The HathiTrust Shared Digital Repository for Preserving and Providing Access to the World's Print Culture

Kevin S. Hawkins (University of Michigan, Ann Arbor; Royal Irish Academy, Dublin)
Jeremy York (University of Michigan, Ann Arbor)

HathiTrust is a shared digital repository created through a partnership of research libraries to archive and share their digitized collections.[i] The project was launched in October 2008 by 25 American research libraries, using their collective experience in digital libraries to construct a robust and scalable infrastructure to house, manage, and provide access to their collections. Unprecedented in scale, this repository reached five million digitized volumes in December 2009 and expects to reach 8 million in 2010.[ii] As an alternative to Google Books, HathiTrust, through its governance and funding models, demonstrates how to build and sustain a repository for the products of mass digitization, and its complex rights management system demonstrates how private enterprise and public institutions can work within the confines of copyright law to serve the needs of copyright holders and researchers. But best of all, the print collections of the partner institutions—added to HathiTrust in digital form as they are digitized—will be of use to researchers around the world, fulfilling the mission of HathiTrust "to contribute to the common good by collecting, organizing, preserving, communicating, and sharing the record of human knowledge."[iii]

What is HathiTrust?

A partnership of research libraries constructing a repository with the same name, HathiTrust is the largest digital preservation initiative in the history of libraries—one that aims to preserve and provide access to millions of volumes from these libraries' print collections.[iv] The

repository is co-owned and jointly managed by the partner institutions, which pay to support the infrastructure in proportion to the amount of content they deposit.[v] HathiTrust uses as its basis the repository developed by the University of Michigan Library to store page images received through its partnership with Google (formerly called MBooks), and continuing development of the repository is meant to comply with digital preservation standards and review processes.[vi]

While much of the content of HathiTrust is protected by copyright and is therefore not publicly viewable, a significant portion is in the public domain, and a small but growing portion has been made publicly available by agreement with rightsholders.[vii] Most of the content added to HathiTrust so far has been generated by Google as part of the Google Books Library Project; however, processes for ingesting non-Google content into HathiTrust are being developed as well.[viii]

How is HathiTrust different from the World Digital Library, Open Library, and Google Books?

HathiTrust, the World Digital Library, and Open Library all have global ambitions, but they differ significantly in scope and mission. While HathiTrust aims to preserve and provide access to research library collections[ix] (including limited access to material in copyright), the World Digital Library contains only selected primary source materials from national libraries and other partner institutions. HathiTrust more closely resembles Open Library, which was launched in 2007, in that both include hundreds of thousands of full-text public domain works. However, OpenLibrary's goal is "to display a page on the web for every book ever published," constructing not "a catalog to share among libraries" (like WorldCat) but "a catalog to share with the public."[x] Each book's page might include links to outside sources of information, digitized copies found online, and links to buy the book.[xi]

HathiTrust bears some similarities with the Google Books portal, through which users can search and—barring copyright restrictions—access the full text of books contributed by Google's publisher and library partners. HathiTrust's corpus overlaps significantly with Google's, but the overlap will likely decrease over time as Google and HathiTrust partners independently add content to their repositories not digitized through the Google Books Library Project. While Google's goal in its digitization is in keeping with its mission "to organize the world's information and make it universally accessible and useful,"[xii] HathiTrust focuses on preservation, with access as a significant part of its preservation strategy and broader mission to provide a public good available to anyone in the world.[xiii]

HathiTrust and Google Books have different methods of displaying search results for works protected by copyright,[xiv] and HathiTrust asserts rights under US copyright law that Google has not, for example by full access to works produced by the US federal government and allowing users with print disabilities who sign out a print copy of a book to receive full text access as well.[xv] Staff at the University of Michigan conduct review of US works published between 1923 and 1964 to see if they really qualify for copyright protection,[xvi] and HathiTrust makes available certain works by agreement with rightsholders.[xvii]

