HathiTrust Digital Library

SPECIAL EDITION - 2011 YEAR IN REVIEW

January 17, 2012

The close of 2011 marked 4 years since the first formal commitments were made to building HathiTrust, a broad collaborative of academic and research institutions that are working together to ensure the long-term preservation and accessibility of the cultural record.

2011 saw the solidification of the HathiTrust repository's position in the library community, as it received Trustworthy certification from the Center for Research Libraries. It also saw the solidification of HathiTrust services, as a new mobile interface was released, significant enhancements were made to the Full-text search, PageTurner, and Collection Builder applications, and a database of print holdings was incorporated into access systems, providing a mechanism to provide lawful access to in-copyright materials that are held by member institutions.

The HathiTrust partnership achieved a new level of cohesion and stability in 2011 as well, as the member institutions came together in a Constitutional Convention to make collective decisions about the structure and priorities of the initiative going forward. Agreements with a variety of entities (organizations, academic presses and vendors) to expand access to materials in HathiTrust and enhance their discovery further magnified the impact of the partnership's work in the broader library community.

2011 offered partners the opportunity to reflect on the accomplishments of HathiTrust in its first years, and make collective plans to address the challenges libraries face in stewarding and provisioning the cultural record in years to come. We move into 2012 with optimism, based on what we have been able to achieve, in our ability to collaborate deeply and effectively to address these challenges, and maintain and even enhance the role that libraries play in the new, shared, digital future.

A summary of HathiTrust activities in 2011 is given below:

Highlighted Achievements and Activities

Details on each item can be found in the monthly updates from 2011, available at http://www.hathitrust.org/updates.

New Partners

HathiTrust grew from 52 to 66 partners in 2011. The new institutions that formally announced partnership include:

- Boston College
- Boston University
- Lafayette College
- University of Arizona
- University of Connecticut
- University of Florida
- University of Notre Dame
- University of Miami
- University of Missouri

New content

HathiTrust partners contributed 2,129,874 volumes to the repository in 2011, for a total of 9,966,572. 753,403 of these (2,712,626, or 27% overall) are either in the public domain or volumes that rights holders have given HathiTrust permission to make publicly available. HathiTrust exceeded 10 million volumes in early January 2012 (see the blog post and timeline of repository development).

Having completed a framework for ingesting volumes from varied sources at the end of 2010, in 2011 HathiTrust began to scale up ingest of locally-digitized content from partner institutions. Large-scale deposits continued as well. New institutions contributing content in 2011 included:

Large-scale digitization

- Library of Congress
- Harvard University
- University of Virginia
- Northwestern University
- Purdue University
- North Carolina State University
- Duke University
- University of North Carolina-Chapel Hill

Local or in-house digitization

- Universidad Complutense de Madrid
- University of Minnesota

- Utah State University Press
- Yale University

Conversations regarding ingest of locally-digitized materials were initiated with

- Columbia University
- Northwestern University
- University of Florida
- University of Illinois
- University of Iowa
- University of North Carolina – Chapel Hill
- University of Utah
- University of Pittsburgh

Constitutional Convention

In October 2011, HathiTrust partners convened a Constitutional Convention to determine directions for the partnership following its first 5-year period, which will conclude at the end of 2012. HathiTrust's Strategic Advisory Board released a review of the partnership's activities and progress over its first 3 years prior to the Convention to set the stage for ballot initiatives and partner discussion. 7 ballot initiatives were considered by partners at the Convention. 5 of these were accepted:

- To establish a distributed archive of print monograph volumes from partner institutions
- To establish an approval process for HathiTrust development initiatives
- To establish a new governance structure (to be in place by mid-April, 2012 – see HathiTrust Governance for more information)
- To initiate coordinated action to expand and enhance access to U.S. federal government documents
- To establish a fee-for-service model of content deposit from non-partner entities

Information about the Constitutional Convention, including notes from the convention, ballot initiatives, attendees, and the 3-year review are available on the Constitutional Convention information page. John Wilkin's opening remarks and the presentation given by representatives of the Strategic Advisory Board are available on the Papers and Presentations page. John Wilkin's remarks were also posted on the HathiTrust blog.

