



## Update On April Activities

In This Newsletter

### Top News

May 14, 2010

**Internet Archive Ingest** – Through the collaborative efforts of California Digital Library (CDL) and the University of Michigan, University of California volumes digitized by the Internet Archive began flowing into HathiTrust in April. This achievement is significant because of the amount of additional public domain volumes UC will make available and preserve in HathiTrust (approximately 200,000), and because of the new channel it opens for partner ingest. HathiTrust has offered deposit of Google-digitized materials at no cost to partners since its inception in 2008. This ingest service is now extended to partner institutions' Internet Archive-digitized volumes. The collaboration has also significantly advanced HathiTrust's progress towards repository-wide standards for digital object packages and guidelines for deposit. The first Internet Archive-digitized volume to enter HathiTrust was entitled "The Dawn of All", appropriately marking this major step in HathiTrust's ability to preserve the wide variety of non-Google digitized content produced by library partners. More than 90,000 of UC's Internet Archived digitized volumes are in HathiTrust as of the time of newsletter release.

**Partner Local Digitization** – A group of staff members at the University of Michigan are developing a formalized process for receiving, analyzing, and preparing locally-digitized content from partner institutions for deposit. Leveraging experience gained with Internet Archive ingest, the group will be working over the next several weeks to publish content guidelines that will aid partners in assessing the readiness of

their content for deposit. The guidelines will also provide consistent benchmarks for UM staff to use in performing the transformations and normalizations that may be necessary to assemble coherent and consistent archival packages.

**Bibliographic Management System** – The University of California is engaged in the process of designing a new bibliographic management system for HathiTrust. HathiTrust team members at CDL and the University of Michigan engaged in multiple teleconferences throughout April to define the scope of services and functions provided in the current management system at UM (Ex Libris' Aleph product), and by the systems that it supports such as HathiTrust's temporary catalog and Bibliographic API. Development of the new system will diversify the management of support systems in HathiTrust. It also presents a valuable opportunity to revisit current practices and assumptions and reengineer existing processes to be more efficient. Team members have worked through a number of architectural issues in the design and May discussions will focus on strategies for transition to the new system. Development of the system has not yet begun and there is no timeline currently for implementation.

**Website Changes** – Staff from UM and UC took part in a usability exercise in April geared towards improving the navigability of the HathiTrust.org "About" website. Improvements are ongoing, and UM staff have implemented the first in a series of changes to occur. Top navigation has changed and sub-navigation is now provided in the left

### Top News

- Internet Archive Ingest
- Partner Local Digitization
- Bibliographic Management
- Website Changes

### Working Groups

- Discovery Interface
- Development Environment

### Ingest

- Penn State and UC Bibliographic Updates

### Development Updates

- Shibboleth
- Large-scale Search
- Collection Builder
- PageTurner

### New Growth

Number of volumes added:

	Month of April	Overall
Indiana Univ.	1,815	176,835
Penn State	10,661	17,274
Univ. of California	39,079	1,203,556
Univ. of Michigan	79,027	3,939,722
Univ. of Minnesota	7,189	73,065
Univ. of Wisconsin	26,317	330,004
Total	164,008	5,740,496

Public Domain (~17% of total)

Total	48,729	903,519
-------	--------	---------

**There's an elephant in the library.**





## Update On April Activities

May Forecast

side of the interface. Search functionality has also been added. New content and additional architectural changes will be made in the coming weeks as documentation assembled for HathiTrust’s audit with CRL for compliance with TRAC draws is made available, and additional usability tests are conducted.

### Working Groups

**Discovery Interface** – OCLC loaded more than 1 million HathiTrust records into WorldCat in April, bringing the total number of records to close to 2.3 million by the end of the month. These constitute over 60% of the total HathiTrust records that are currently available. Loading of these records will continue throughout May. In mid-May, OCLC will release version 1 of the HathiTrust catalog internally to the Discovery Interface Working Group. Staff at the University of Michigan are preparing to make necessary changes to HathiTrust websites to accommodate the new catalog. OCLC provided a first glimpse of the new catalog to the working group in a recent WebEx session.

