HathiTrust
A Research Library at Web Scale
Heather Christenson

Research libraries have a mission to build collections that will meet the research needs of their user communities over time, to curate these collections to ensure perpetual access, and to facilitate intellectual and physical access to these collections as effectively as possible. Recent mass digitization projects as well as financial pressures and limited space to store print collections have created a new environment and new challenges for large research libraries. This paper will describe one approach to these challenges: HathiTrust, a shared digital repository owned and operated by a partnership of more than forty major libraries.

The activities of research libraries in the next five to 10 years will define the role of libraries in the digital age. The library community must now ensure that these collections not only retain their research value in a digital platform, but also realize their potential as users adjust their information needs and expectations.

—HathiTrust FAQ, July 2010 (www.hathitrust.org/faq)

In an era of mass digitization of library collections, research libraries are confronting an array of new challenges to continuing their traditional role as stewards of library collections. How will libraries ensure perpetual preservation of these sometimes massive new digital library collections, a promise Google does not make? How will libraries provide wide access to their digital collections in an appropriate manner, unencumbered by commercial interests and in support of the activities of scholars? What new possibilities for services are opened up by digital formats, and how can libraries bring those new services to their user communities? How do these new large digital collections relate to print collections, and what opportunities are available for libraries to coordinate collection management between print and digital materials? This paper will consider these challenges and then describe how HathiTrust, a shared digital repository owned and operated by a partnership of more than forty major research libraries, offers answers to some of these questions and an opportunity for libraries to collectively explore this new territory.

Literature Review

Simultaneous with lively reporting and debate in prominent popular news sources and magazines regarding Google Books and the outcomes of mass digitization projects, researchers have explored the implications of mass digitization for libraries and the collaborative possibilities for addressing the challenges of digital preservation, access, support for scholarly research, and collection management in light of new, massive digital collections. The specter of commercial hosting of research library content by Google juxtaposed with the responsibility of libraries to uphold their users' right to access information, as well as their
mission to preserve it, is a theme addressed by a number of researchers and library leaders. Hahn concluded that “it may be foolish to expect that commercial companies will share librarians’ values and commitment to digitized material preservation” and that “research libraries alone will be held accountable for fulfilling that vital preservation mission.”

In 2008, Bruntley, then executive director of the Digital Library Federation, urged libraries to “trade for our own account” because libraries “stand for what no other organization in this world can: the fundamental right of access to information, and the compulsion to preserve it for future generations.” Leetaru made a case that the output of mass digitization is “access digitization” rather than “preservation digitization.” He acknowledged that placing responsibility for long-term storage with libraries “is a legitimate argument, especially in light of Microsoft’s recent withdrawal from book digitization,” but concluded that the academic community has so far failed to provide good access service for mass digitized books.

Dougherty also explored the question of what happens if Google goes away and pointed to HathiTrust as an example of libraries taking this question seriously.

The utility of collaboration and scale for addressing problems of access and scholarly use of the mass digitized corpus is an idea that resonates with researchers. The Council on Library and Information Resources (CLIR), among others, has invested in moving research forward on the outcomes of mass digitization projects, and a number of CLIR-sponsored reports have been produced to this end. A 2007 report described ideas originating from a seminar on promoting digital scholarship and the “so-called ‘million books’ problem.” The report examined characteristics of the mass digitized corpus compared to local digitization methods (as they existed at the time), such as the greater heterogeneity of collections included, the variability of error rates that occur because of the optical character recognition (OCR) used in mass projects across texts and languages, and the lack of granular markup for logical pieces of text (e.g., chapters and sections, proper names). The report pointed to a potential model that combines “massive scale with the flexibility for particular domains to manage data and provide services that suit their needs.”

In a 2008 paper, Rieger, addressing mass digitization projects, examined the “issues that influence the availability and usability, over time, of the digital books that these projects create,” and recommended a balance of preservation and access requirements as well as collaboration amongst cultural institutions. The concept of leveraging collaboration for cost savings in the development of repositories was mentioned by Furlough as he surveyed the repository landscape from a user-services perspective: “If content management and delivery services have a limited audience on a given campus, it may be better to partner with others.”

The implications of a library-owned aggregation of mass digitized materials for managing print and digital collections at the local institution level have emerged as a research theme in recent years. Sandler, in a thoughtful article, considered “a world where a single digital copy of an article or book can be delivered to multiple users, anytime, anywhere” and speculated that core resources could be “served up centrally,” saving costs to individual libraries and enabling them to focus on needs specific to their institutions and user communities. In the conclusion of a 2010 report, Henry pointed to a new collaborative cloud library model for collection development and management in which multiple libraries share the costs of maintaining both print books and their digital surrogates. A recent project led by Malpas of OCLC Research and funded in part by a grant from the Andrew W. Mellon Foundation explored the proposition that outsourcing management of portions of monographic print collections because of replication in both shared digital and shared print storage may be cost effective for libraries.

