HathiTrust Digital Library

Update On April Activities

May 11, 2012

Top News

Automatic Partner Login

Staff at the University of Michigan have developed functionality that allows users from partner institutions to be “automatically” logged into HathiTrust when following links from local institutional catalogs or other resources. Permanent links to HathiTrust volumes can now be wrapped with a single sign-on URL that automatically passes users through their own institution’s authentication service. Users who are not already authenticated are prompted to do so. Documentation of the new functionality is available at http://www.hathitrust.org/automatic_login. Thanks to Johns Hopkins University for suggesting this enhancement.

Ingest

Local Digitization

Staff at Michigan continued to work on tools that content depositors can use to create and validate locally-created content packages prior to submission to HathiTrust. The tools will available to partner institutions in May.

Google

HathiTrust began ingest of Google-digitized content from the University of Illinois in April, bringing in more than 80,000 volumes.

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

Communications

The Communications Working Group continued regular activities and development of a briefing for the new Board of Governors. New communication initiatives are awaiting the transition to the new Board.

User Experience Advisory Group

The UX Advisory Group revisited issues related to the labeling of PDF download options in PageTurner. The group’s recommended changes aim to clarify when full PDF downloads are or are not available. The changes are under development and will be implemented in May.

May Forecast

Rebuild the full-text index with CJK (Chinese, Japanese, Korean) indexing improvements; to be completed in May or June

Papers & Presentations

Stacy Kowalczyk, DLP: The HathiTrust Research Center: An Overview, April 4, 2012.

Jeremy York, Access Services in the Age of Mass Digitization, April 20, 2012.

See http://www.hathitrust.org/papers for all papers, presentations, and reports.

HathiTrust Digital Library

Update On April Activities

User Support Working Group

The adjacent table shows a summary of the issues received by the User Support Working Group in April.

User Support Issues	April	March
Content	**231**	**203**
Quality	222	193
Non-partner Digital Deposit	1	0
Collections	4	9
Cataloging	**33**	**49**
Access and Use	**112**	**195**
Copyright	76	137
Permissions	7	17
Takedown	2	1
Print on Demand	0	0
Inter-library loan	0	2
Full-PDF or e-copy requests	10	19
Datasets	4	2
Data Availability and APIs	1	2
Reuse of content	1	6
Web applications	**14**	**11**
Functionality problems	4	4
Problems with login specifically	1	1
General questions about login	4	3
Partners setting up login	3	3
Usability issues	0	0
Feature requests	0	0
Partner Ingest	**0**	**5**
General	**129**	**101**
Partnership	5	7
Infrastructure	0	0
Miscellaneous	124	94
Total	**519**	**559**

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

Projects

Bibliographic Data Management

California Digital Library created prototype exports of the metadata that will be used to populate HathiTrust's tab-delimited inventory files ("hathifiles") and bibliographic catalog. Timing tests for these exports were also conducted. The CDL team continued to reconcile bibliographic records in Zephir with records in the current system at the University of Michigan to ensure all the data is accounted for, addressing record discrepancies and ingest errors as encountered. The team has also begun development of a process to sync rights information in Zephir (the new management system) with the HathiTrust rights database.

jPach (formerly HathiTrust Publishing)

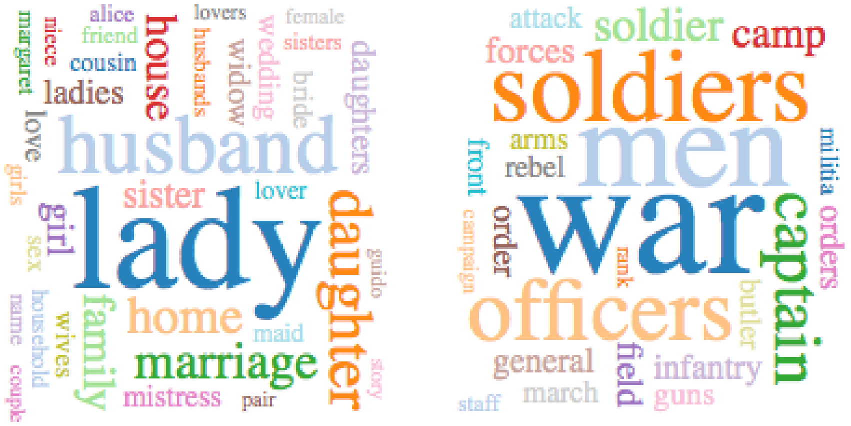
University of Michigan staff continued work on modifications to the HathiTrust PageTurner to display JATS XML. jPach's Norm module (see descriptions of all jPach modules) can now extract 15 common components of a journal article, plus embedded media, from a DOCX file and create valid JATS with references to associated media files. A specification for a Submission Information Package for jPach content is nearly complete and will be posted to the jPach website soon. Work has begun on developing wireframes for the Dashboard module. A timeline for the project is available on the HathiTrust jPach project page.

HathiTrust Research Center (HTRC)

The HTRC completed the agreements necessary with Google to receive Google-digitized public domain volumes from the HathiTrust repository and make them available for computational purposes. With the Google agreements and a Memo of Understanding with HathiTrust in place, the HTRC is actively working with staff at Michigan to bring in the complete set of more than 2.9 million public domain volumes in HathiTrust. Preparation for the transfer includes setup of disk storage and compute nodes at Indiana University (IU), which is being done in collaboration with IU Research Technologies. All computation on HathiTrust volumes will be carried out on HTRC machines; the HTRC itself will not make content available for download. Users interested in receiving texts should follow the directions at http://www.hathitrust.org/datasets.