Fulfillment of Functional Objectives

With the ingest of image content from Minnesota, the establishment of a HathiTrust Research Center, progress to enable HathiTrust as a platform for digital publishing, certification by the Center for Research Libraries for compliance with TRAC, and the establishment of infrastructure to offer access to in-copyright works for users who have print disabilities (see further information on these below), HathiTrust has provided a meaningful deliverable for each of the initial objectives set by the founding partners (see HathiTrust Functional Objectives).

Lawful Uses of In-copyright Materials

- In 2011, Utah State University Press and Duke University Press agreed to open back file publications in HathiTrust in exchange for perpetual archiving of the deposited volumes.

- HathiTrust added support for Creative Commons licenses in the repository, giving rights holders the ability to use these licenses to open access to materials. Support includes the inclusion of Creative Commons licensing information as RDFa in the PageTurner application. The Brooklyn Museum, Society of American Archivists, University of Texas at Austin and many others were early adopters.

- HathiTrust released the first iteration of a database of print holdings information from partner institutions. The database will act as the basis for the new pricing model to be implemented in 2013, and expanded access to in-copyright materials for members of partner institutions. See the Update on July 2011 Activities for more information. Note: expanded access has not yet been released.

- HathiTrust leveraged work in the University of Michigan's IMLS-funded Copyright Review Management System to begin to identify orphan works in the repository. Michigan, the University of Wisconsin, Cornell, Duke, Johns Hopkins, Emory University, and the University of California announced plans to provide access to orphan works identified through this process on a limited basis to faculty, students, and staff at their institutions. Information about the terms of access is available on the HathiTrust website. See also the Orphans Works Project page on the University of Michigan Library website.

- The Authors Guild and others filed a lawsuit against HathiTrust alleging copyright infringement. HathiTrust partners are convinced of the value and legality of preserving in-copyright materials and continue their work as the lawsuit proceeds. Further information about the lawsuit

is available on the HathiTrust website.

Discovery

HathiTrust signed agreements with ProQuest, OCLC and EBSCO to make the HathiTrust full-text index searchable through their discovery services.

Datasets

HathiTrust began to make datasets of public domain materials available on a large scale. See HathiTrust Datasets for more information.

Repository

- HathiTrust was certified by the Center for Research Libraries as a Trustworthy Digital Repository in March 2011.

- University of Michigan staff, with feedback from the UX Advisory group and significant contributions from California Digital Library, made a number of enhancements to the Collection Builder, PageTurner, and Full-text search applications. These included:

Collection Builder

o The application was re-architected to leverage the full-text search index, allowing users to create collections of any size.

o Interface enhancements were made to improve browsing and discovery of collections, including the abilities to search collections by title and description, and filter collections by their featured status, last time of update, number of items, and whether or not they belong to the current authenticated user.

PageTurner

o Views were added to allow users to scroll through volumes, flip pages similar to a physical book, and view thumbnail images of all pages in a volume. The views were accomplished through backend enhancements and the integration of the Internet Archive's open source BookReader in the PageTurner. Initial work on BookReader integration, including the development of thumbnail views, was completed by staff at California Digital Library.

o The interface was reorganized and streamlined to improve use. New features include more prominent display of copyright status, re-positioning of navigation features and volume information, and the ability to expand the viewing area to full-screen.

o Quick-copy links to volume pages and permanent volume URLs were added.

o A progress bar was added to improve the user experience for full-volume PDF downloads.

Full-text Search

o The full-text search index was enhanced to include bibliographic metadata, allowing for improved relevance ranking and faceted display of full-text search results. The search engine used to search inside a book was upgraded from the XPAT search engine, to Solr, improving results display for multiword searches when searching within a book. Michigan staff developed a prototype for advanced full-text search and performed a preliminary user interaction/usability walkthrough. California Digital Library completed substantial work toward the implementation of a full-text search spelling suggestion feature.

- HathiTrust began posting weekly reports on the ingest of partner volumes.

- Michigan staff completed the first cycle of storage replacement for HathiTrust, on storage purchased in 2007, as well as the first replacement of HathiTrust database and ingest servers (see HathiTrust Technology for more information on storage and replacement).

- University of Michigan staff drafted new security specifications for the Data API which will allow additional access to members of partner institutions for specific purposes.

- Staff at Michigan implemented new procedures to perform periodic, generalized auditing of the repository, including checksum validation of deposited volumes and other types of analysis, such as investigation into usage of preservation or other associated metadata.

