HathiTrust Digital Library

Update On February Activities

March 9, 2012

Top News

Board of Governors Elections

The elections for the HathiTrust Board of Governors are taking place the first two weeks of March. There are twelve candidates in the running, and six will be elected. Each institution or consortia is able to vote for up to six individuals, and each vote will be assigned the full voting weight of the partner. The 12 candidates are as follows:

- Richard W. Clement, Dean of Libraries, Utah State University
- Susan Gibbons, University Librarian, Yale University
- Elizabeth Kirk, Associate Librarian for Information Resources, Dartmouth College
- Carol A. Mandel, Dean of Division of Libraries, New York University
- Sarah C. Michalak, Associate Provost & University Librarian, University of North Carolina at Chapel Hill
- Judith Russell, Dean of University Libraries, University of Florida
- Helen Shenton, Executive Director for the Harvard Library, Harvard University
- Martha R. Sites, Deputy University Librarian, University of Virginia
- Patricia A. Steele, Dean of Libraries, University of Maryland, College Park
- Diane Parr Walker, Edward H. Arnold University Librarian, University of Notre Dame
- Betsy A. Wilson, Dean of University Libraries, University of Washington
- Robert Wolven, Associate University Librarian for Bibliographic Services & Collection Development, Columbia University

Changes to tab-delimited metadata files

Staff at Michigan added 5 new fields to HathiTrust's tab-delimited inventory files: publication date, publication location, language, bibliographic format, and an indication of whether or not a volume has been identified as a U.S. federal government document. A description of the new fields, which are included in the inventory files as of March 1, is available at http://www.hathitrust.org/hathifiles_description.

Data API Modifications

Effective May 1, support for legacy Data API URLs in the following form will be removed:

http://services.hathitrust.org/api/htd/pathinfo-arguments

After May 1, URLs should be submitted according to the current Data API schema without the "api" path element:

http://services.hathitrust.org/htd/pathinfo-arguments

Michigan staff resumed work on Data API security enhancements, which was postponed in August as staff prepared systems to support access capabilities for

March Forecast

Governing Board Elections

Continue work on advanced search features for full-text search

Continue work on Data API security enhancements.

Papers & Presentations

Tom Burton-West "HathiTrust Large Scale Search: Scalability meets Usability". Code4Lib, February 7, 2012.

Jeremy York "HathiTrust: Issues and Challenges in Preserving the Published Record". Amigos Online, February 8, 2012.

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

Update On February Activities

users who have print disabilities and access to orphan works. The specification for the new Data API security is available at http://bit.ly/jozHQK.

Print on Demand Reports

HathiTrust is now posting reports of public domain and open access volumes in HathiTrust that are available for print on demand. The reports can be found at http://www.hathitrust.org/pod_reports, and will be released on the first of every month beginning in April.

Ingest

Local Digitization and Internet Archive

HathiTrust began working with the University of Utah and continued conversations with Northwestern University on ingest of locally-digitized volumes. Staff at Michigan completed ingest of a second set of open access volumes from the Utah State University Press.

Google

Michigan Staff received bibliographic metadata for approximately 180,000 volumes from the University of Illinois. These volumes were part of growth projections that are made yearly by partners, on which annual storage purchases are made. Ingest of the Illinois volumes will begin after the 2012 additional storage is in place, likely in April.

Total Volumes Added	February	Overall
Columbia University	1	64,177
Cornell University	6,855	391,460
Duke University	0	4,522
Harvard University	233	53,674
Indiana University	205	187,155
Library of Congress	0	89,411
North Carolina State University	0	3,196
University of North Carolina - Chapel Hill	0	8,087
Northwestern University	210	6,266
New York Public Library	40	259,506
Penn State University	316	43,262
Princeton University	939	250,618
Purdue University	23,053	23,940
University of California	36,848	3,329,011
University of Chicago	1,198	12,897
University of Illinois	0	14,503
Universidad Complutense	57	108,740
University of Michigan	13,190	4,525,854
University of Minnesota	1,787	92,368
University of Wisconsin	4,995	533,573
University of Virginia	1,525	48,921
Utah State University	44	90
Yale University	4	23,678
Total*	91,500	10,074,909

Public Domain (~28% of total)

Total**	59,574	2,791,223

* Does not include archival and image materials in the Minnesota Digital Library project

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

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 UX Advisory group provided feedback on several interface- and usability-related projects in February: the addition of a volume version (date of last ingest) in the PageTurner interface, a potential change of the default view in PageTurner, and the best way to encourage the creation of high-quality public collections. Work on all three of these projects will continue in March.

