HathiTrust Digital Library

Update On September Activities

October 8, 2010

Top News

September Meeting in Chicago – Staff from a number of HathiTrust institutions gathered in Chicago on September 23 and 24 for meetings of HathiTrust's governing committees, its operational and planning working groups, and teams working on specific projects including ingest of locally digitized partner content, full text search, user interface collaboration, and others. Activities in the meetings are reported throughout the newsletter, and will be posted in the Executive Committee and Strategic Advisory Board meeting minutes.

Local Digitization Ingest – The first draft of a policy and specifications framework for receiving content from a broad array of digitization sources and workflows into HathiTrust was completed in September and shared with several partner institutions. The framework is now available publicly online in two parts: the HathiTrust Guidelines for Digital Object Deposit and the HathiTrust Deposit Form, which includes detailed specifications for submitted content. We would like to formally request comments and feedback on the framework from partner institutions and interested parties. To be included in our review and revisions, please send comments to hathitrust-info@umich.edu by October 31, 2010.

Copyright Review – For the last two years, staff at the University of Michigan have been conducting review of volumes in HathiTrust that were published in the United States from 1923 to 1963, releasing materials into the public domain that do not comply with U.S. copyright formalities. Over the summer, this work was expanded to additional HathiTrust partner institutions and staff at Indiana University, the University of Wisconsin, and the University of Minnesota were trained to work as reviewers. As of September 1st, 18 staff members from the four institutions are contributing to the project. The increase in staff has resulted in a larger number of volumes being opened up in HathiTrust on a monthly basis, from 470 volumes in June to over 2500 volumes in September. Approximately 85,000 of 188,000 current candidate volumes in HathiTrust have been reviewed since the project began. Close to 50,000 of these, or about 55% have been determined to be in the public domain and opened in HathiTrust.

Shibboleth – Four new partner institutions configured access to HathiTrust via Shibboleth in September. Logging into HathiTrust provides students, faculty, and other affiliates at partner institutions the ability to download a full-PDF of all public domain materials. It also enables use of HathiTrust's Collection Builder tool with a local sign-on. HathiTrust plans to use Shibboleth to offer additional features and services to partner institutions in the future.

October 31 Partnership Deadline – A number of partners have joined HathiTrust in the last several weeks, and we will be announcing several more throughout October. Institutions are joining ahead of an October 31 deadline, by which institutions must become members in order to participate in a constitutional convention that HathiTrust will hold in 2011. In this convention partners will conduct a formal review of HathiTrust governance and sustainability and shape future directions for the partnership.

In This Newsletter

Late Breaking News

HathiTrust Welcomes TRLN and Dartmouth – The Triangle Research Libraries Network (TRLN) and Dartmouth College have joined HathiTrust. TRLN will be contributing public domain volumes digitized through in-house initiatives and partnerships with the Internet Archive. Dartmouth joins HathiTrust as the first partner under HathiTrust's new cost model. Visit http://www.hathitrust.org for more details.

HathiTrust Digital Library

Update On September Activities

Working Groups

Communications – The Communications Working Group's activities in the past month continued to focus on plans and processes for receiving new partners. In its in-person meeting in Chicago on September 24, the group made progress towards a communications and marketing plan and provided feedback to the website redesign project in an interactive session. An additional "HathiTrust 101" presentation was held on September 9. Slides from the presentation are available at http://www.hathitrust.org/documents/HathiTrust101-201008.ppt.

Development Environment – The transition of all ongoing HathiTrust development to the new development environment is in its final stages. Work in September focused on developing testing and release processes, and the transition is expected to be complete by mid-October.

Discovery Interface – The Discovery Interface Working Group (DIWG) convened in Chicago on September 24th, as part of the larger two-day face-to-face meeting of HathiTrust partners. The group had a productive discussion, largely focused on re-structuring the working group, scoping its future work, and clarifying the DIWG's relationship to other HathiTrust working groups. Three areas of future focus are phase 2 of the HathiTrust-OCLC catalog, full-text search services, and usability for both of these projects (in collaboration with the Usability WG). Several other areas of potential development falling under the topic of "end-user services" were identified for further investigation. As a follow up to this meeting, the chair of the group took the DIWG's ideas and questions to the HathiTrust Strategic Advisory Board, to whom the working group reports, for consultation. Meanwhile, a beta release of the phase 1 HathiTrust-OCLC catalog is expected by the end of 2010.

Usability – The Usability Working Group also met in a face-to-face meeting in Chicago in September. The group has been working on forming connections with other working groups, understanding existing HathiTrust interfaces and functionality, and determining the scope of the work the group will undertake. Formal liaisons to the Communications and Discovery Interface working groups were established in September.

Ingest

NYPL and Illinois – HathiTrust began ingest of content from New York Public Library and the University of Illinois in September, including more than 80,000 Google-scanned volumes from NYPL and more than 14,000 Internet Archive-scanned volumes from Illinois. Illinois is the third institution following the University of California and Columbia University to contribute volumes digitized by the Internet Archive. Ingest of content from Yale University will begin in October.

