HathiTrust Digital Library

Update On December Activities

January 14, 2011

Top News

Minnesota Image Ingest

From September through December 2010, the University of Minnesota worked with HathiTrust on a prototype project to add 50,000 image objects and associated metadata from the collections of the Minnesota Digital Library, and another 8,000 from the Minnesota Historical Society. To date, numerous lessons have been learned regarding format standards, identifiers, and rights issues related to image data sourced from different institutions. The project is also expected to shed some light on the costs of archiving image data in HathiTrust relative to that for published books and journals. Completion of the project and release of the final report are expected in the next month. For more information, please contact John Butler (j-butl@umn.edu).

Local Digitization Ingest

Staff at the University of Michigan incorporated feedback received from a variety of sources in October and November into the policy and specifications framework for scaling ingest of locally-digitized partner materials. The framework was finalized and approved, and is available at http://www.hathitrust.org/ingest. The bulk of the enhancements to ingest systems to support this work were completed as well, and ingest of Minnesota images and a sample of Yale content have occurred in the new ingest environment. The new environment will eventually be used for Ingest of all materials, including those downloaded from Google and the Internet Archive.

CC licenses

Developers at Michigan began implementing changes to support Creative Commons licenses in the repository's rights management scheme. Development is expected to be completed in February. Beginning March 1, CC licenses will be included in the "Rights" and "Rights determination reason code" fields of tab-delimited files HathiTrust makes available for download. These files contain copyright, identifier, and limited bibliographic information for all volumes in the repository.

Print Holdings Information

At the beginning of December, HathiTrust requested information from partners about the print holdings of their respective libraries. The information is being used to assemble a database that will support the new cost model all partners will participate under in 2011, facilitate legal access uses of materials in HathiTrust (e.g., section 108 uses and access for users with print disabilities), and form a base for collaborative collection management and collection development activities among the partnership. Partners are requested to provide this information by the end of February.

January Forecast

- Complete Minnesota Image Ingest Pilot
- Complete draft of marketing and communications plan
- Complete full-text re-indexing
- Continue work on BookReader integration

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

Working Groups

Collections

The recently-formed HathiTrust Collections Committee is a new standing committee reporting to the Strategic Advisory Board charged with establishing strategic directions related to the collection, including collection building and management (see charge and membership). The Committee held its first meeting in October 2010. Examples of issues currently under consideration include the role of duplicates in HathiTrust, models for shared management of print collections, and a variety of rights-related concerns. A more general area of investigation will be an exploration of specific collection development opportunities that the partnership might pursue and recommendations for how such activities should be prioritized and carried out, including considerations relating to non-book formats and collaboration with other initiatives. The Committee is considering a survey of the membership in order to assemble a better picture of partner expectations and aspirations. Input from other HathiTrust partners is welcomed; feel free to contact Ivy Anderson, chair (ivy.anderson@ucop.edu) or another member of the Committee with comments and questions.

Total Volumes Added

	December	Overall
Columbia University	34	57,316
Cornell University	161,803	215,610
Indiana University	322	179,351
NYPL	191	258,019
Penn State University	1,035	34,400
Princeton University	364	208,506
University of California	130,891	2,048,246
University of Chicago	4	2,444
University of Illinois	0	14,428
University of Madrid	78,256	78,256
University of Michigan	14,956	4,249,620
University of Minnesota	2,507	76,371
University of Wisconsin	6,922	413,987
Yale University	144	144
Total	397,429	7,836,698

Public Domain (~25% of total)

Total	138,527	1,959,223

Communications

The Communications Working Group continued to craft a marketing and communication plan for 2011, and expects to send a draft to the Executive Committee and Strategic Advisory Board by the end of January.

Discovery Interface

Despite holiday vacations, December was a busy month as the Discovery Interface Working Group (DIWG) worked with OCLC to take the final steps towards releasing the version 1 prototype catalog. With endorsement by the Strategic Advisory Board, the DIWG is now pleased to announce that the public release will go forward as planned in mid-January. Keep an eye out for the official announcement from OCLC. In addition to planning for the scheduled release, the DIWG is also developing post-release processes for managing user feedback and monitoring the system, an operational responsibility that will be supported by the California Digital Library. A three-month period of user testing will take place post-release, which will provide valuable input and help shape version 2 of this important effort.

