HathiTrust Digital Library

Update On November Activities

December 9, 2011

Late Breaking News

Boston College Joins HathiTrust

HathiTrust is pleased to welcome Boston College as its newest member. The full announcement is avalable on the Boston College website.

Research Center on HathiTrust.org

A new "Our Research Center" portion of HathiTrust.org was launched in early December, containing information about the governance, timeline and deliverables, architecture, and access and use policies for the HathiTrust Research Center (HTRC), which is jointly led by Indiana University and the University of Illinois. "Our Research Center" also includes information about research partnerships and a demonstration tool that allows users to create tag clouds and perform limited analysis on a small number of works. The HTRC welcomed two new members from the University of Illinois library to the HTRC technical team in November: Kirk Hess and Harriett Green. Kirk and Harriett bring experience in user interfaces and services, areas that complement the technical strengths of Indiana staff currently working on the HTRC.

Is that the library in your pocket?

A new Perspectives on HathiTrust blog post authored by Suzanne Chapman, chair of the User Experience Advisory Group, was released in early December, highlighting HathiTrust's new mobile interface.

Top News

Expansion of "Buy a Copy"

The "Buy a copy" option has now been expanded to include over 30,000 public domain volumes from the University of California. UC will incrementally add new volumes to the service. UC has partnered with Hewlett-Packard to create the reprints and make them available for purchase via Amazon.com.

User Support: New member and call for nominations

The User Support Working Group is very pleased to welcome a new member, Kathryn Stine from the California Digital Library. The working group is seeking nominations from partner institutions for up to 4 additional positions. Nominations should be sent to Jeremy York (jjyork@umich.edu) and include the name, title, and a short description of current job duties. Additional information that might be relevant to participation in the group may be included as well. User Support members are on call at least one day per week and follow up on inquiries throughout the week, requiring between 2-4 hours of work. Staff that participate on the group will

- Gain knowledge about HathiTrust's user base, typical problems and ques-

December Forecast

- Complete orphan works project pilot phase
- Enable updated throttling policies
- Complete installation of two replacement servers

Papers & Presentations

Jeremy York "HathiTrust Infrastructure and Information Organization". November 7, 2011.

Jeremy York "HathiTrust Current Work, Challenges, and Opportunities for Public Libraries". November 16, 2011.

Partner-specific

Magán, Palafox, Tardón, Sanz, "Mass Digitization at the Complutense University Library: Access to and Preservation of its Cultural Heritage". Liber Quarterly, Vol. 21 (2011), No.1.

Update On November Activities

tions that are raised and how they are resolved.
- Become aware of new ways HathiTrust is being used, and features and functionality that users desire.
- Gain knowledge of HathiTrust organizational and technical infrastructure, and policies and procedures relating to copyright, access, collection development, deposit of materials, and preservation.

The charge for the working group is available at http://www.hathitrust.org/wg_user-support_charge.

Volunteer Lucene Developer

HathiTrust is seeking a volunteer Lucene developer (from partner institutions or not) to work directly through the Lucene contribution process to improve indexing capabilities for Chinese-, Japanese-, and Korean-language (CJK) materials; more specifically, to add overlapping bigram functionality for CJK languages to the Lucene ICUTokenizer (view the Lucene JIRA ticket for this issue). A new HathiTrust large-scale search blog post on word segmentation for CJK languages provides additional context. Please contact Tom Burton-West for more information.

Changes to Tab-delimited files

On February 1, HathiTrust will be adding additional columns to the tab-delimited inventory files ("hathifiles"). A final description of the changes will be posted in the update on December activities. Proposed additions include the publication date and publication location of volumes, as well as an indication of whether volumes have been identified as U.S. federal government documents.

Updated Permissions Agreement

University of Michigan staff have updated the permissions agreement by which rights holders can open access to their works in HathiTrust. The agreement, which is now also available as a fillable PDF, is available at http://www.hathitrust.org/permissions_agreement, with instructions on completion and submission.

Total Volumes Added

	November	Overall
Columbia University	123	64,172
Cornell University	5,833	374,089
Duke University	15	4,501
Harvard University	163	53,006
Indiana University	393	186,588
Library of Congress	0	73,642
North Carolina State University	0	3,194
University of North Carolina - Chapel Hill	0	8,087
Northwestern University	57	5,412
New York Public Library	212	259,377
Penn State University	281	41,096
Princeton University	413	249,329
Purdue University	886	887
University of California	27,759	3,172,748
University of Chicago	822	8,875
University of Illinois	0	14,503
Universidad Complutense	296	108,640
University of Michigan	34,744	4,481,254
University of Minnesota	728	89,323
University of Wisconsin	6,190	511,432
University of Virginia	54	47,384
Utah State University	0	46
Yale University	0	23,674
Total	78,971	9,781,261

Public Domain (~27% of total)

Total*	6,032	2,662,192

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

Update On November Activities

Ingest

Volume Projections for 2012

HathiTrust sent a call to partners in November for projections of volumes to be deposited in 2012. The projections will be used to estimate storage needs and fees for partners in the coming year. A variety of locally-digitized collections were identified for deposit, in addition to volumes digitized through Internet Archive and Google. More information on these and continuing work on ingest will be included in coming months.

