HathiTrust Digital Library

Update On June Activities

July 9, 2010

In This Newsletter

Top News
- Shibboleth and Full-PDF Download
- SEASR

Working Groups
- Discovery Interface
- Development Environment
- Quality

Development Updates
- Large-scale Search
- PageTurner
- Collection Builder
- Storage Upgrade

Top News

Shibboleth and Full-PDF Download – HathiTrust released Shibboleth as a mechanism for partner authentication in June. Authenticated users can now download full-PDFs of all public domain volumes in HathiTrust, and access the Collection Builder feature through local sign-on. Shibboleth also lays the groundwork for future augmented services to partner institutions, potentially including the ability to make uses of digital volumes allowed by Section 108 of U.S. copyright law, and allow full access to in copyright volumes for users with print disabilities.

Full-PDF Download: The release of Shibboleth was made in conjunction with improvements to PageTurner that enabled delivery of high-resolution PDF files with embedded OCR for entire volumes. While only individuals at member institutions have access to this service across the repository, all public domain volumes that were not digitized by Google are available for full-PDF download to members and non-members alike. Right now these include nearly 100,000 Internet Archive-digitized volumes that have been contributed by the University of California, and thousands of volumes digitized locally by the University of Michigan. The partners are poised to significantly increase the amount of non-Google-digitized content preserved in HathiTrust in the near future, making many more public domain volumes freely available for download and distribution.

SEASR – HathiTrust is in the process of investigating SEASR, the Software Environment for the Advancement of Scholarly Research, as a means to provide computational access to materials stored in the repository. Staff at the University of Michigan began installation of SEASR in the HathiTrust development environment in June, and expect to gain more knowledge about SEASR and what would be involved in applying it to HathiTrust over the next several weeks.

Working Groups

Discovery Interface – As of the end of June, there are nearly 3.1 million HathiTrust records in WorldCat. Record loading is now continuing at a quicker pace, and is nearly complete. Meanwhile, the working group is in the process of configuring the HathiTrust-OCLC catalog interface to make branding and design consistent with the existing HathiTrust Digital Library system. OCLC is also making several alterations to the catalog's functionality to fully meet HathiTrust's requirements. This work is expected to extend into early August, after which time the interface will be re viewed for public beta release.

With the working group's charge expanding to include development of the HathiTrust Full Text Search, the group plans to restructure its membership in order to specifically target different areas of focus. While the new structure is still being finalized, the goal is to form various task forces to address different aspects of the HathiTrust Discovery Interface: full text search, bibliographic data management, and the HathiTrust-OCLC catalog interface.

Development Environment – University of Michigan staff continued the migration of HathiTrust applications into the new development environment in June, performing testing and config-

New Growth

Number of volumes added:

	Month of June	Overall
Indiana Univ.	236	177,333
Penn State	328	22,824
Univ. of California	616	1,509,169
Univ. of Michigan	34,605	4,056,835
Univ. of Minnesota	173	73,856
Univ. of Wisconsin	10,073	353,639
Total	46,031	6,193,386

Public Domain (~20% of total)

Total	32,805	1,208,351

Update On June Activities

uration of the GlusterFS distributed file system that will be used as the storage back-end for the environment as well. Michigan staff are in the process of setting up and testing the virtual MySQL and web service provisions of the new environment. An initial version of the development environment is being used currently by staff at California and at Michigan to make improvements to the existing PageTurner application. When configuration is complete, the environment will support HathiTrust development efforts broadly across the partnership.

Quality, Ingest, and Error Rate – The quality working group is still working through a set of scenarios for gating volumes of poor quality from entering HathiTrust, and developing a justification and recommendation for the best approach to follow. A set of larger issues around quality has also been identified, some of which deal with larger policy considerations.

Development Updates

Large-scale Search – The full text search index in Indiana was put into production by Michigan staff in early June, making the infrastructure for full text search fully redundant. Two new index build servers were also put into production in Michigan. All of the new systems have been functioning well, and the new build servers have substantially improved the performance of index building and maintenance.

Michigan staff began running tests in June to determine the effects of cache-warming on performance, as well as tests relating to scaling strategy and indexing speed. The goal of scaling tests is to determine the optimum size to use for index shards, or sections of the search index, that are stored on each index server, the optimum number of shards per server, and optimum memory allocation per server. Indexing speed is of critical importance for deploying new searching features, which often requires the entire search index to be rebuilt.

Michigan staff also developed a Lucene utility in June (Solr uses Lucene) to read an index and print out the total number of occurrences of a term. The code has been contributed and committed to the stable Lucene development branch (3.x).

PageTurner – Additional progress was made on GnuBook integration with the current HathiTrust PageTurner. Michigan investigated in particular ways to optimize the serving of thumbnails. Performance optimization for the new page image server also continued, with a focus on common CGI performance mechanisms, including FastCGI.

Collection Builder – Integration of Collection Builder functionality with large-scale search is in the final stages of testing and will be deployed in July.

Storage Upgrade – Michigan staff have ordered and received additional storage for the Indiana and Michigan sites and will be putting it into service during July and August. The upgrade requires the installation of a new, larger storage network switch, so staff will be using the opportunity to introduce a new cabling layout for the entire system. In Indiana, the upgrade and recabling work will be combined with a recommended relocation of all server equipment to another area of the data center for improvements in air handling and a transition to high-voltage power distribution. No outage is expected for this maintenance work.

Outages – HathiTrust services were unavailable on Monday, June 7 from 7:10-10:00am and on Tuesday, June 8 from 5:00-5:30pm due to a connectivity problem with one of the web servers; and on Saturday, June 25 from 8:30-10:00am due to a database server disk space shortage.

July Forecast

- Explore capabilities and requirements of SEASR
- Continue configuration of the new development environment and migration of current development activities
- Install storage upgrade at Indiana site

Presentations

ALA NISO/BISG Forum	June 25

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