- Michigan implemented improvements to repository throttling mechanisms to mimimize interruptions to normal use while ensuring compliance with third-party restrictions on bulk download of materials.

- Michigan staff designed and developed a mo-

bile-friendly interface to the catalog search and PageTurner (read the blog post).

- Staff at Michigan installed new infrastructure in the HathiTrust Development Environment to support performance requirements of the new print holdings database. Michigan also implemented services that allow code administrators to see differences between the last deployed versions of repository code when staging new code for development, and improve the process by which developers stage new code for testing. Partners interested in exploring the development environment should contact feedback@issues.hathitrust.org.

Governance, Working Groups, and Committees

Strategic Advisory Board

The Strategic Advisory Board welcomed two new members from the University of California in May 2011: Todd Grappone, Associate University Librarian for Digital Initiatives and Information Technology at UCLA, and Julia Kochi, Director, Digital Libraries and Collections at UC San Francisco. Todd and Julia took the place of Bernie Hurley, UC Berkeley and Bruce Miller, UC Merced.

Collections

The Collections Committee submitted a proposal to the Constitutional Convention to establish a distributed print monographs archive. This proposal was the first submitted and served as a model for subsequent proposals. It was accepted by the partners and is a foundational piece in HathiTrust's strategy to coordinate shared storage strategies among the partnership.

The Collections Committee also completed a report on treatment of duplicates in HathiTrust that is currently being reviewed by the Strategic Advisory Board, and began to consider processes for handling user requests to contribute volumes to the repository.

Tom Teper of the University of Illinois joined the Committee in 2011, stepping in for Kim Armstrong of the Committee on Institutional Cooperation.

Communications

The Communications Working Group organized a webinar, given several times in the spring, targeted towards members of the large number of new part-

ner institutions that joined HathiTrust at the end of 2010. The webinar reviewed basic elements of the partnership and discussed current activities and future work. The group also organized a webinar for non-partner institutions interested in learning more about HathiTrust. Other working group activities included:

- The development of a prioritized communications and marketing plan

- The launch of a new "Perspectives from HathiTrust" blog and establishment of a Facebook page

- Coordination of several new partner announcements, announcements of the WorldCat Local HathiTrust catalog prototype and the positive outcome of the TRAC audit, and HathiTrust's statement on the Google Settlement ruling

Three new members joined the group in 2011: Robin Bedenbaugh from Texas A&M University, Oya Rieger from Cornell University, and Stacy Kowalczyk from Indiana University, joining as a representative from the HathiTrust Research Center.

The group begins 2012 continuing work on a public services-oriented communications package, highlighting ways HathiTrust can be used to address a variety of research and reference inquiries.

Discovery Interface Working Group

OCLC released a prototype of the HathiTrust-OCLC catalog in beta in January. The catalog was the result of nearly two years of collaborative work between OCLC and HathiTrust, coordinated by the Discovery Interface Working Group. The effort included the loading of all HathiTrust records into WorldCat. After the catalog was released, the DIWG moved its focus to usability testing of the new prototype system and defining areas for subsequent improvement.

The DIWG launched a Full-text search subgroup to develop a prioritized list of features to implement in the full-text search application (see below).

The DIWG fulfilled the work in its charge and was disbanded in June 2011.

User Experience Advisory Group

The User Experience Advisory group provided feedback on the interface improvements mentioned above for the HathiTrust PageTurner, Collection Builder, and Full-text search applications.

The group also released a set of HathiTrust User Personas to help staff working on HathiTrust learn more about HathiTrust users, discover how to better meet their needs, and identify areas in which to do more in-depth research.

One new member, Darcy Duke from MIT, joined the advisory group in 2011. Darcy had been an active contributor to the UX discussion list, which was launched by the group in 2011.

User Support Working Group

In March, The Executive Committee launched a new User Support Working Group to respond to feedback submitted through HathiTrust's help email addresses and user interfaces. Staff at Michigan who had been managing the process previously helped to establish a new partner-wide ticketing system. An 8-member group began an on-call rotation to address user issues beginning in April. The group has posted statistics on inquiries in the monthly newsletter since that time. As of the end of December, 3 members, Nancy Spiegel and Todd Ito of the University of Chicago and Bob Kackley of the University of Maryland have had to leave the group. The group was pleased to welcome Kathryn Stine from the California Digital Library in November and is open to nominations from partner institutions. Please contact Jeremy York (jjyork@umich.edu) for information.