Local Digitization

HathiTrust has ingested nearly all of approximately 200 rare manuscripts and incunabula from the Universidad Complutense de Madrid. Issues with some of the submitted volumes that prevented ingest will be investigated further by Michigan staff.

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 Working Group

The Communications Working Group continued to make progress on a public services-oriented communications package, highlighting ways HathiTrust can be used to address a variety of research and reference inquiries.

User Experience Advisory Group

The User Experience Advisory Group finalized the user personas it began to develop over the summer. The personas and an accompanying overview of the project are available at http://www.hathitrust.org/personas. The purpose of the personas is to help HathiTrust staff and partners (developers, policy makers, user experience designers and researchers, reference and instruction librarians, etc.) envision different types of HathiTrust users in a more concrete way in order to inform our work. The group welcomes any questions or comments about the personas to be sent to Suzanne Chapman (suzchap@umich.edu), chair of the UX Advisory Group.

User Support Working Group

The table on the following page contains a summary of the issues received by the User Support Working Group in November.

HathiTrust Digital Library

Update On November Activities

Strategic

Collections Committee

The Collections Committee made several revisions to its draft recommendations for handling duplicates in HathiTrust and has submitted these to the Strategic Advisory Board for approval. Once the revisions have been approved by SAB and endorsed by the Executive Committee, the full document will be posted on the HathiTrust website. The committee is currently discussing mechanisms for responding to user-submitted requests and offers and is formulating plans to address other items on its work agenda. Suggestions for additional work items are always welcome and can be sent to the committee chair, Ivy Anderson (ivy.anderson@ucop.edu).

Projects

Bibliographic Data Management System

The California Digital Library team worked to develop an infrastructure to compare bibliographic records in Zephir, its core metadata management system, with records in HathiTrust. Some of the challenges are determining when a record has been updated in HathiTrust, and managing multiple (non-HathiTrust) identifiers for volumes. The Zephir team loaded and tested records, and refined the timeline for migrating records into Zephir as they work with staff at the University of Michigan. Further information about the project can be found at http://www.hathitrust.org/htmms.

HathiTrust Publishing (HTPub)

Staff in MPublishing began work in November on a tool to convert DOCX files to JATS XML and worked with broader stakeholders at the University of Michigan Library to specify additional design requirements and agree on a set of design principles for HTPub (available on the HTPub project page). MPublishing staff also reviewed notes from a session at THATCamp Publishing 2011 dedicated to shared infrastructure for publishing, to consider how such an infrastructure might affect the architecture of HTPub tools, and services that might be offered in the future using those tools.

IMLS Quality Grant

Physical review of volumes in the first 1,000-volume sample continued in No-

User Support Issues

	October	November
Content	**154**	**107**
Quality	142	102
Non-partner Digital Deposit	0	0
Collections	1	5
Cataloging	**44**	**43**
Access and Use	**136**	**103**
Copyright	75	55
Permissions	4	10
Takedown	4	1
Print on Demand	2	2
Inter-library loan	0	0
Full-PDF or e-copy requests	23	15
Datasets	1	1
Data Availability and APIs	2	2
Reuse of content	2	1
Web applications	**29**	**24**
Functionality problems	6	5
Problems with login specifically	3	1
General questions about login	4	2
Partners setting up login	1	3
Usability issues	2	2
Feature requests	5	2
Partner Ingest	**1**	**3**
General	**59**	**47**
Partnership	8	6
Infrastructure	0	0
Miscellaneous	51	41

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

Update On November Activities

vember, with volumes requested through inter-library loan continuing to arrive at Michigan. Plans are being made for staff at partner libraries to conduct physical review of volumes in cases where the volumes are not available for inter-library loan. Preliminary information on the quality review performed on the first sample of digital volumes will be posted to the grant project website in December.

Quality review on the second sample of 1,000 volumes from HathiTrust was completed in mid-November. Measures to evaluate inter-coder consistency required re-review of some volumes in the sample, as well as individual pages within specific volumes. This review began in late November and should be complete in the first week of December. As review of the second sample of volumes was completed, project staff prepared to begin review of a third sample of 1,000 volumes, which will include pre-1923 English-language monographs digitized by the Internet Archive.