In parallel to finalizing the version 1 catalog, the HathiTrust team is turning its attention to the full-text search application. The Discovery Interface Working Group will assume responsibility for further developing HathiTrust’s full-text search service, and is in the process of finalizing a formal charge and roadmap for this project.

**Collaborative Development Environment** – Michigan staff, in consultation with working group members, have completed an initial design of the development environment. The design includes specific plans and conven-

tions for version control, file layout and naming, virtualization provisions for developers, and multiple test and beta instances. UM staff are now planning the details of migrating current development into the new scheme as well as configuring and building out the server resources for the environment.

### Ingest

**Penn State** – HathiTrust recently received updated bibliographic records from Penn State University for several hundred PSU-contributed volumes. The volumes, which are all in the public domain, had received an incorrect bibliographic rights determination in HathiTrust because of problems with metadata, including missing a flag indicating that volumes were government documents. With the corrected records, all of these volumes are now freely accessible. HathiTrust will be monitoring rights determinations for volumes from institutions depositing public domain-only materials. If metadata corrections are required to make volumes available, HathiTrust will notify the institutions appropriately.

**UC Delivery of Bibliographic Records** – UC recently modified the way it makes bibliographic records available to HathiTrust for volumes that are being scanned on an ongoing basis. Records will now be pulled from UC by HathiTrust, rather than pushed by UC, which will simplify the flow of ingest for these materials.

### Development Updates

**Shibboleth** – Based on input from partners, staff at UM have further refined the final list of attributes needed to provide Shibboleth services. These

- Deploy redundant hosting of large-scale search service at Indiana site
- Release Shibboleth authentication and full-PDF public domain download for HathiTrust partner institutions

### Presentations

Princeton Forum on Preservation	April 9
Bilkent University, Ankara Turkey	April 20
CDL Resource Liasons and Users Council Webinars	April 20 and 30
Lucid Imagination Webinar	April 29
Association of Research Libraries	April 30

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

There’s an elephant in the library.





## Update On April Activities

attributes, and other information about Shibboleth in HathiTrust are available at <http://www.hathitrust.org/shibboleth>. UM staff also registered HathiTrust a service provider with the InCommon Federation. Two early adopter institutions successfully tested access to HathiTrust development systems via Shibboleth, and HathiTrust is planning to release the service publicly in mid-May.

**Large-scale Search** – University of Michigan staff developed a pair of index tools for reporting specific statistics about a Solr index in April. These tools help to identify frequently occurring terms, which can be used to improve performance. The tools have been committed to the Solr code base and will be part of future Solr releases.

Work by UM team members continues on installing new servers at the Indiana site. New electrical and networking capacity has been installed, and firewalling is being reconfigured to support new remote administration and installation capabilities. Once complete, the new servers will be configured and brought online, anticipated for late May.

**Collection Builder** – Staff at UM have developed functionality that allows a user to add multiple items at once to a Collection from the full-text search results. Cosmetic changes to the user interface are required to complete this effort.

**PageTurner** – Developers from UM and CDL have been collaborating to integrate new image serving capabilities at UM with the GnuBook reader. A prototype application combining these services has been developed and next steps will involve merging the proto-

type with the current PageTurner application. Code developed by CDL to produce thumbnail views of volumes in GnuBook has been incorporated into the mainline code, maintained by the Internet Archive.

**Outages** – The beta\* large-scale search service was unavailable from 10:00am - 12:40pm EDT on Friday, April 9 to troubleshoot a hardware problem on an index server.

\*Beta services are typically non-redundant and/or volatile, and while we strive to minimize down time and report any that occurs, we do not attempt to adhere to non-peak outage windows for maintenance.

### Partner News

**UC-eLinks (SFX)** – HathiTrust books free of copyright restrictions, contributed by both UC and other HathiTrust partners, are now available via a link in UC's UC-eLinks (SFX) menu window. CDL has developed a target for SFX that exposes HathiTrust public domain books utilizing the HathiTrust Bibliographic API. CDL plans to review statistics for the new target and work with HathiTrust staff to measure the load placed on HathiTrust APIs by UC usage. There are plans to share the target with HathiTrust partners, and in the future potentially contribute it back to ExLibris. UC Davis has also created a test implementation of this functionality within Aleph.

There's an  
elephant in  
the library.

