HathiTrust Digital Library

Update On June Activities

July 13, 2012

Top News

Update to Bibliographic Ingest Specifications and Checklist

HathiTrust has updated its bibliographic metadata specifications and minimum bibliographic metadata requirements in preparation for moving to Zephir (under development by California Digital Library) as the bibliographic metadata management system for HathiTrust. The requirements are in effect immediately for institutions that have not previously deposited content in HathiTrust. Institutions that have already deposited content are requested to meet the minimum requirements, but it is not required (we will continue to accept bibliographic metadata as it has been submitted). The primary difference from previous requirements is that records from new depositors must at a minimum include a MARC Leader, 008 field, title field, and OCLC number in order to be loaded. Further details are provided at http://www.hathitrust.org/bib_specifications. The HathiTrust Ingest Checklist page has been revised in conjunction with these changes.

Tools to Support Ingest of Locally-digitized Materials

The University of Michigan made the first iteration of tools available to aid institutions in transforming, validating, and packaging digital content for deposit in HathiTrust. The tools can be downloaded at http://www.hathitrust.org/ingest_tools. Notifications of updated versions of the tools will be sent to a Google Groups email list, and we recommend anyone who will be using the tools to subscribe.

New HathiTrust Project Librarian

by Jeremy York

I am pleased to announce the appointment of Angelina Zaytsev to the position of HathiTrust Project Librarian. Angelina has worked for HathiTrust part-time for the last year, assisting in the coordination of numerous HathiTrust activities including ingest of content from partner institutions, processing permissions to open access to materials, general user support, and multiple duties "as assigned". Angelina will continue these duties while taking on a greater role in managing and coordinating projects for HathiTrust.

Registration Open for HathiTrust Research Center UnCamp

The HathiTrust Research Center opened registration for the HTRC UnCamp, to be held in Bloomington, Indiana on September 10-11, 2012. More information can be found at http://www.hathitrust.org/htrc_uncamp2012.

New Copyright Status (IC-US)

Michigan staff completed the majority of development necessary to support a new rights status in HathiTrust Web applications. The status will apply to works that were restored to being in copyright in the United States by the General Agreement on Tariffs and Trade (GATT), but are now in the public domain in the rest of the world. An increasing number of these volumes are being identified as part

Papers & Presentations

Jon Rothman: HathiTrust Supplemental Metadata, June 23, 2012.

Partner-specific:

Jackie Bronicki: Validating the Quality of Digitized Books for Long-Term Preservation: Preliminary Findings on Digitization Error, June 23, 2012 (slides pending)

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

of CRMS-World, the IMLS-funded continuation of the CRMS project.

Ingest

Internet Archive

HathiTrust continued working with Boston College and began working with Penn State and the University of Illinois on ingest of volumes digitized by the Internet Archive.

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 Experience Advisory Group

The User Experience Advisory Group continued discussions about a new home page design and provided feedback on mockups created by the University of Michigan.

User Support Working Group

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

User Support Issues	June	May
Content	**237**	**168**
Quality	228	159
Non-partner Digital Deposit	0	0
Collections	6	3
Cataloging	**31**	**51**
Access and Use	**123**	**129**
Copyright	69	64
Permissions	20	12
Takedown	2	2
Print on Demand	0	1
Inter-library loan	0	0
Full-PDF or e-copy requests	20	22
Datasets	2	4
Data Availability and APIs	0	4
Reuse of content	2	2
Web applications	**19**	**12**
Functionality problems	2	3
Problems with login specifically	3	0
General questions about login	0	0
Partners setting up login	1	3
Usability issues	0	0
Feature requests	6	1
Partner Ingest	**3**	**2**
General	**72**	**81**
Partnership	14	6
Infrastructure	0	1
Miscellaneous	59	74
Total	**485**	**443**