User Support Working Group

The User Support Working Group decided to postpone submission of its report on recommendations to the Executive Committee until later in the year. This will give more time for changes that have or might be implemented as a result of the group's recent evaluation process to be assessed and incorporated into more formal recommendations.

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

User Support Issues	February	January
Content	**106**	**144**
Quality	97	117
Non-partner Digital Deposit	3	0
Collections	2	10
Cataloging	**24**	**38**
Access and Use	**131**	**79**
Copyright	73	33
Permissions	20	20
Takedown	1	0
Print on Demand	1	0
Inter-library loan	0	0
Full-PDF or e-copy requests	17	15
Datasets	1	0
Data Availability and APIs	0	0
Reuse of content	0	1
Web applications	**22**	**24**
Functionality problems	7	7
Problems with login specifically	0	1
General questions about login	5	3
Partners setting up login	3	1
Usability issues	1	5
Feature requests	0	4
Partner Ingest	**5**	**4**
General	**152**	**127**
Partnership	11	7
Infrastructure	2	1
Miscellaneous	139	119

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

Strategic

Collections Committee

The Executive Committee and SAB have approved the Collection Committee's recommendations for the treatment of duplicates in HathiTrust; the final report will be posted online shortly. The report discusses various categories of duplicates that exist in the repository and attempts to assess their scope and cost, while also noting some of the difficulties in the precise identification of duplicates. The report recommends that HathiTrust retain all duplicate copies ingested into the repository for the time being, with periodic reassessment. Some categories of duplicates are recommended for permanent retention (e.g. early published books). The SAB has requested that the Committee make further recommendations about the criteria that should be applied in future assessments and identify the future costs and risks of retaining duplicates in the corpus.

Projects

Bibliographic Data Management System

California Digital Library has nearly completed testing of HathiTrust records received from the University of Michigan in Zephir, the new management system, and is making significant progress on reconciling records ingested in both systems. Staff from Michigan and CDL finalized the minimum record submission standard to be used prospectively for records submitted by partners to HathiTrust. The standard will be integrated into the HathiTrust ingest checklist. CDL and Michigan also worked to address issues related to integration planning and reconciliation of records in Michigan's system and in Zephir. CDL performed a dry run load test on ingest of records into Zephir.

HathiTrust Publishing (HTPub)

Staff at Michigan completed the first iteration of a tool that is able to create valid JATS XML from simple DOCX files, and continued development on PageTurner to render JATS XML. Staff clarified the goals of the project to include implementation of a publishing system (allowing management of an editorial workflow) in addition to mechanisms for ingest, display, and discoverability of born-digital journal materials in the HathiTrust repository. More information is available at http://www.hathitrust.org/htpub.

HathiTrust Research Center (HTRC)

The HathiTrust Research Center released a beta version 0.1 of the HTRC Data API. The API is a RESTful API through which the HTRC Solr index and volume store are accessed. It cannot be used to download volumes, but can be used to move data to a location where computation takes place. It can also be used to search the Solr index for a set of volume IDs and pass the volume IDs to a service for access and computation. Access to the API will require OpenID authentication and appropriate authorization. The Data API is installed on two sandbox machines, one at UIUC and another at IU, for internal testing. Both sandbox installations work against a small subset of non-Google scanned volumes.

