HathiTrust Digital Library

Update On October Activities

November 12, 2010

Top News

New Partners – HathiTrust is pleased to announce that membership for the 2011 Constitutional Convention has been finalized. More than fifty institutions have joined HathiTrust and will take part in a collective process next year to determine the governance structure for HathiTrust in its next phase, and shape future directions for the partnership. The official announcement of partners will be made in the coming week.

Minnesota Image Ingest – The University of Minnesota and its statewide partners – the Minnesota Digital Library (MDL) and the Minnesota Historical Society (MHS) – are working with the University of Michigan technical team to lead a project to develop a prototype workflow for depositing images and associated metadata into the HathiTrust system for access, storage, and preservation purposes. This effort will help HathiTrust meet one of its key functional objectives, support for non-book/non-journal digital content. The project, scheduled to run from September 2010 to December 2010, involves a variety of content types, from simple continuous tone images, to compound objects made up of a series of images in a specified structural relationship. Demonstration content will include several tens of thousands of images from the MDL database and a 10,000 image subset of the MHS collection management system. Significant progress has been made in defining and testing the METS, PREMIS, and XMP data that are required for ingest into HathiTrust. Consultants to the project are Eric Celeste and Katherine Skinner, who are working very closely with the Minnesota and Michigan technical teams. Partner colleagues from Wisconsin and Northwestern are also involved in providing input and review to the project. For more information, please contact John Butler <j-butl@umn.edu>.

Bibliographic Data Management – The UC project team has begun the first phase of work on the new metadata-management system, with an overall goal of fostering a transparent model for HathiTrust metadata management. The project will address decoupling of the HathiTrust and University of Michigan production systems to enable the HathiTrust development environment to support the development, integration and deployment of a major HathiTrust Repository component. This endeavor is exemplary of how sub-projects may be delegated and managed within the HathiTrust collaborative structure.

IMLS Grant for Validating Quality – HathiTrust will be participating in a grant received by Paul Conway, Associate Professor at the University of Michigan School of Information, from the Institute of Museum and Library Services to validate the quality of products produced in large-scale digitization projects. See the full announcement for more details.

Developing Support for Publishing – The University of Michigan is beginning development of tools that will enable the use of HathiTrust as a publishing platform for both encoded-text and page-image content. This 2-year effort will include the creation of ingest, management, and presentation tools for journals; support for books is planned as a later stage of development. Publish-

In This Newsletter

Presentations

LuceneRevolution	October 7
ARL Fall Forum	October 15

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

HathiTrust Digital Library

Update On October Activities

ed materials will be included in large-scale search and viewable in a new interface that can be configured to reflect individual brands. The MPublishing division of the University of Michigan Library will partner in the development of this platform with the intention of making it the permanent host of MPublishing's Open Access journals. This effort is also a significant step toward the fulfillment of HathTrust's objectives in supporting formats beyond digitzed book and journal content; electronic publications in particular.

Search Widgets – California Digital Library has developed a search box that can be placed on any web page to search HathiTrust directly from websites, learning management systems, library guides and more. There are a number of different versions to choose from, all variations on bibliographic search and full text search. The code for the search box is available at http://www.hathitrust.org/widgets. Anyone may take this code and embed it in a web page. The search box may be especially useful in contexts where users are seeking to discover full-text materials published before 1923, government documents, and historical information.

Website Redesign – The HathiTrust.org website was launched in October 2008, when the new partnership was officially announced. At that time, HathiTrust did not have its own bibliographic catalog, and the ability to search the full text of materials in the repository was more than a year away. HathiTrust.org was created as a separate project site with the expectation that it would one day be integreated with HathiTrust repository and search services in a single interface. That time has arrived, and with significant coordination between the Communications and Usability working groups, and developers at the University of Michigan, the project website has been restructured and redesigned, and the interface integrated with HathiTrust applications to provide a single portal for all HathiTrust activity. Please visit the new website at HathiTrust.org.

Partner Local Digitization – Comments were received from a number of institutions on drafts of HathiTrust's policy and specifications framework for accepting content from a variety of digitization sources. The comments are being collated and incorporated into the framework, which staff at the University of Michigan hope to finalize by the end of November. At that time, staff also expect to have the technical systems fully in place to begin routine ingest of locally digitized content.

Working Groups

Communications – The group continued their work on evaluating a draft of the new website design and content, and on announcing many new partners.

