Rhiannon Bettivia

Digital Preservation

Multi-Institutional Repository Report

The HathiTrust

The HathiTrust is an access program for digitized books and journals. It is a repository for academic research libraries, and currently consists of elements of the library collections from 34 universities in the United States and the New York Public library. The project started with a core group of universities, primarily those involved in the Committee on Institutional Cooperation (CIC) conference, which is an academic conference based in the Midwest that encompasses all the Big Ten sports conference schools in addition to the University of Chicago. The University of Michigan and Indiana University are leading the project, fielding contact and questions about the HathiTrust and providing storage. The stated goals of the project are to create an accessible archive of digitized materials and ensure preservation of digitized materials (2). The need it seeks to fill is to ensure the access to and longevity of major library institutions' digitized materials, and other digital materials in the long run as it has a goal of including born digital content in the future.

The project is aimed at providing access to the materials in major research libraries and preservation of digitized content. The idea is to make digitized content searchable. The items currently in the collection are submitted by the partner schools, and include materials digitized by the schools and their partners, in addition to materials from Google books and the Internet Archive. The collection consists of digitized copies of books and journals and does not currently contain materials that are born digital. However, the project does profess an interest in expanding to include born digital journals in particular and is set to launch a fact finding project in 2010 to ascertain how these can be added to the current catalog.

The HathiTrust launched in 2008 to run for 5 years, and is currently scheduled to run until 2012. In the upcoming year, the efficacy of the project will be analyzed to see if and how the project should be perpetuated. The repository was created to fulfill the TRAC guidelines for a trusted repository, and has attempted to build sustainability into the governance of the project, by tying funding to major libraries rather than simply to grants, for example. This project should be undergoing review very soon after about 3 years in operation.

The project is funded by charges to the partner schools, which currently include the following institutions (3):

California Digital Library
Columbia University
Cornell University Library
Dartmouth College
Duke University
Indiana University
Michigan State University
University of California Riverside
University of California San Diego
University of California San Francisco
University of California Santa Barbara
University of California Santa Cruz
The University of Chicago
University of Illinois

New York Public Library
North Carolina Central University
North Carolina State University
Northwestern University
The Ohio State University
Penn State University
Princeton University Library
Purdue University
University of California Berkeley
University of California Davis
University of California Irvine
University of California Los Angeles
University of California Merced
University of Illinois at Chicago
The University of Iowa
University of Michigan
University of Minnesota
The University of North Carolina at Chapel Hill
University of Wisconsin-Madison
University of Virginia
Yale University Library

As institutions join the HathiTrust, they pay a fee per gigabyte of submitted materials annually in addition to a one-time annual fee that is 25% of their overall cost per gigabyte for that year (4). This money sustains the project. University of Michigan and Indiana University also provide substantial funding. The funding structure is set to be reviewed every three years assuming that the project runs in successive 5 year phases. The project also received a grant from the Institute of Museum and Library Services to search the works already in the repository to determine copyright status and find if any additional works may actually be in the public domain (1). This project is built to include funding from university library projects and seeks to include the library and technology communities from these partner schools in the project. In doing so, they hope this will help ensure the longevity of the project because libraries are guaranteed budgets by their universities to buy and store books and journals, and this is a natural extension of that (5).

The types of materials that go into the repository can be broadly described as digitized books and academic journals. Each contributing institution decides what of its digitized materials it will submit to the repository. I suspect funding and staffing may play a large role in selection, as the project charges per gigabyte submitted.

HathiTrust currently accepts 4 major file formats of material: TIFF, JPEG2000, UTF-8, and OCR text in Google's HTML format (6). There is a deposit for member institutions. It seems as though METS wrappers are preferred and bibliographic information should be submitted in MARC or MARCXML. However, it seems that part of the ingest process entails the repository itself creating METS and PREMIS metadata packages and wrappers for the objects. The project uses library of congress call number standards.

The HathiTrust says it uses an OAIS framework for its repository and meets most or all of the TRAC specifications for a trusted digital repository (7). To ensure the safety of the content, the digital files are stored at both the University of Michigan in Ann Arbor with a mirror site in Indianapolis at the University of Indiana. In addition, there is an encrypted tape backup stored at a third location in Michigan but away from other site (8).