*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 (CDL) staff completed a portion of development needed to sync Zephir with rights information in the HathiTrust rights database. Staff also reloaded records for HathiTrust items into Zephir as part of an iterative process to be sure rights and other necessary administrative metadata are being properly loaded. CDL staff completed documentation of Zephir metadata ingest and workflow guidelines, and will be working with HathiTrust project staff to add the information to the HathiTrust website as the launch of Zephir gets nearer.

mPach

University of Michigan staff rewrote the list of modules to be included in mPach, a package of tools Michigan is developing for publishing open access journal content in HathiTrust. Staff divided modules into three categories: Editorial Workflow and

Peer Review, Content Preparation, and HathiTrust. Work continues to adapt PageTurner to handle full-text XML content and to develop wireframes for the Dashboard module. Staff began developing code to validate the Submission Information Package.

IMLS Quality Grant

Project staff completed review of the 4th and final 1,000-volume sample, consisting of non-Roman language materials. This concluded the data collection phase of the project. The project team finalized plans for a specialized study of errors in digitized illustrations, and began to assemble and review select illustrations from each of the Library of Congress classifications. The study is designed to be an in-depth investigation into errors in digitized illustrations; it is not meant to describe or characterize the extent of errors in HathiTrust as a whole.

Jackie Bronicki presented current findings of the project at the ALA Annual Meeting in June. The new interface for the project website, which has been undergoing a redesign in the last couple of months, was released at the same time.

The project team also made progress on a framework for certifying the quality of volumes in HathiTrust.

Development Updates

Full-text Search

Staff at Michigan completed the first phase of work to improve indexing and searching of CJK (Chinese, Japanese, and Korean) languages. The first phase involved re-indexing all 10.4 million volumes in the repository using the new CJKBigramFilter available in Solr 3.6, and a custom unigram filter. The new index was put into production in mid-June. The improvements in search precision for CJK queries turned out to be smaller than anticipated. Investigation revealed the cause to be a bug in Solr's edismax query processor, and a bug report was filed in the Solr JIRA bug-tracking system. Michigan staff are investigating both temporary workarounds and a long-term fix to the bug.

Michigan staff indexed the INEX Book Track corpus and conducted a series of relevance ranking experiments. From the experiments, 6 "runs" were chosen and submitted to the INEX Book Track "Prove It" task. Three of the runs were designed to simulate users searching the HathiTrust full text index and three were baseline runs to measure the impact of the default HathiTrust search configuration on queries of different types and lengths. The results from the INEX Book Track will be used to tune relevance ranking in HathiTrust's repository-wide, and single-volume, full-text search.

California Digital Library refined the algorithm used to score spelling suggestions based on queries extracted from HathiTrust log files, and improved the way suggestions are made when stop words and words that are inappropriately combined

You can follow HathiTrust on Twitter or Facebook

Subscribe to email updates (via Google Groups) are present in the query. The next step will be to experiment with making suggestions in different languages.

Michigan removed a long-standing bottleneck in the full-text indexing process, effectively doubling throughput. Under ideal conditions staff believe it should be possible now to index approximately 100,000 documents per hour.

Outages

No outages were reported in June.

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.

Total Volumes Added	June	Overall
Columbia University	0	64,184
Cornell University	85	399,956
Duke University	0	4,523
Harvard University	34,607	234,346
Indiana University	0	187,664
Library of Congress	0	89,416
North Carolina State University	0	3,196
University of North Carolina - Chapel Hill	0	8,088
Northwestern University	4	7,207
New York Public Library	1	259,560
Penn State University	0	43,322
Princeton University	0	250,849
Purdue University	2,906	27,687
University of California	3,446	3,340,228
University of Chicago	1,210	22,031
University of Illinois	0	96,151
Universidad Complutense	1	111,828
University of Michigan	7,100	4,546,468
University of Minnesota	830	100,300
University of Wisconsin	3	539,211
University of Virginia	0	48,922
Utah State University	0	90
Yale University	0	23,678
Total	50,193	10,408,905

Public Domain (~29% of total)

Total*	72,332	3,105,587

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