An added advantage of HathiTrust over Google Books is its integration with library catalogs, making it possible for users to locate particular issues of serials and take advantage of subject headings and other catalog metadata, which are often not displayed correctly in Google Books.[xviii] HathiTrust also makes available metadata for public-domain items in MARC21 and Dublin Core formats through OAI-PMH for any library to add to its catalog.[xix]

Governance and funding model

While cultural and educational institutions often desire to collaborate in their services or even operations, it is often difficult to make such collaborations effective. Barbara McFadden Allen describes two types of bad collaboration: collaboration where there is "goal drift" (movement away from originally stated aims) and collaboration in which there is no buy-in from administrative bodies, resulting in insufficient funding and support. [xx] She also describes two types of tensions that can arise in collaborations—a stakeholder's fear of loss of independence and a stakeholder's fear that collaboration will slow decision-making—and suggests overcoming these tensions by establishing collaborations that are voluntary and flexible, "allowing participants to invest at varying levels as warranted," with decision-making structures "that allow for broad input without dramatically slowing decision-making.[xxi] HathiTrust attempts to set up such a successful collaboration. HathiTrust partners pay per gigabyte of data deposited, plus an annual fee for each year in which new content is deposited.[xxii]Such a model allows institutions to pay in proportion to their use of the repository, ensuring a low barrier of entry for new partners. The project is currently directed by representatives of the founding institutions (the Executive Committee), with outreach and planning guidance provided by representatives of all member institutions (the Strategic Advisory Board). The governance and funding models will be revisited as HathiTrust grows.[xxiii]

Rights management

As discussed above, HathiTrust makes available to users the full text of many works not available through Google Books. Each item in HathiTrust has a record in the rights database giving the rights attribute (controlling to whom the full text may be displayed), the reason for this attribute, and the source of the decision.[xxiv] An interface to this database for staff reviewing the copyright status of items has allowed for many

items to be made available whose copyright status was previously unknown.

Access to the world's culture

When digitization of library collection first began, the cost of reformatting, storing, and providing access to the content was so high that care had to be taken to select for digitization only the rarest materials or those materials with the broadest appeal. While American libraries were among the first to digitize parts of their collections, this second stage of selection meant that nearly all digitized material was in English. However, thanks to decreasing digitization costs and to a shift in practice towards "mass digitization", all sorts of content that was previously available only in a few select research libraries is becoming available to users around the world. HathiTrust contains digitized public domain items in 190 languages (based on MARC language codes), and since only about 14% of HathiTrust items are in the public domain, there are doubtless items in many more languages digitized but not yet made available to the public.[xxv]

HathiTrust provides an innovative model for bringing together the expertise and financial resources of individual libraries to construct a shared digital repository that works within the constraints of copyright law. While preservation is its main goal, this effort also greatly increases access to the world's cultural heritage.

---

[i] The authors wish to thank Boris Orekhov for translating this article from the original English.

[ii] HathiTrust Reaches 5 Million Volumes [Электронный ресурс] / John Weise // [BLT] Blog for Library Technology. — 2009. — Режим доступа: mblog.lib.umich.edu/blt/archives/2009/12/hathitrust_reac.html.

[iii] Mission & goals [Электронный ресурс] // HathiTrust : a shared digital repository. — 2009. — Режим доступа: http://www.hathitrust.org/mission_goals.

[iv] This library never forgets / Jeremy York // Archiving 2008 [2009] final program and proceedings. — Режим доступа: http://www.hathitrust.org/documents/This-Library-Never-Forgets.pdf.

[v] How to join [Электронный ресурс] // HathiTrust : a shared digital repository. — 2009. — Режим доступа: http://www.hathitrust.org/join.

[vi] Accountability [Электронный ресурс] // HathiTrust : a shared digital repository. — 2009. — Режим доступа: http://www.hathitrust.org/accountability.

