HathiTrust Digital Library

# Update On Spring Activities

June 21, 2016

## Top News

### HTRC Expanding Services to Entire HathiTrust Collection

HathiTrust Research Center is proud to announce expanded services to support computational research on the entire corpus of HathiTrust— over 14 million digitized volumes, including more than 7 million books, more than 725,000 U.S. federal government documents, and more than 350,000 serial publications. Previously, HTRC supported analysis of only the public domain subset of the HathiTrust collection, but is now the only place where scholars can perform text mining on the entire HathiTrust collection.

“The big data infrastructure of HTRC ensures that researchers will retain access to the collection even as it grows in size,” said Beth Plale, Indiana co-director of HTRC and professor of informatics and computing at Indiana University. “A researcher carrying out text mining on millions of texts needs both tools and the help of HTRC experts in high performance mining techniques. HTRC research staff bridge the gap between the researcher and the data.”

At first, researchers will be able to access the HTRC collection through its Advanced Collaborative Services grants. This peer-reviewed grant process gives awardees dedicated HTRC staff time, and will be the initial path for non-consumptive research of the full corpus. HTRC expects to make the full collection available through its secure HTRC data capsules in spring 2017. A features data set, derived from the full collection at both volume level and page level, will be released in fall 2016. For more, see this article from Indiana University.

### HathiTrust Print Disability Service Expansion

As of April 25th, print disability service proxies are able to access all copyrighted material in HathiTrust on behalf of users with print disabilities. We made this change after a thorough review of our policies and U.S. copyright law.

Previously, proxies were limited to accessing only the HathiTrust materials that matched books held in their library. This situation frequently created confusion about what was and was not available at a given HathiTrust member institution. The change to a uniform service for all members will broaden access and make the process significantly more straightforward.

Users with print disabilities will now have access to the entire HathiTrust collection, whereas previously they may have only been able to access somewhere between 46,000 and 6.9 million items (depending on the size of their library). This vastly increases the value of this service, and we hope the change will encourage greater use of this service on your campuses.

SAVE THE DATE: 2016 HathiTrust Member Meeting

**When:** Thursday, November 10, 2016
**Where:** Big Ten Center, Chicago

HathiTrust on the Road

HathiTrust staff will be attending the following events in summer 2016. Please contact us if you wish to meet us at any of these events:

ALA Annual Conference, Orlando, FL, June 23-28 - Heather Christenson, Valerie Glenn, Lizanne Payne, Mike Furlough

Digital Humanities 2016, Krakow, Poland, July 11-14 - J. Stephen Downie, Peter Organisciak, Sayan Bhattacharyya

### CRMS Wins ALA's Ray Patterson Copyright Award

The Copyright Review Management System (CRMS) is the winner of the L. Ray Patterson Copyright Award from The American Library Association (ALA). The award is given to a person or group that demonstrates dedication to a balanced U.S. copyright system through advocacy for a robust fair use doctrine and public domain. ALA's announcement can be read here: http://www.ala.org/news/press-releases/2016/05/ala-announces-2016-winner-l-ray-patterson-copyright-award.

This is the first time a group has been named winner of the award acknowedgement that CRMS has truly been a team effort. Through the investment and commitment of your staff who have dedicated a portion of their time to participate in copyright review, we have collectively reviewed over 600,000 items, identifying and making available 320,000 public domain works. This project has been generously funded by the Institute for Museum and Library Services starting in 2008, based on an initial proposal submitted by John Wilkin, then Founding Executive Director of HathiTrust and now Juanita J. and Robert E. Simpson Dean of Libraries and University Librarian, University of Illinois at Urbana-Champaign.

Dozens of people have been involved with CRMS as advisors, managers, or reviewers since its launch in 2008. A full list of these can be found online at: https://www.hathitrust.org/copyright-review-management-system-crms-partner-institutions.

*A Comic History of England.* *http://hdl.handle.net/2027/uc1.l0059606491 ...*

### Electronic Access and the "Collective Collection"

"Electronic Access and the 'Collective Collection,'" a talk delivered by Executive Director Mike Furlough at the 2016 Center for Research Libraries Collections Forum, "@Risk: Stewardship, Due Diligence, and the Future of Print" has been published on our Perspectives from HathiTrust blog. In the paper Mike discusses HathiTrust's plans for shared print monographs archiving and speculates how research libraries can collectively develop vision for both print preservation and digitization that will sustain us over the next twenty years.

### Board of Governors Update

The Board of Governors held its Winter 2016 meeting by phone on March 7th and Spring 2016 meeting in person at the Big Ten Center in Chicago on June 2nd.

During its March meeting, the Board reviewed the calendar for 2016, took action to extend the term of Program Steering Committee members from two years to three, set the term of the PSC chair to two years, and discussed the pending change to policies governing access for users with print disabilities.

