HathiTrust Digital Library

# Update On July Activities

August 10, 2012

## Top News

### HathiTrust Training and Information Sessions Survey

HathiTrust has offered webinars in the past to orient new partners and provide updates on HathiTrust services and initiatives. For our next series, or program, we are considering offering information sessions led by staff throughout the partnership on topics of interest. To begin to plan for these sessions, we would like to receive feedback from members of partner institutions on 4 questions related to session topics, venue, and participation. A form with the questions is available at http://tinyurl.com/8n3k9nr. Although feedback is especially sought from partner institutions, others may provide input. If there is sufficient interest we will consider offering sessions open to anyone, whether affiliated with HathiTrust partner institutions or not. Responses are requested by September 21, 2012.

### HathiTrust Accessibility Initiative

University of Michigan developers, under the guidance of Michigan's User Experience Department, are in the process of reviewing and making improvements to the accessibility of HathiTrust Web applications. The first phase of the work, which began in July, involves ensuring compliance of HathiTrust interfaces with the Web Content Accessibility Guidelines (WCAG) 2.0, Level A. The second phase will target compliance with WCAG 2.0 Level AA, and include usability testing by users who have print disabilities. It is expected as part of this work that Michigan staff will begin to draft policies and guidelines to ensure that future coding for HathiTrust applications maintains these standards.

Staff from partner institutions with Web accessibility expertise who are interested in being involved with this initiative are encouraged to contact Suzanne Chapman (suzchap@umich.edu).

### Copyright Review Update

Over the last several of years, staff from HathiTrust partner institutions have been manually reviewing the copyright status of volumes published in the United States from 1923 to 1963, as part of CRMS-US, an IMLS funded grant project. The grant has come to an end but review by partner institutions continues, and has been expanded through a second IMLS grant, CRMS-World, that is targeting review of non-US-published works, beginning with English-language works published in the United Kingdom, Canada, Australia, and Spain. Through this process tens of thousands of works have been discovered to be in the public domain and opened in HathiTrust to viewers world-wide. Reports on the number of volumes reviewed and opened as of early Augsut are shown below and will be included in future updates.

| | Reviewed | Opened |
|---|---|---|
| CRMS-US | 309,090 | 164,222 |
| CRMS-World | 9,459 | 4,169 |
| Total | 318,549 | 168,391 |

### August Forecast

Continue accessibility work

Complete and release revised Data API documentation

Begin work on Data API enhancements to serve content derivatives

Continue work on full-text search relevance ranking

### Papers & Presentations

Jeremy York: "HathiTrust and TRAC", July 25, 2012.

Jeremy York: "HathiTrust: On TRAC", July 26, 2012.

# HathiTrust Digital Library

## Update On July Activities

## Ingest

### Local Digitization

Michigan staff provided support as needed for the new ingest tools that have been made available. Staff who have questions about using the tools, or who would like to initiate deposit of materials should contact feedback@issues.hathitrust.org.

### Internet Archive

HathiTrust began ingest of a first set of Internet Archive-digitized volumes from Penn State and a second set of Internet Archive-digitized volumes from the University of Illinois.

## Working Groups and Committees

Working groups and committees in HathiTrust may have an operational or strategic focus. See http://www.hathitrust.org/working_groups for more information.

## Operational

### User Support Working Group

A summary of the issues received by the User Support Working Group in July is provided at the end of the update.

| Total Volumes Added | July | Overall |
|---|---|---|
| Columbia University | 0 | 64,184 |
| Cornell University | 3,492 | 403,448 |
| Duke University | 0 | 4,523 |
| Harvard University | 0 | 234,346 |
| Indiana University | 5 | 187,669 |
| Library of Congress | 305 | 89,721 |
| North Carolina State University | 0 | 3,196 |
| University of North Carolina - Chapel Hill | 0 | 8,088 |
| Northwestern University | 1 | 7,208 |
| New York Public Library | 3 | 259,563 |
| Penn State University | 661 | 43,983 |
| Princeton University | 14 | 250,863 |
| Purdue University | 0 | 27,687 |
| University of California | 6,355 | 3,346,583 |
| University of Chicago | 408 | 22,439 |
| University of Illinois | 4,027 | 100,178 |
| Universidad Complutense | 0 | 111,828 |
| University of Michigan | 5,302 | 4,551,770 |
| University of Minnesota | 96 | 100,396 |
| University of Wisconsin | 25 | 539,236 |
| University of Virginia | 0 | 48,922 |
| Utah State University | 0 | 90 |
| Yale University | 0 | 23,678 |
| Total | 20,694 | 10,429,599 |