Usability

The group reviewed a plan for a second round of usability for the HathiTrust-OCLC prototype catalog to be conducted in conjunction with OCLC and the Discovery Interface Group. The group also provided feedback on some proposed designs for a new PageTurner and revised home page.

Ingest

Content from Yale

A sample of digitized content from Yale University Library was ingested in December. The content is being reviewed by staff at Yale ahead of full ingest, which is expected to begin in January.

Development Updates

Bibliographic Metadata Management

In December 2010, the California Digital Library (University of California) and HathiTrust solidified business arrangements and posted a Principal Metadata Analyst position to support development of the HathiTrust Metadata Management System. The University of Michigan transferred input files and scripts describing current bibliographic metadata transformation practices so work can begin at CDL on developing routines for metadata ingest. Progress is also being made on the development of the core metadata storage system. Project information, including overview, milestones, and timeline, is available at http://www.hathitrust.org/htmms.

First Storage Replacement Cycle Begins

The original storage equipment purchased in late 2007 for HathiTrust has reached its retirement age of approximately 3 years. HathiTrust uses modular storage, and modules may be removed and replaced without disrupting service. Data migration is handled in the background and is fully automatic, though the process does take time to complete. Michigan staff have developed a plan for the upgrade at the Michigan site, and once in progress, will start a similar upgrade process at the Indiana site. The replacement of storage hardware will now be an annual or semi-annual process, shadowing historical patterns of growth and storage purchases.

Update On December Activities

Repository Auditing

Michigan staff have begun developing audit mechanisms to verify the integrity of content stored in the repository. These processes will augment existing features of the storage system that routinely scan, detect, and repair hardware-level data storage errors (commonly referred to as "bit rot"). As part of this initiative, a preliminary integrity check of all repository Zip archives--which are used as containers for image, text, and metadata files--was run. The check revealed an error in one page of one volume resulting from a problem with data synchronization from Michigan to Indiana; this was easily corrected. Developers are now coding and testing a comprehensive set of audit routines to ensure that all items recorded as being present in the repository are stored properly and are fully intact, including checksum validation.

Full-text Search

Work to re-index the full text of all volumes in the repository continues, and after encountering some out-of-memory problems, additional tuning and upgrades were made to Solr servers. Performance is almost an order of magnitude better than expected, owing to new optimizations that are being tested for the first time. This effort is on schedule for completion by the end of January.

PageTurner

Staff at Michigan outlined final steps for integrating BookReader with PageTurner. Changes to the user interface layout and performance testing are the main areas of remaining work. The layout design was completed in December, and will be coded in January.

Data API

In November, developers at Michigan made a change to the structure of the URL used to retrieve content and metadata from the repository through the Data API. The old structure will no longer be supported as of March 1. Users of the Data API should consult the URL structure specified in the current Data API documentation.

Outages

There were no outages in December.

Letter from the Executive Director

HathiTrust and Growth in 2011

HathiTrust grew rapidly in 2010, increasing the relevance of the HathiTrust collection to the management of our print collections. The HathiTrust collection will reach two milestones in early 2011: in January, HathiTrust will reach 2 million public domain volumes and, soon afterwards, the collection as a whole will pass 8 million volumes. Thanks to ongoing collection analysis work by OCLC Research, we know that North American research libraries overlapped with the HathiTrust collection at a median rate of just over 33% and that the median rate of overlap with the Oberlin Group of Libraries was closer to 40%. In each of the last two years, the repository grew by nearly 3 million volumes, and the rate of overlap between ARL libraries and HathiTrust grew at about 1% per 240,000 volumes of HathiTrust growth.