Project staff continued to define requirements for a new quality review interface, targeted specifically for review of volume-level errors such as missing, duplicate, and out-of-order pages. The project developer began coding basic elements of the system. Combining this new interface and procedures with those in the first interface, which was designed to review page-level errors, will lead to a system for comprehensive review that will enable certification of volumes at different quality levels. The project team is in the process of drafting specifications for certifying volumes. The final model will be based on the findings from statistical sampling and manual review at the page and volume levels.

Orphan Works

Work continued on the Orphan Works Project pilot phase, which will continue through the end of December. Reviewers from the University of Michigan and the University of California, Los Angeles have now researched the same set of approximately 50 works. Staff from both institutions are looking at the results and reviewing the process for accuracy. The pilot phase of the OWP is intended to serve as a test for an orphan works identification process, through which we will document examples and further define parameters for research.

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

Development Updates

Full-text Search

Staff at Michigan began a re-indexing process in November for all 9.8 million volumes in HathiTrust. The purpose was to correct an error in the "Original Location" metadata facet, and to provide additional metadata for advanced search, relevance ranking, and to determine the viewability status of volumes (see below). The re-indexing was 98% complete at the end of November and is anticipated to go into production in early December. This re-indexing, and the discovery of a bug in the way Solr processes Boolean queries, slowed development of the advanced search feature that was planned for release in November. A workaround

Update On November Activities

for the bug will be implemented until the bug is fixed. The advanced search feature is now planned for release in January.

As the indexing enhancements were put in place, Michigan staff completed the coding necessary for full-text search results to reflect whether or not a user is able to view items in situations that depend on institutional print holdings and other factors. This will apply to search results that include orphan works (when available), volumes that may be available under Section 108 of U.S. copyright law, and volumes that are accessible to users at partner institutions who have print disabilities. In order to see the availability of these volumes, and access them, users from partner institutions will need to be logged in using their institutional account.

Michigan developers continue to work with staff at the California Digital Library on the development of a spelling suggestion feature. CDL is testing various algorithms on sample HathiTrust data including the Solr/Lucene Levenshtein Automaton and Martin Reynaert's anagram hashing algorithm. The work is focusing both on the speed and scalability of the algorithms and on the accuracy of the suggestions. Experimental code to extract useful bigrams from existing HathiTrust indexes is in the works, which will obviate the need to maintain multiple indexes to support spelling correction, as is currently the case.

PageTurner

HathiTrust has implemented new policies regarding access to in-copyright works, where lawful access is permitted. Access for authorized users at partner institutions who have print disabilities is now only possible from IP addresses within the United States. Access is limited to one user per physical volume held by the user's institution. Access to in-copyright works is also now recorded in HathiTrust system logs, in accordance with HathiTrust's privacy policy: http://www.hathitrust.org/privacy.

In connection with the HTPub effort, Michigan staff continued work to adapt the HathiTrust PageTurner to display XML content based on initial specifications.

Throttling

Michigan staff tested and refined application-specific policies for throttling (e.g., in the PageTurner, Full-text search, and Collection Builder applications), and expect to enable the new policies in December.

New Web Servers

Michigan staff purchased and began installing two new replacement web servers for HathiTrust in November. These are the last of 8 eight servers targeted for replacement this year (the other six were replaced in July).

Storage Hardware Upgrade

The new storage brought online in June of this year was discovered by Isilon Systems, the storage provider, to have a subtle hardware issue requiring all drives

and some internal components in eight nodes to be removed and re-installed in a new chassis. The upgrade was preventative in nature; the minor symptom that caused the hardware issue had not been observed in HathiTrust. The maintenance was covered under the existing support agreement, and carried out without any interruption to service by Isilon's field service technicians under close supervision by Michigan staff.

Security Risk Assessment and Vulnerability Test

As part of a regular program for continuous improvement in IT security, Michigan Library staff have been working with analysts in University of Michigan central IT to conduct a thorough risk assessment and vulnerability penetration test of the HathiTrust infrastructure. The scope of the risk assessment, which follows a framework developed at the University, consists primarily of servers and storage hardware, but also includes coverage of aspects such as facilities, management practice and policy, and workflows involving sensitive data. The vulnerability test focuses on network security, and is a hands-on exercise conducted by a trained security expert who attempts to discover flaws in network security and evaluate their potential for exploit. Final reports on both analyses are due in December.

Outages

No outages were reported in November 2011.

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.