In reading the technical reports on disaster recovery (9), I also noticed that they discuss the dangers of hardware and software obsolescence in the index/contents, but do not seem to address this particular issue in the body of the report itself. Rather, those sections focus more on what would happen in the case of hardware failure or bugs in software programming.

Preservation is a key stated goal of the project (10). In working with the Google books project, one of the stated goals of this partnership is to ensure the long term existence of the works digitized by Google no matter what happens to the company itself (11). They do engage in a number of fixity checks like checksums to ensure data integrity (12). The project does not seem to have a concrete plan for migration in place, choosing to say instead that it has "mitigated" the need for migration by using widely adopted and openly available technology (13). This in conjunction with the failure (in my opinion) to adequately discuss obsolescence issues as part of their TRAC-inspired disaster recovery project makes me think that they want to be a preservation archive but now are more in the limited access archive stage.

I did try playing with the access interface. It allows a general catalog search and a search specifically for full text public domain materials. Like a traditional library search, you can input key words, authors, titles, etc. The University of Michigan Press site notes a feature wherein a search for a book published by U-M Press will include a link for "easy" purchase of the materials (14). I was interested in seeing if this was true, however a search of the only U-M Press titles I knew rendered no results: these books, like *Guitars, Bars, and Motown Superstars* by Denis Coffey, were not available on HathiTrust, meaning I guess that they either haven't been digitized yet or that they haven't been ingested into this particular collection yet. The search interface also allows for people to create logins and with this login create their own special collection of materials, including full text items and catalog links for items with limited access. Limited access items (items controlled by copyright restriction) cannot be view full text but can be searched using a tool like Google's look inside this book feature. Libraries can also submit materials that will not be accessible at all and will simply be stored by the repository without access granted to anyone outside that institution (15).

The quality of images available is also discussed on the website's FAQ. They advocate using the highest quality possible, but say that lower quality materials will be accepted and part of the ingest workflow will focus on improving quality (15).

The HathiTrust calls its platform unique and cites its "special curation" as one of the benefits (15). The project has engaged in creating a page turning application and collection building applications to allow people to search through items and create their own collections. They are also developing a discovery catalog in conjunction with OCLC (16). I think this project is probably unique in its scope in that it is trying to specifically link and target particular institutions and particular aspects of their collections. However, it doesn't seem so unique—it is not so different from something like WorldCat, only its scope is more limited; on that note, HathiTrust is working with WorldCat, uploading its own information into WorldCat's discovery interface (17

Webography (all materials accessed in 10/2010):

1. http://www.libraryjournal.com/lj/home/887388-264/hathitrusts_copyright_detectives.html.csp
2. http://www.hathitrust.org/mission_goals
3. http://www.hathitrust.org/about
4. http://www.hathitrust.org/cost
5. http://www.hathitrust.org/documents/This-Library-Never-Forgets.pdf
6. http://www.hathitrust.org/technology
7. http://www.hathitrust.org/documents/hathitrust-ifla-201008.pdf
8. http://www.hathitrust.org/documents/hathitrust-ifla-201008.pdf
9. http://www.hathitrust.org/technical_reports/HathiTrust_DisasterRecovery.pdf
10. http://www.hathitrust.org/mission_goals
11. http://bits.blogs.nytimes.com/2008/10/13/an-elephant-backs-up-googles-library/
12. http://www.hathitrust.org/preservation
13. http://www.hathitrust.org/objectives
14. http://ns.umich.edu/htdocs/releases/story.php?id=7354
15. http://www.hathitrust.org/faq
16. http://www.hathitrust.org/access
17. http://www.hathitrust.org/updates_march2010

Additional reading and listening:

http://www.nypl.org/press/press-release/2010/05/24/nypl-takes-giant-step-preserving-its-digitized-collections

http://www.press.umich.edu/digital/hathi/

http://www.cic.net/Home.aspx

http://en.wikipedia.org/wiki/HathiTrust

http://www5.oclc.org/downloads/programsandresearch/parcasts/20090506RT-Wilkin.mp3