Development Updates

Bibliographic Metadata Management – During the month of September, HathiTrust teams in Michigan and California focused on producing a planning document for a HathiTrust Metadata Management Service to be developed, hosted, and run by the University of California. The document codifies the goals, success criteria, assumptions, and requirements of the system, as well

New Growth

Number of volumes added:

	Month of September	Overall
Columbia Univ.	38	54,983*
Indiana Univ.	140	178,102
New York Public Library	81,228	81,228
Penn State	0	33,357
Univ. of California	37,868	1,807,095
Univ. of Illinois	14,428	14,428
Univ. of Michigan	19,820	4,149,828
Univ. of Minnesota	20	73,694
Univ. of Wisconsin	4,544	383,655
Total	158,086	6,778,155

Public Domain (~21% of total)

Total	108,028	1,419,306

*Incorrectly reported as 56,730 in the previous update

Presentations

HathiTrust 101	Sept. 9
Indiana Univ.	Sept. 27
Library of Congres	Sept. 27

Please see http://www.hathitrust.org/papers for links to all HathiTrust presentations, papers, and reports.

HathiTrust Digital Library

Update On September Activities

as strategies for migration, integration, and acceptance testing. The planning document provided a focus for face to face meetings in Chicago on September 23-24. The University of California expects to begin developing the system later in the fall.

Large-scale Search – Staff at the University of Michigan conducted additional testing in September to better understand scalability and memory issues in full text indexing and to tune searching and indexing process. As a result of the testing, Michigan developers were able to solve issues related to the size of the index and memory use, improving the speed of full text searches.

PageTurner – Michigan staff completed work to add a progress bar for full-volume PDF generation in the PageTurner application. The new feature will be put into production in October. Staff at Michigan also began light experimentation with the coordinate OCR text format to investigate possibilities for use.

Collection Builder – Staff members at Michigan and California Digital Library discussed improvements that could be made to the collection builder interface.

Improvements to Ingest – Work on architectural improvements to ingest that was reported in the Update on July 2010 Activities is nearly complete. The major areas of enhancement are more thorough barcode validation, generalization of routines that create METS and PREMIS markup, an improved logging framework, and the use of XPath for XML validation. Along with these changes, a regression testing methodology is being developed to exercise all validation logic.

Outages – HathiTrust was unavailable on Tuesday, September 7 from 4:20pm to 5:00pm due to a software error that was undetected during release testing.

Partner News

SFX HathiTrust Target – As reported in the Update on August Activies, the University of California has created an SFX "target" to link to the HathiTrust Digital Library. HathiTrust partner libraries using Ex Libris' SFX scholarly linking who implement the new target will be able to include a link to HathiTrust books in their SFX menu window. Library users will be able to see immediately whether a HathiTrust book is available electronically and if so, link to the full text in the HathiTrust Digital Library. For a copy of the code, email Margery Tibbetts, California Digital Library, at CDL-SFX-Tech-l@ucop.edu.

Special Message on Security and HathiTrust

From: John Wilkin
To: HathiTrust Partners

Dear Colleagues,

In recent months, we have seen an increase in the number of incidents of large-scale downloading of HathiTrust resources, and even the availability of applications to aid in downloading and circumventing limitations on access. HathiTrust is strongly committed to openness; even so, we occasionally encounter issues related to security. For example, overly aggressive crawlers can consume such large amounts of system resources that they affect access for typical user access. External agents may have a negative affect intention-

October Forecast

- Finalize membership for the 2011 constitutional convention
- Continue work toward redesign of the HathiTrust.org website.
- Begin ingest of Yale content
- Complete transition to new development environment
- Receive feedback on ingest policies and specifications

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

Update On September Activities

ally (a sort of "denial of service attack") but most are simply poorly designed or the persons running them do not understand why limits have been put in place.

In addition to general system resource concerns, HathiTrust holds many resources that have contractual obligations for limiting some types of systematic or large-scale downloading. Although the most famous example of this is Google-digitized content, which requires the participating library and HathiTrust to prevent uncontrolled robotic activity, e.g., "to implement technological measures (e.g., through use of the robots.txt protocol) to restrict automated access to any part of such entity's website where substantial portions of such Digital Copies are available" (http://www.lib.umich.edu/files/services/mdp/Amendment-to-Cooperative-Agreement.pdf). Publicly available publisher resources may also have these types of constraints.

HathiTrust uses strategies to enforce some limits on use, but balances this with strategies to provide more open access. As a general preventative measure, HathiTrust employs forms of "throttling" as one mechanism to protect the system from malicious external forces. To be frank, this sort of approach is fairly coarse and is insensitive to whether the user's activities are appropriate or inappropriate; we hope to increase the sophistication of these mechanisms as time goes on to better distinguish and permit legitimate heavy use. One approach we take to give more generous access is through user authentication at partner institutions. In this case, Google-digitized content, which can typically only be viewed one page at a time, is downloadable by the authenticated user as a whole volume (or parts of a volume). We can also use the Data API as a vehicle for broader access. In cooperation with a partner institution, we can permit a specific IP address for authorized larger-scale uses. For example, we also hope to add functionality to the Data API to enable authorized uses, perhaps even of in-copyright materials. For most partner-digitized resources, where limitations are not required, the Data API is an excellent tool, and we have deployed a demonstration tool that facilitates large-scale downloading (temporarily available at http://www.lib.umich.edu/two-over-threehundred/).

We want to remind partner institutions that responsible management of the repository requires us to collaborate in managing types and levels of access. Many factors, including copyright law and contracts, come into play in guiding our strategies. Institutional representatives may be asked to investigate and help resolve issues of apparent violations. We can also work together to provide broader access where possible. I hope you'll be able to aid us in managing this balance.

Sincerely,

John Wilkin, Executive Director
HathiTrust Digital Library