Context

The volume of books digitized from library collections has grown rapidly during the last decade. Output from small-scale, in-house library scanning operations was dwarfed when Google initiated its project to digitize books from libraries, first announced in December 2004. Google’s stated goal for the Google Books Library Project is to “make it easier for people to find relevant books—specifically, books they wouldn’t find any other way such as those that are out of print—while carefully respecting authors’ and publishers’ copyrights. Our ultimate goal is to work with publishers and libraries to create a comprehensive, searchable, virtual card catalog of all books in all languages that helps users discover new books and publishers discover new readers.”

The Google Books Library Project was followed by the Open Content Alliance (OCA), a coalition of libraries, nonprofit organizations, and corporations formed in October 2005 with the goal of digitizing public domain works. While a member of the OCA, and via its Live Search Books program, Microsoft funded the digitization of more than 750,000 books from libraries from December 2006 to May 2008 via its Live Search Books program. By early 2008, the number of books digitized under the auspices of these programs began to approach many millions across the participating libraries.

With libraries facing enormous economic pressures and with Google’s projects to digitize libraries’ collections continuing to increase the amount of digitized content, a number of research libraries joined together to address these issues. In October 2008, HathiTrust was launched as a collaborative effort by the Committee on Institutional Cooperation (CIC)—then a consortium of thirteen universities, two of which (Michigan and Wisconsin) were already
Google Library partners—and the University of California Libraries—to create a shared repository of digital collections. The University of Virginia became a participant in January 2009, with many other libraries joining since then. These partners have joined with the common understanding that the massive scale of library digitization enterprises, along with the high costs of digital preservation, demand a web-scale collaborative solution for ensuring long-term access to the digital output and a new vision for a collective collection. Because of the size of the HathiTrust repository and the depth of the collaboration involved, the participating libraries are uniquely positioned to leverage technical infrastructure and collective expertise for digital preservation, services, and collection management on an unprecedented scale. The presence of a critical mass of research institutions in the HathiTrust partnership enables an aggregation of digital resources not seen before, hosted by libraries for the long term in a continuation of their traditional role as stewards of the scholarly record and supporters of research and other scholarly pursuits.

What is HathiTrust?

At the heart of HathiTrust is a shared secure digital repository owned and operated by a partnership of major research libraries. The repository is best known as a means of preserving digital materials created via large-scale digitization projects. By pooling their collective resources and expertise, the partners have created a robust and scalable infrastructure to efficiently store, manage, and preserve their collections of digital books and journals in common. HathiTrust, however, has positioned itself as more than a simple aggregation of digitized library material. Its stated mission is much broader: "to contribute to the public good by collecting, organizing, preserving, communicating, and sharing the record of human knowledge."

To that end, the HathiTrust repository now likely contains the largest collection of digital volumes outside of Google Books. Because most of the U.S.-based Google library partners are members, the collections of the current HathiTrust members can be estimated to constitute a majority of all of the content contributed by U.S. libraries to Google Books. The partnership is open to institutions internationally. The first partner from outside the United States is the Universidad Complutense de Madrid, also a Google Books library partner. As HathiTrust adds members, the repository also will encompass a growing number of the volumes digitized from U.S. libraries by Microsoft under the auspices of the now defunct Microsoft Live Search Books service. In addition, the repository now contains tens of thousands of volumes digitized by the Internet Archive and additional volumes digitized by the partners themselves.

Growth has been rapid, and the repository (as of this writing) holds more than 8 million volumes, including 2 million public domain volumes. As the combined output of mass digitization accumulates, the sheer number of digital volumes aggregated in the repository will foster the partner libraries' collective ability to leverage a digital version of library collections assembled and curated by generations of librarians across the nation's research libraries. In addition to including the digital volumes derived from print, plans are underway to include other types of digital publications within the repository. For example, HathiTrust is in discussions with university presses about putting new books and book backlists online via open access and plans to extend that model. Currently, hundreds of current university-press titles are available online with open access permissions. The partners intend that the repository eventually will encompass materials beyond books and journals. Because new content types will demand new access, management, and preservation requirements, much remains to be resolved.

Goals and Values

The name HathiTrust was chosen to express the fundamental values of the organization. Hathi (pronounced hah-tee) is the Hindi word for elephant, an animal noted for its memory, wisdom, and strength. While HathiTrust's intent is to build a reliable and increasingly comprehensive digital archive of library materials converted from print that is co-owned and managed by a number of research institutions, the enterprise has a number of other important goals:

- To dramatically improve access to these materials in ways that, first and foremost, meet the needs of the co-owning institutions.
- To help preserve these important human records by creating reliable and accessible electronic representations.
- To stimulate efforts to coordinate shared collection management strategies among libraries, thus reducing long-term capital and operating costs of libraries associated with the storage and care of print collections.
- To create and sustain this "public good" in a way that mitigates the problem of free riders.
- To create a technical framework that is simultaneously responsive to members through the centralized creation of functionality and sufficiently open to the creation of tools and services not created by the central organization.