The HTRC technical team prototyped Blacklight (http://projectblacklight.org), an open source library catalog search and retrieval system, for deployment in the HTRC. Blacklight is designed to support data that is both full text and bibliographic, it is built on Solr, the same technology used to index HTRC data, and Blacklight supports faceted searches, a known need of researchers. The test implementation of Blacklight was deployed on a shared server at UIUC. In the next quarter, the HTRC expects to use the customization options to configure the look and feel of the interface and perhaps extend the functionality to show snippets of the text to help researchers refine their results. Any new functionality that is developed will be shared back with the larger Blacklight community. Blacklight is expected to be a significant component of the public face of the HTRC.

Members of the HTRC performed a study recently on quantifying OCR errors in the HathiTrust corpus. Scholars are interested in doing quality text analysis, but results can be confounded by OCR errors. Information on which books (or pages) in the collection have significant rates of OCR errors could help. The HTRC explored a couple of approaches to OCR error detection and have results for one approach that uses machine-generated and expert-evaluated rules. Starting with a large dictionary of correctly spelled words, HTRC members identified outlier words that were in the HathiTrust corpus but not in the dictionary. As a check on identified words, the rules by which outliers were detected were verified by a human expert. Using this approach, HTRC formulated 48,308 rules that identified outlier words and provided corrections. HTRC members applied the rules to 256,000 non-Google digitized volumes from HathiTrust, which took 4 hours using the National Center for Supercomputing Applications' Ember supercomputer. The results showed that the probability of a word having an OCR error (detected by the rule set) was 0.20%. The average number of errors per page was 0.57. The average number of errors per volume was 156. The probability that a page had one or more errors on it was 11%. The probability that any volume had one or more errors was 84.9%. Overall, 217,754 of the 256,416 volumes had one or more OCR errors and 7,745,034 of the 69,297,000 pages had one or more errors.

You can follow HathiTrust on Twitter at http://www.twitter.com/hathitrust

IMLS Quality Grant

Project staff completed page-level review of a third production sample, consisting of 1,000 volumes digitized by the Internet Archive. More than 85,000 pages were reviewed in all. Approximately 9,400 of these (about 10%) were coded by two reviewers for quality assurance purposes. The focus of project work shifted then to finalizing training materials and data collection systems and procedures for whole-volume error review (review for errors that apply to an entire volume, such as missing, out-of-order, or duplicate pages). Project staff reviewed approximately 300 test volumes in a new whole-volume review interface to surface issues in using the interface and applying the new error model, and to develop an initial training manual. Whole-volume review began in mid-February on the same volumes reviewed in the first page-level production sample (1,000 English language, public domain, Google-digitized volumes).

Although the primary focus of work shifted to whole-volume review, physical review of the volumes sampled in the first production run continued in February. 870 of the 1,000 volumes in the sample were obtained and reviewed by volunteer graduate students at the University of Michigan. Students also began physical review of 600 Michigan volumes included in the second page-level sample (1,000 English language, Google-digitized volumes published post-1923). More than 400 of these volumes were reviewed by month's end.

Development Updates

PageTurner

Michigan staff made a number of adjustments to the PageTurner application. These included fixing a bug in the RFDa emitted in the PageTurner bibliographic metadata that had prevented license information from being included appropriately; enhancing the access control mechanism for items that are public domain in the United States to better detect whether a user is on U.S. soil when access is proxied; updating the back-end process by which user feedback is submitted from HathiTrust applications (including PageTurner, the HathiTrust bibliographic catalog, Full-text Search, Collection Builder) to the central HathiTrust ticketing system; and implementing a process to detect cases where multiple tickets are submitted on identical HathiTrust items or records.

Full-text Search

Michigan staff began work on the next iteration of advanced full-text search, which will allow users to build queries with greater Boolean complexity and enhance the ability to revise advanced searches. Staff made progress as well on plans to improve search results relevance ranking. This work is planned to begin after the next release of advanced full-text search.

California Digital Library staff completed dictionary-building work for the spelling suggester feature. The code can now build a language-sensitive dictionary of unigrams and bigrams from any Lucene index, automatically choosing a frequency cut-off to constrain the size of the dictionary. Focus will now shift to implementing fast-lookup and suggestion ranking.

Outages

Page viewing of volumes classified as “Public Domain in the United States” was unavailable on Tue 2-7 from approximately 5:30-9:45pm EST due to a software problem.

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.