Public Domain (~30% of total)

| | | |
|---|---|---|
| Total* | 22,057 | 3,127,644 |

*Includes volumes opened through copyright review and rights holder permissions.

## Projects

### Bibliographic Data Management

California Digital Library (CDL) staff continued working with staff at the University of Michigan to develop processes to sync rights determination information between Zephir and the HathiTrust rights database. CDL also worked with Michigan to test new bibliographic submission guidelines and a new workflow for submitting bibliographic metadata to Zephir via FTPS. HathiTrust members currently depositing content will soon be asked to participate in a test of this new submission process, which will be put in place when the cutover from the bibliographic management system at the University of Michigan to Zephir takes place. Zephir development is in its final phase, and in the coming months Michigan and CDL will move to an integration phase that will involve extensive testing and operation of the two systems in parallel before a final cutover.

# Update On July Activities

## HathiTrust Research Center

The HTRC UnCamp is a little over a month away. The UnCamp is part informational, part community building, part boot-camp, and part unconference, designed to show the research capabilities that the HathiTrust Research Center can offer and garner feedback from a broad range of interested users. An exciting list of speakers includes John Wilkin, Executive Director of HathiTrust, Colin Allen, Professor of History and Director of the Cognitive Science Program at Indiana University, and Ted Underwood, Associate Professor of English at the University of Illinois. Details on the UnCamp can be found at http://d2i.indiana.edu/htrc/uncamp2012/.

The UnCamp marks the 12-month point in the development of the HTRC, and contributes to a milestone set out in the MOU between the HTRC and HathiTrust of a demonstration of HTRC functionality 12 months into development. The HTRC is scheduled to transition into production in Spring 2013, at the 18-month mark from its inception, so the UnCamp is timely for gathering community input.

The HathiTrust Research Center is pleased to have recently received two allocation awards from XSEDE for computational resources: one for exploratory research and the other in support of educational and outreach activities.

## IMLS Quality Grant

Project staff focused work in July on finalizing the quality review datasets assembled over the course of the grant for analysis, and on developing a framework for reviewing and certifying the quality of volumes in HathiTrust. Staff also finished assembling a catalog of frequently-observed errors in illustrative content for in-depth analysis by an expert in digital conversion errors. The second meeting of the grant Advisory Board and project collaborators is scheduled for the end of August. The project team will present findings at the meeting and solicit feedback and input on direction from the attendees.

## mPach

University of Michigan staff completed a diagram of the mPach system architecture, which has been added to the mPach website. Staff also refined the schema for mapping bibliographic data from JATS XML (the format to be used for encoded text) to MARC records, which are required for ingest into HathiTrust. Work continued on wireframes for the Dashboard module (see a description of all modules), on the profile for the METS file that will accompany digital objects in Submission Information Packages, and on rendering XML articles in the HathiTrust PageTurner.

# Development Updates

## Collection Builder

University of Michigan staff made changes to the Collection Builder Web application to allow longer collection titles and descriptions, and notify users when entries exceed the allowed lengths.

You can follow HathiTrust on Twitter or Facebook

Subscribe to email updates (via Google Groups)

### Data API

Michigan staff worked to update documentation of the Data API to reflect changes reported in the Update on April Activities, as well as other recent changes. The documentation will be made available in August. Documentation of the interactions of existing clients with the Data API is also being developed.