At the June meeting, the Board received reports on several topics, including 1) the Subcommittee on Membership and Finance, which is working to assess the HathiTrust financial model and membership criteria; 2) plans for shared print programs by the new Shared Print Program Officer Lizanne Payne; 3) a brief report on the U.S. Federal Documents Registry from Heather Christenson, the new Program Officer for Federal Documents and Collections 4) copyright review planning, and 5) Digital Preservation Network replication services. The Board also discussed and approved recommendations from the Collections Committee and the Program Steering Committee resulting from the 2015 survey of members' collection priorities, and reviewed potential future policy changes to services for users with print disabilities.

*Her Ladyship's Elephant by D.D. Wells http://hdl.handle.net/2027/uc2.ark:/13960/t1rf5mz65 ...*

## Program Steering Committee Addresses Planning Briefs

John Butler from the University of Minnesota began a two-year term as chair of the Program Steering Committee in March, taking over from Robert Wolven from Columbia University.

The Program Steering Committee (PSC) continues to establish and monitor the work of its committees, and advisory and working groups, as progress continues on the major areas of focus identified in the four Planning Briefs presented to the membership in fall 2014 (i.e., Quality and Validation Issues, Print Disability Services, Metadata Strategy and Policy, and Investigating Format Expansion). The Committee also has been giving renewed attention to the last of the 2011 Constitutional Convention Ballot Proposals to be acted on: Framework for Development Proposals.

Specific highlights include:

The Collections Committee submitted to the PSC its Collection Priorities Survey Analysis Final Report, which provides results and recommendations on the Fall 2015 members-wide survey on HathiTrust collections issues and priorities. The report is currently under review with distribution planned by the end of the summer.

Revision of the HathiTrust Commitment to Quality statement, which includes identification of stakeholders, prominent use cases, related quality issues and potential improvement strategies.

Appointment of the following individuals to three-year terms on the Collections Committee, beginning in July:

- Mildred Jackson, Head of Collection Strategy and Development, University Alabama
- Jeff Kosokoff, Head of Collection Strategy and Development, Duke University
- Michael Neubert, Supervisory Digital Projects Specialist, Collections Services Directorate, Library of Congress
- Nicholas Wolf, Research Data Management Librarian, New York University

There's an elephant in the library.™

www.hathitrust.org

## HathiTrust Research Center

### HTRC Welcomes Yu “Marie” Ma as new Development Operations Manager

HathiTrust Research Center is delighted to welcome to their team the new DevOps Manager, Yu “Marie” Ma. Dr. Yu (Marie) Ma, whose Ph.D. is in Computer Science from Indiana University (2006), joined Indiana University’s University Information Technology Services in 2006 where she has been supporting and leading academic research activity as a member of the Science Gateways Group. She has played leadership and collaborative roles in a wide range of research projects within Indiana University and across the nation including those funded by NASA, USGS, and the NSF-funded large-scale Extreme Science and Engineering Discovery Environment (XSEDE) project. Marie brings years of rich experience in both research and user support in areas such as scientific data management, computational cyberinfrastructure, science gateways and cloud computing, and has written numerous publications on these topics as well. HTRC is extremely pleased to have someone of her caliber and accomplishments in this central role, and look forward to her successful work with the team now and in the future.

## Ingest

### Zephir Update

In March and April 2016, the Zephir Metadata Management System loaded 1,203,237 new and 442,573 updated records from HathiTrust contributors for 49 unique content streams.

## Projects

### Copyright Review

A summary of the determinations from HathiTrust copyright review activities in Spring 2016 is given below. See CRMS-US and CRMS-World for further information.

| | March-May | | Overall | |
|---|---|---|---|---|
| | Public Domain Determinations | All Determinations | Public Domain Determinations | All Determinations |
| CRMS-US | 1,680 | 2,417 | 179,078 | 334,316 |
| CRMS-World | 5,021 | 9,317 | 149,754 | 281,617 |
| Total | 6,701 | 11,734 | 328,832 | 615,933 |

You can follow HathiTrust on Twitter or Facebook

Subscribe to email updates (via Google Groups)

*Cactus culture for Amateurs.* William Watson, 1903. *http://hdl.handle.net/2027/msu.31293201080128 ...*

### Volumes Added

Ingest numbers and Collection statistics are updated daily and can be found on our website here: https://www.hathitrust.org/visualizations_deposited_volumes_current

# Update On Spring Activities

June 21, 2016

### U.S. Federal Documents Registry

The U.S. Federal Documents Registry is now undergoing testing to be a beta version. The Registry is updated daily with new or updated records from the HathiTrust repository. The interface has been enhanced so that it is as accessible as other HathiTrust interfaces, and each Registry record has a persistent unique identifier.

The Registry contains roughly 5.5 million records, with many known duplicates. Project staff continue to work on refining duplicate detection, focusing mostly on item description (enumeration and chronology). Staff have also begun to develop sample needs lists based on Registry records that do not contain a HathiTrust ID.

## Development Updates

### Full-text Search

The Core Services team continued an exploratory analysis of query and click logs of HathiTrust usage. The combined work of characterizing user tasks and analyzing the click logs will lay the groundwork for future testing of new features, simplify user tasks and future testing of measures to improve relevance ranking. Preliminary results indicate that some additional logging features need to be added to the logging framework.