[vii] Google & the future of books [Электронный ресурс] : an exchange / by Paul N. Courant [и др.] // The New York review of books. — Т. 57, № 1 (14-го янв. 2010). — Режим доступа: http://www.nybooks.com/articles/23565.

[viii] Update on November 2009 activities [Электронный ресурс] // HathiTrust : a shared digital repository. — 2009. — Режим доступа: http://www.hathitrust.org/updates_november2009.

[ix] Welcome to the shared digital future [Электронный ресурс] // HathiTrust : a shared digital repository. — 2009. — Режим доступа: http://www.hathitrust.org/about.

[x] Frequently asked questions [Электронный ресурс] // OpenLibrary. — 2009. — Режим доступа: http://openlibrary.org/about/faq.

[xi] Open Library developer's meeting : one Web page for every book ever published // Musings on information and librarianship / Eric Lease Morgan. — 2009. — Режим доступа: http://infomotions.com/musings/open-library/.

[xii] About Google Books [Электронный ресурс] // Google. — 2009. — Режим доступа: http://books.google.com/intl/en/googlebooks/history.html.

[xiii] This library never forgets / Jeremy York // Archiving 2008 [2009] final program and proceedings. — Режим доступа: http://www.hathitrust.org/documents/This-Library-Never-Forgets.pdf.

[xiv] Universities add their own search of Google Books / Jeff Young // Wired campus / Chronicle of higher education. — 2009. — Режим доступа: http://chronicle.com/blogPost/Universities-Add-Their-Own/8901/?sid=wc&utm_source=wc&utm_medium=en.

[xv] HathiTrust accessible interface [Электронный ресурс] // HathiTrust : a shared digital repository. — 2009. — Режим доступа: http://mblog.lib.umich.edu/blt/archives/2009/10/hathitrust_acce.html.

[xvi] Copyright review management system – IMLS National Leadership Grant [Электронный ресурс] // MLibrary. — 2009. — Режим доступа: http://www.lib.umich.edu/copyright-review-management-system.

[xvii] HathiTrust rights database [Электронный ресурс] // HathiTrust : a shared digital repository. — 2009. — Режим доступа: http://www.hathitrust.org/rights_database.

[xviii] Google Books : a metadata train wreck / Geoff Nunberg // Language log. — 2009. — Режим доступа: http://languagelog.ldc.upenn.edu/nll/?p=1701.

[xix] Data distribution & APIs [Электронный ресурс] // HathiTrust : a shared digital repository. — 2009. — Режим доступа: http://www.hathitrust.org/data.

[xx] Come together right now / Barbara McFadden Allen // Digital libraries : a vision for the 21st century : a festschrift in honor of Wendy Lougee on the occasion of her departure from the University of Michigan. — Ann Arbor, MI : Scholarly Publishing Office, University of Michigan Library, [s.a.]. — Режим доступа: http://hdl.handle.net/2027/spo.bbv9812.0001.001.

[xxi] Come together right now / Barbara McFadden Allen // Digital libraries : a vision for the 21st century : a festschrift in honor of Wendy Lougee on the occasion of her departure from the University of Michigan. — Ann Arbor, MI : Scholarly Publishing Office, University of Michigan Library, [s.a.]. — Режим доступа: http://hdl.handle.net/2027/spo.bbv9812.0001.001.

[xxii] Cost calculator [Электронный ресурс] // HathiTrust : a shared digital repository. — 2009. — Режим доступа: http://www.hathitrust.org/cost.

[xxiii] Governance [Электронный ресурс] // HathiTrust : a shared digital repository. — 2009. — Режим доступа: http://www.hathitrust.org/governance.

[xxiv] HathiTrust rights database [Электронный ресурс] // HathiTrust : a shared digital repository. — 2009. — Режим доступа: http://www.hathitrust.org/rights_database.

[xxv] Public domain by language [Электронный ресурс] // HathiTrust : a shared digital repository. — 2009. — Режим доступа: http://www.hathitrust.org/public_domain_language.