If you manage a rich library collection, you will find a significant percentage of your holdings online in HathiTrust; moreover, because of the size and diversity of the HathiTrust collection, you can add over one million new public domain volumes to your collection through the addition of HathiTrust links to your catalog. It is certainly true that the obstacles to using the in-copyright volumes for the delivery of mainstream library services are immense, but just as certainly this phenomenal collection can help us change the way that we administer the storage of little-used print collections. We can confidently say that in 2010 we made progress on HathiTrust's mission-related goal "[t]o stimulate redoubled efforts to coordinate shared storage strategies among libraries, thus reducing long-term capital and operating costs of libraries associated with the storage and care of print collections." With 2011 also comes a change in HathiTrust's growth trajectory and the need for a better understanding of the challenges and opportunities for future growth as a tool in shaping our collection management. To date, the two largest depositors in HathiTrust have been California and Michigan, representing approximately 26% and 57% of HathiTrust's total deposits. The growth of new content from these institutions will slow in 2011 and, barring significant changes, HathiTrust's collection will grow by fewer than 2 million volumes in 2011. Even this more modest growth is good news: it may lead to a 40% overlap between HathiTrust and ARL libraries.

We know from OCLC's analysis (Cloud-sourcing Research Collections: Managing Print in the Mass-digitized Library Environment," by Constance Malpas) that even 33% overlap is of significant value to many of our libraries. Still, HathiTrust *needs* growth. HathiTrust's value as a pivotal resource in viewing the aggregation of our collections benefits from growth. Building comprehensive and accessible online collections is a necessary part of our strategy for designing effective print storage and access strategies. This is true, for example, for US federal government publications, and is just as true for the large volume of mid-20th century publishing, much of which languishes in suboptimal off-site storage facilities in our libraries. While a 33% overlap between the HathiTrust collection and the collections of ARL libraries is valuable, 50% and 60% overlap can be a powerful catalyst to major changes in print storage.

Our growth is key for a broad array of library access and management opportunities. The case for HathiTrust as a catalyst for changed print management has become clear to our partners. There are other important reasons as well:

- Large numbers of **titles appear to be protected by copyright but are in fact in the public domain**. Digital availability has been a necessary piece of the strategy that has helped HathiTrust partners open access to 55% of the books published in the US between 1923-1963. A new effort will also open access to large numbers of non-US works. Because of our investments to date, adding to the US 1923-1963 collection will also increase what we know to be in the public domain.
- Partners are now working to assign resources to **securing permissions** for use of books and journals now online. Preliminary efforts have opened access to thousands of volumes. Online availability ensures that opening access is merely a matter of flipping a switch once permission is secured, increasing our incentive to work on the problem.
- The richness of the collection makes possible important lawful uses of in-copyright materials. Many library volumes are eligible for **uses under Section 108** provisions in US copyright law, and the online availability of a volume can help a library provide lawful access to an out-of-print work that is damaged, deteriorated, lost or stolen. HathiTrust partners are poised to follow Michigan's lead and use the online volumes to provide **services for their users with print disabilities**. Again, this is an activity that can only happen when the volume in question is online.

Despite the likelihood of lower growth for 2011, the possibility of future HathiTrust growth remains great. Overlap between HathiTrust and ARL libraries will probably grow to "only" 40% in 2011, but based on current prospects, that overlap could grow to 60% in the coming year. The impediments to that growth are significant but tractable. For example, several newer HathiTrust partners who have also invested in local repository infrastructure have millions of volumes of digital content that would enrich the collection. Needless to say, the prior investments by these institutions make the additional cost of deposit in HathiTrust expensive, but because of a predominance of pre-1978 materials, this content is a rich resource for copyright determination work and would significantly increase overlap. Some partners also face legal and contractual obstacles. The majority of volumes digitized from CIC are embargoed under the presumption that they are in copyright. As we have learned through our copyright determination work, significant percentages of this content are actually in the public domain, and again these volumes would also increase overlap.

Update On December Activities

The growth of HathiTrust and the nature of the collection have created critical opportunities, but we must continue to push toward the goal of a nearly comprehensive digital collection in order to benefit fully from what that collection can offer. Copyright determination work, securing rights, and especially print storage management will all be furthered by growth. We will continue to address existing impediments and urge our partners to help round out HathiTrust's large and increasingly comprehensive collection.

John Wilkin
Executive Directory, HathiTrust