Development Environment – The new development environment is now being used actively for all HathiTrust development, testing, and production release processes at Michigan. The working group, which met less frequently during the intensive migration process, will now be discussing the current provisions of the environment and welcoming partners interested in using it to contact us for access. Incremental improvements to the environment will continue. Major planned refinements include increasing storage capacity, upgrading the MySQL service to new serv-

New Growth

Number of volumes added:

	Month of October	Overall
Columbia Univ.	511	57,279
Indiana Univ.	232	178,334
New York Public Library	175,534	256,762
Penn State	6	33,363
Princeton Univ.	77,494	77,494
Univ. of California	18,107	1,825,202
Univ. of Chicago	2,204	2,204
Univ. of Illinois	0	14,428
Univ. of Michigan	30,022	4,179,850
Univ. of Minnesota	29	73,723
Univ. of Wisconsin	16,636	400,291
Total	321,011	7,099,166

Public Domain (~24% of total)

Total	252,970	1,672,276

HathiTrust Digital Library

Update On October Activities

ers, and formalizing processes for refreshing sample content.

Discovery Interface – With the beta release of the phase 1 HathiTrust-OCLC prototype catalog quickly approaching, the Discovery Interface working group (DIWG) is working with OCLC and the Communications working group to prepare a public announcement of the catalog. HathiTrust staff members are also making adjustments to the current HathiTrust interface to accommodate the new beta catalog.

As of the end of October, the Full-Text Search Working Group, a subgroup reporting to the DIWG, now has a finalized charge and membership. This subgroup, to be chaired by Tom Burton-West and involving members from four partner institutions, will have a two-part focus: 1) the group will work on short-term improvements to the HathiTrust full text search, and 2) it will evaluate user needs and draft a long-term work plan for the full text search functionality and interface. The subgroup is expected to begin work in November.

Usability – The Usability Working Group has been actively participating in other committees via liaison roles. The Communications group liaison helped refine the information architecture of the HathiTrust.org website. The Discovery Interface Working Group liaison began discussions with OCLC for a second round of usability testing to evaluate WorldCat Local for HathiTrust. The group also consulted on a usability issue regarding login and made recommendations for improvements.

Ingest

New Partner Ingest – Ingest began in October of content from Princeton University and the University of Chicago.

Development Updates

Large-scale Search – Staff at the University of Michigan are preparing to regenerate the complete full-text index of volumes in the repository to take advantage of new and improved functionality in Solr 1.4.1, extend metadata support in the schema, improve journal volume metadata, and improve overall performance with better pre-filtering to reduce the number of unique terms. The actual process of re-indexing is expected to take approximately 40 days and is targeted for completion by the end of January.

A temporary solution was put in place to solve the problem of too many unique terms described in a Large-scale Search Blog post in February. A more permanent solution is in the works. Details are posted in the Large Scale Search blog from October 5.

PageTurner – Michigan made progress on the integration of the Internet Archive BookReader software (formerly known as the GnuBook) into the HathiTrust PageTurner. Building on the advanced working prototype developed by the California Digital Library, the present task is to fully integrate the code for production use. Some enhancements, such as the display of OCR text, are in the works. During October, the University of California wrapped up its involvement in integration of the BookReader with the HathiTrust PageTurner and handed off development to staff at Michigan. UC's resources will be redeployed to the HathiTrust Metadata Management System development.

November Forecast

- Continue work on BookReader integration into PageTurner, full-text re-indexing, and Data API security enhancements
- Finalize framework for ingest of locally-digitized content and establish technical systems for routine ingest
- Begin ingest of Yale volumes and content from new partner institutions

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

Update On October Activities

John Wilkin praised UC's involvement as "incredibly helpful in moving forward this critical piece of our environment".

Collection Builder – Developers at Michigan discussed development options for proposed changes to the Collection Builder application interface. The immediate objective is to make the list of collections easier to use, based on guidance from the Usability Working-Group. Preliminary conversations among Michigan staff about how to support full text search of significantly larger collections were also initiated.

Data API – Michigan staff began work in October to specify a security layer for the HathiTrust Data API, and enhanced the API to support dissemination of coordinate OCR.

Outages – HathiTrust's search-within-a-book feature was unavailable from Tuesday, October 12 at 3:00pm EDT to Wednesday, October 13 at 12:00pm EDT due to a software change that had an unanticipated impact on this functionality.

Partner News

CDL Object Validation Tool – Staff at CDL demonstrated the Object Validation Tool to team members at Michigan in late October. HathiTrust teams at CDL and Michigan discussed strategies for generalizing the tool for other partner use. CDL is currently planning to check the code into the new code repositories in the HathiTrust development environment and solicit participation from another HathiTrust partner to experiment with running the tool.