HathiTrust differs from Google and from organizations such as the Internet Archive in a number of ways. Structurally, HathiTrust is not a corporation or even a nonprofit organization nor is it a "trust" in the legal sense of the word. The
partnership is a collaborative enterprise of research libraries that depends on funding and in-kind contributions from members. As “an enterprise principally driven by a scholarly mission,” HathiTrust is committed to the principle that “creating a digital research library for the research community is the responsibility of research libraries.” In accordance with its research mission, HathiTrust embraces values long held by libraries, such as preservation, quality, privacy, and public access, and formally commits to long-term digital preservation. Although Google and the Internet Archive both maintain and provide access to large amounts of data, as a matter of course, neither organization is formally committed to digital preservation of digitized books over time. HathiTrust also differs from national or regional projects such as the Joint Information Systems Committee (JISC) (www.jisc.ac.uk) in the United Kingdom or the Europeana (www.europeana.eu/portal) initiative of the European Union in that currently no government-supported mandate or national cultural institution supports its existence.

In keeping with a public access mission, HathiTrust has put mechanisms in place to support greater access to the works in the repository. Although HathiTrust must follow copyright law and restrict access to volumes that are not in the public domain, the organization’s philosophy is to open up materials to the greatest extent legally permissible. Most of the digital volumes within the repository are the result of Google’s digitization, but HathiTrust may assign a viewability status to the library copy that is different from that of the copy in Google Books. In general, HathiTrust takes a less conservative stance regarding providing full-view access to government documents. In addition, a growing number of HathiTrust institutions provide mechanisms for rights holders to release their works into full view within the HathiTrust.

HathiTrust partner libraries also are actively working to move orphan works (works that can be assumed to be in-copyright but whose copyright owner cannot be located) into the public domain as another route to greater access. As of this writing, 6 million volumes within the repository are considered in-copyright or potentially in-copyright orphan works and are not viewable, but Lovoi and Dempsey note that many of those fall into the orphan works category and actually may be in the public domain. By collaborating within HathiTrust, research libraries plan to begin tackling this problem collectively. With support from an Institute of Museum and Library Services (IMLS) National Leadership grant, the University of Michigan has developed a Copyright Review Management System (CRMS), the expansion of which is currently being piloted by several HathiTrust partner libraries under the aegis of the IMLS grant. This system will be a means to scale and propagate book-by-book copyright determination by human beings, a process that can be arduous and complex. The intent is to expand use of the CRMS for copyright review activities across research libraries. In the meantime, Michigan is making progress, having used the CRMS to analyze more than 123,000 books (as of this writing) and moved approximately 54 percent of them into the public domain. HathiTrust makes these rights determinations available as part of a set of downloadable metadata called the “Hathifiles.”

Any discussion of copyright and digitized books invariably leads to the Google Books Settlement Agreement. The October 2008 agreement between the Authors Guild, the Association of American Publishers, and Google settled Authors Guild et al. v. Google, a class-action lawsuit alleging that Google's digitization and indexing of in-copyright works constitutes copyright infringement. In November 2009, an amended version was filed in response to a Department of Justice brief suggesting that the original version violated antitrust laws. The amended settlement is complex and has engendered discussion much broader than the scope of this paper, and, of this writing, the presiding judge has yet to rule. The aims of HathiTrust predate and are independent of the settlement and the amended settlement. However, HathiTrust could be affected by some of the provisions of the amended settlement, including those that would allow Google to sell institutional subscriptions to libraries for full view of books within Google, provide for libraries to host a research corpus of books, and prescribe the establishment of a “book rights registry.” In a December 2008 interview, John Wilkin, HathiTrust Executive Director, addressed some of the more positive potential effects:

Much of what HathiTrust proposes to do—preserve content, support access by print-disabled users, generate print replacement copies from the digital files when original print copies are damaged or lost, and serve as a body of content for large-scale computational needs—explicitly sanctions the settlement agreement, thus protecting this fundamental library-based effort from legal threats.

HathiTrust service development will need to take the settlement outcomes into account, respecting mandated constraints where they exist. If the amended settlement is approved, HathiTrust may leverage services such as the institutional subscription within the access services it offers where appropriate and considered valuable to the partners.

Collaboration

Owning and managing the repository is of inherent benefit to the participating libraries, and such an enterprise demands a thoughtfully structured collaborative infrastructure that accounts for the interests of all partners. In addition to cost savings for digital preservation and services resulting from economy of scale, a key benefit of collaboration is the ability