In April, work began on preparation and testing for re-indexing all 14 million volumes. The new index will include fields which will allow us to provide more accurate language facets, and test several new features. The new index will use a Solr index plug-in that will use less memory. This will allow us to begin testing alternative relevance-ranking algorithms.

### Collection Builder

We have added functionality to the Collection Builder that allows users to download the item metadata for any collection; the download reflects the displayed/filtered item list (all, full-text, search results, etc.).

### PDF Downloads

We have updated how we monitor the building of requested PDFs to improve how load balancing affects download functionality.

After discussions with University of Michigan accessibility consultants, we have deployed changes to the PDF production pipeline: PDFs of scanned image volumes use layers for watermarks and add contents outlines. HathiTrust now can serve born-digital PDFs from the repository after attaching watermarks (currently limited to Knowledge Unlatched items).

*Original pen and ink sketches* by Phil May. Original book price: one shilling. *http://hdl.handle.net/2027/uc2.ark:/13960/t58c9tf0m ...*

### Repository Availability

Cumulative 12-month availability of repository access: 99.975%.

There's an elephant in the library.™

www.hathitrust.org

### Architecture and Engineering

Work was completed to replace all HathiTrust storage. At each of the Michigan and Indiana sites, staff retired 30 Isilon X200 nodes with a total of 1PB of storage and replaced them with 13 Isilon X410 nodes with a total of 1.6 PB storage.

Architecture & Engineering continued planning to implement an improved storage networking and data center layout for HathiTrust equipment.

Michigan staff deployed expanded usage of the HTTPS protocol. Access to all HathiTrust services now uses HTTPS.

### Forecast

Continue work on a unified logging framework for HathiTrust applications.

Work to take fuller advantage of Shibboleth and remove isolated institution specific dependencies on Cosign.

Research ways to support alternative text formats.

Accessibility audit of HathiTrust apps/interfaces in Production.

## Papers and Presentations

### Presentations

Bhattacharyya, Sayan. "Small data and big data: The reflective in the context of text analysis and the humanities classroom." Part of panel on "What Do Comparative Literature and Digital Humanities Have To Say To Each Other? A Critical Approach." Annual Conference of the American Comparative Literature Association (ACLA), Harvard University, March 17-20, 2016. Abstract (Google Doc), slides (PDF)

Underwood, Ted, "Literary History and Machine Learning in Dialogue about Genre," Spring Symposium, UIUC Center for Advanced Study, 4 April 2016.

*Ruins of Dunluce Castle, Antrim* http://hdl.handle.net/2027/uc2.ark:/13960/t24b2zj11 ...

*Her Majesty's Tower by Hepworth Dixon* http://hdl.handle.net/2027/uc2.ark:/13960/t2t43pv37 ...

There's an elephant in the library.™

www.hathitrust.org

# HathiTrust Digital Library

## Update On Spring Activities

June 21, 2016

### Most-accessed volumes Spring 2016

- The surnames of Scotland, their origin meaning and history, by George F. Black.
- Quicksand, by Nella Larsen.
- Text-book of ordnance and gunnery, by William Freeland Fullam.
- McClure's Magazine, v.11, 1898 May-Oct.
- America is in the heart, a personal history, by Carlos Bulosan.
- Solid mensuration, by Willis F. Kern and James R. Bland.
- Il Poligrafo Domenica, v. 3, Jan.-June 1812.
- A text-book of the construction and manufacture of the rifled ordnance in the British service, by Captain Francis S. Stoney and Lieut. Charles Jones
- Return to life through contrology, by Joseph H. Pilates and William John Miller

*Hospital Construction and Management, 1883.* *http://hdl.handle.net/2027/gri.ark:/13960/t86h8tf95 ...*

| User Support Issues | Mar-May | Jan-Feb |
|---|---:|---:|
| **Content** | **127** | **272** |
| Quality | 108 | 250 |
| Collections | 17 | 21 |
| **Cataloging** | **155** | **227** |
| **Access and Use** | **346** | **298** |
| Copyright | 156 | 120 |
| Permissions | 28 | 19 |
| Takedown | 2 | 1 |
| Print on Demand | 0 | 1 |
| Inter-library loan | 2 | 2 |
| Full-PDF or e-copy requests | 69 | 50 |
| Datasets | 8 | 8 |
| Data Availability and APIs | 2 | 7 |
| Reuse of content | 19 | 13 |
| **Web applications** | **74** | **66** |
| Functionality problems | 46 | 31 |
| Problems with login specifically | 6 | 10 |
| General questions about login | 0 | 1 |
| Partners setting up login | 1 | 1 |
| Usability issues | 0 | 0 |
| Feature requests | 1 | 4 |
| **Partner Ingest** | **65** | **37** |
| **General** | **225** | **237** |
| Partnership | 19 | 25 |
| Miscellaneous | 206 | 212 |
| **Total** | **1150** | **1137** |

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