Michigan staff have begun planning to extend the Data API to support requests for derivative forms of content including scaled, individual, page images and PDFs. These features will initially serve quality review and print-on-demand applications but are expected to have other uses as well. Access will be subject to Data API authorization requirements.

### Full-text Search

Programmers at the California Digital Library added language-aware relevance ranking to the search spelling suggestion feature under development for full-text search. Staff also built a regression framework for testing algorithm changes and began to make changes using heuristics to improve spelling suggestion quality for a test set of 100 HathiTrust queries. Over the coming weeks, additional changes to the scoring system for suggestions are expected to further improve the quality of suggestion results.

Michigan staff completed the second phase of planned improvements to indexing and searching of Chinese, Japanese, and Korean (CJK) languages in HathiTrust. The second phase involved re-indexing all volumes in the repository with a new schema to provide better searching over bibliographic data for CJK materials. Staff made corresponding changes in HathiTrust Web application code to take advantage of the new schema. Staff observed that the bug discovered in the first phase of work (reported in the Update on June 2012 Activities) caused the level of improvement to be less than expected in the second phase as well. Work will continue to address this issue. In the meantime, the improvements made to full-text indexing in June reduced the time needed to index all 10.4 million volumes by nearly half – to approximately one week – despite the fact that there were interruptions.

### First Full-Repository Metadata Upgrade

University of Michigan staff made preparations to begin the first repository-wide upgrade of metadata for HathiTrust objects. The upgrade applies primarily to PREMIS metadata, though metadata in other areas of the HathiTrust METS file will be affected as well. In conjunction with this upgrade, HathiTrust will begin moving toward a formalized model of publicly communicating planned full-repository changes.

# HathiTrust Digital Library

## Update On July Activities

### Process for Item Deletion

Deletions from the HathiTrust repository are rare, occurring in instances where

- A volume is either wholly unusable due to quality problems, or a superior copy of the volume is available in HathiTrust (such deletions are authorized by the depositing institution) or
- Removal is requested by the rights holder.

Michigan staff are in the final stages of implementing an automated process (though the process must be initiated manually) to remove items from repository storage, as well as catalog and full-text indexes. In cases where volumes are deleted, a "tombstone" is created for provenance purposes and to maintain permanent links for references users may have created.

### Outages

HathiTrust may have been unavailable for some users on Monday, July 16 from 8:15pm to 8:30pm due to a database locking problem at one repository instance, and from Monday, July 16 at 11:45pm to Tuesday, July 17 at 8:15am due to a problem with a web server at one instance.

HathiTrust sends notice upon discovery and resolution of unscheduled outages and in advance of scheduled outages and maintenance work that may result in an outage. We welcome and encourage additional recipients for these notices. If your institution is not receiving outage notifications and would like to, please contact feedback@issues.hathitrust.org.

| User Support Issues | July | June |
|---|---|---|
| **Content** | **326** | **237** |
| Quality | 318 | 228 |
| Non-partner Digital Deposit | 0 | 0 |
| Collections | 4 | 6 |
| **Cataloging** | **113** | **31** |
| **Access and Use** | **112** | **123** |
| Copyright | 66 | 69 |
| Permissions | 16 | 20 |
| Takedown | 1 | 2 |
| Print on Demand | 4 | 0 |
| Inter-library loan | 6 | 0 |
| Full-PDF or e-copy requests | 16 | 20 |
| Datasets | 4 | 2 |
| Data Availability and APIs | 0 | 0 |
| Reuse of content | 3 | 2 |
| **Web applications** | **27** | **19** |
| Functionality problems | 3 | 2 |
| Problems with login specifically | 0 | 3 |
| General questions about login | 1 | 0 |
| Partners setting up login | 0 | 1 |
| Usability issues | 12 | 0 |
| Feature requests | 2 | 6 |
| **Partner Ingest** | **2** | **3** |
| **General** | **108** | **72** |
| Partnership | 7 | 14 |
| Infrastructure | 0 | 0 |
| Miscellaneous | 101 | 59 |
| **Total** | **688** | **485** |

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

There's an elephant in the library.™

www.hathitrust.org