Full-text Search Subgroup

The full-text search subgroup of the Discovery Interface Working Group researched and estimated technical feasibility and implementation effort for potential new full-text search features. The group's report, which includes a prioritized list of full-text search enhancements, is available online. 3 of the top enhancements have been implemented (1a, 2, 3b) and 3 are currently under development (5a, 7, 1b).

Special Initiatives

HathiTrust Research Center (HTRC)

The HTRC initiative was formally launched by Indiana University and the University of Illinois in July 2011. The research center will offer computational access for nonprofit and educational users to public domain and, in the future, in-copyright works in HathiTrust.

The Research Center received a 3-year, $600,000 grant from the Sloan Foundation in July to investigate "non-consumptive" research on the full-text of materials in the HathiTrust corpus. Particularly in relation to in-copyright works, "non-consumptive" research is research that allows computation of results about a body of works but not significant reading or "consumption" of the works.

The HTRC technical team worked throughout 2011 to establish core cyberinfrastructure and data analysis tools for the Research Center, and develop access policies. A full demonstration of the HTRC is scheduled to be available in July 2012.

Michigan Staff developed an initial model for transferring data from HathiTrust to the HTRC (using rsync), and Indiana University staff began performing tests with sample data. Transfer of texts for public domain texts to the HTRC will occur in 2012.

More information about the HTRC is available on the HathiTrust website.

Government Documents and Copyright

Maliaca Oxnam of the University of Arizona initiated research in collaboration with HathiTrust to improve access to U.S. Federal government documents. Details on her work are available in the Update on October 2011 Activities.

IMLS Quality Grant

HathiTrust began its participation as a testbed for research in a 2-year IMLS-funded project to investigate quality in large-scale digital repositories. In its first year the project established a preliminary review interface and methodology for reviewing the quality of digital volumes, and procedures for reviewing the quality of corresponding physical volumes. The project team completed digital review of two samples of 1,000 volumes drawn from the HathiTrust repository. These were the first of several samples to be taken throughout the project to investigate quality within different date ranges and languages, and from different digitization sources. The team is performing an additional physical review of volumes in the first sample to investigate relationships between the physical condiation of volumes and errors observed in their digital surrogates. This review is largely complete. The project is working toward a distributed system that will allow partners to review and certify the quality of volumes in HathiTrust. This certification will inform and facilitate a number of partner activities, including handling of duplicate volumes in the repository,

and partners' local and collaborative strategies for managing their print collections. See the project website for more information.

University of California Print on Demand

The University of California began offering reprints of UC-digitized public domain materials via HathiTrust

Minnesota Digital Library Image Preservation Prototype Project

Nearly 60,000 images and associated metadata from the University of Minnesota and its statewide partners, the Minnesota Digital Library and Minnesota Historical Society, were ingested into HathiTrust in a project to develop a prototype workflow for depositing images and associated metadata into HathiTrust for access and preservation. More about the project can be found on the MDL Images project page.

Bibliographic Data Management

California Digital Library made significant progress toward the establishment of a new bibliographic metadata management system for HathiTrust. CDL staff completed development of the core system, Zephir, and began to load bibliographic metadata for volumes transferred from the metadata management system at the University of Michigan. CDL worked closely with Michigan to understand existing processes for transforming and managing records, and requirements for generating various outputs, including metadata for the HathiTrust catalog, OAI feed, and tab-delimited inventory files. CDL and Michigan have established ongoing processes for syncing of bibliographic data from Michigan to CDL, and will begin to prepare for integration testing and deployment in 2012.

HTPub

HTPub, an effort by the University of Michigan Library's MPublishing Division to develop mechanisms for ingest, display, and discovery of born-digital materials in HathiTrust, moved from initial planning to design and development stages in 2011. Team members at Michigan hired two new staff members to support the initiative, defined goals, requirements, and design principles for the project, began to design system architecture and determine archival package specifications for deposited content, and began initial development of content transformation tools and mechanisms for displaying content. Information about the project is available on the HTPub project page.