HTRC was represented at the recent Committee on Institutional Cooperation Digital Humanities Summit in Nebraska. Many attendees were already aware that the HTRC was a digital scholarship initiative of HathiTrust; brochures were on hand to provide a deeper level of detail.

The HTRC has created Meandre workflow components (Meandre is part of the SEASR infrastructure) that retrieve texts from the HTRC using the HTRC data API, spell-check the texts, correct OCR errors, and then perform topic modeling on the texts. The HTRC has demonstrated this functionality, creating topic models of all pages returned from the data API from single-word queries on a full-text index of volumes. For example, a search for "dickens" in the non-Google digitized public domain corpus returns more than 100 topics with associated keywords. The diagrams below show tag clouds of keywords for the topics "lady" and "men".



IMLS Quality Grant

Project staff completed whole-volume review of the first 1,000-volume sample (1,000 English language, pre-1923 volumes digitized by Google), and over 70% of the second 1,000-volume sample (1,000 English language, post-1923 volumes digitized by Google). Approximately 150 volumes (15%) from each of the two samples were coded by two reviewers for quality assurance. The project team decided to perform whole-volume review on the same volumes sampled earlier in the project for page-level review in order to allow for comparison and more in-depth analysis of the data, and yield a better understanding of error within the volumes.

As of the end of April, staff had completed page-level review of approximately 50% of the fourth 1,000-volume sample (1,000 non-Roman language volumes including Korean, Chinese, Japanese, Arabic, Hebrew and Cyrillic).

In the months to come, the focus of the project team will shift away from data collection to data analysis and reporting, and use-case studies research. More information about this research is forthcoming. In May, the team will focus on developing a sub-study to better identify and describe errors in illustrative content. The project website has been updated to report initial findings. See the links under “Quality Review” at http://hathitrust-quality.projects.si.umich.edu/results.htm.

Total Volumes Added

	April	Overall
Columbia University	1	64,184
Cornell University	4,181	396,537
Duke University	0	4,523
Harvard University	0	53,675
Indiana University	3	187,638
Library of Congress	0	89,416
North Carolina State University	0	3,196
University of North Carolina - Chapel Hill	0	8,088
Northwestern University	383	7,203
New York Public Library	20	259,557
Penn State University	28	43,308
Princeton University	50	250,839
Purdue University	799	24,780
University of California	202	3,329,971
University of Chicago	7,251	20,457
University of Illinois	80,642	96,146
Universidad Complutense	4	111,827
University of Michigan	5,011	4,534,989
University of Minnesota	86	95,150
University of Wisconsin	1	534,871
University of Virginia	1	48,922
Utah State University	0	90
Yale University	0	23,678
Total	98,663	10,189,045

Public Domain (~28% of total)

Total*	96,091	2,880,037

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

Development Updates

Data API

University of Michigan staff deployed the security enhancements described in the Update on March 2012 Activities, and the Data API now supports the use of oAuth 1.0-signed requests. As outlined in the March update, there will be a transition period, ending October 1, 2012, during which signed access to the Data API will be possible but not required. After October 1, all requests to the Data API will need to be properly signed with an access key provided by HathiTrust. HathiTrust has created a Web client that employs a user’s login credentials as a proxy for these keys to facilitate non-programmatic uses. Complete documentation of the security enhancements, methods of obtaining keys, signing requests, and accessing the Web client is forthcoming.

Also effective October 1, the host “services.hathitrust.org” will no longer exist for the Data API. The new host will be “babel.hathitrust.org”, the same host as the PageTurner and other HathiTrust services. Calls to the Data API will therefore need to use URLs such as the following (note the additional “cgi” in the path):

http://babel.hathitrust.org/cgi/htd/meta/mdp.39015019203879

rather than

http://services.hathitrust.org/htd/meta/mdp.39015019203879

Full-text Search

HathiTrust released the second phase of advanced full-text search functionality in April. Users can now combine up to four different fields connected by the “AND” or “OR” operators. Search parameters are retained when users click on the “Revise this advanced search” on the search results page. The advanced search interface also allows complex Boolean expressions in the query box, for example:

(dog OR cat) AND (food OR drink)

If a user enters unbalanced parenthesis, quotes or operators, for example

dog OR OR cat

the application strips out the operators and does a default Boolean AND search and provides a message informing the user.

Several bugs in advanced search were also fixed:

- The reset button now actually clears the form.
- Queries with the characters "<,>", or "&" are now handled correctly.
- The words "and" and "or" are now only interpreted as Boolean operators if the query is in lower case or mixed case and the operators (AND|OR) are all upper case.

PageTurner

The HathiTrust PageTurner now displays a version for items in the repository (at the bottom of the left column when viewing an item). The version is the date the item was last updated. Items are updated when improvements such as higher quality or more complete scans have been made.

Web Hosting Infrastructure Changes

HathiTrust's Drupal-based informational website was successfully moved from Michigan library web hosting infrastructure to the existing dedicated HathiTrust web hosting infrastructure. Work continued on the move of HathiTrust's VuFind-based bibliographic catalog, which is expected to be completed in early May.

Outages

No outages were reported in April 2012.

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.

You can follow HathiTrust on Twitter or

Subscribe to email updates (via Google Groups)