact the generation of derivatives (such as PDF) for items currently in HathiTrust, and will facilitate the creation of derivatives of born-digital articles submitted via mPach. Staff began to make minor improvements to the EPUB download option available from HathiTrust's mobile interface. The improvements will be tested and then released in April.

Spelling-suggester

Staff tested a new spelling-suggestion feature using several large extracts from HathiTrust query logs and lists of commonly misspelled words and their corrections. The spelling suggester provided correct suggestions for nearly all queries. Some outliers revealed needs for further development, however, which is being pursued by staff at the University of Michigan and the California Digital Library.

Server Replacement Cycle

Staff completed security wipes and prepared retired equipment for return to the vendor.

User Support Issues	March	February
Content	**181**	**220**
Quality	168	200
Collections	13	18
Cataloging	**203**	**165**
Access and Use	**212**	**130**
Copyright	144	82
Permissions	13	16
Takedown	2	0
Print on Demand	0	0
Inter-library loan	4	0
Full-PDF or e-copy requests	20	21
Datasets	2	7
Data Availability and APIs	1	0
Reuse of content	4	2
Web applications	**18**	**29**
Functionality problems	7	13
Problems with login specifically	2	13
General questions about login	2	2
Partners setting up login	0	3
Usability issues	0	0
Feature requests	2	2
Partner Ingest	**16**	**2**
General	**101**	**112**
Partnership	5	5
Infrastructure	0	0
Miscellaneous	96	107
Total	**731**	**658**

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

Availability

Repository

Cumulative 12-month availability of repository access: 99.827%*

No outages were reported in March.

* 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

Total Volumes Added	March	Overall
Boston College	315	3,111
Columbia University	0	65,037
Cornell University	43	444,374
Duke University	0	7,258
Harvard University	0	237,435
Indiana University	67	195,647
Keio University	2	88,956
Library of Congress	0	107,929
New York Public Library	3,268	291,640
North Carolina State University	0	3,196
Northwestern University	38	37,639
Ohio State University	1,641	21,068
Penn State	6,599	77,928
Princeton University	0	251,710
Purdue University	0	44,698
Texas A&M	0	1,201
Universidad Complutense	1	112,148
University of California	14,675	3,476,598
University of Chicago	79	39,156
University of Florida	0	9,765
University of Illinois	7,863	134,466
University of Massachusetts, Amherst	680	9,411
University of Michigan	2,078	4,670,559
University of Minnesota	91	119,859
UNC - Chapel Hill	0	17,025
University of Virginia	0	50,821
University of Wisconsin	26	555,973
Utah State University	0	117
Yale University	0	23,678
Total	37,475	11,098,430

Public Domain (~33% of total)

Total*	6,887	3,682,091

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

Most-accessed volumes

- Quintus Curtius [History of Alexander], Vol. 1, with an English translation by John C. Rolfe.
- Quicksand, by Nella Larsen.
- Science and life, by Robert Andrews Millikan.
- The Human Figure, by John H. Vanderpoel
- Quintus Curtius [History of Alexander], Vol. 2, with an English translation by John C. Rolfe.
- With the Turk in wartime, by Marmaduke Pickthall.
- Consumption of the Lungs and Kindred Diseases, Treated and Cured by Kerosene, by Charles Oscar Frye.
- Roster of the Confederate soldiers of Georgia, 1861-1865, v.3.
- 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.1.

There's an elephant in the library.™

www.hathitrust.org