LIBRARY ISSUES

BRIEFINGS FOR FACULTY AND ADMINISTRATORS

Vol. 32, No. 3 **January 2012**

HathiTrust: The Elephant in the Library

By Jeremy York

Since its launch in 2008, the HathiTrust partnership has dramatically changed the outlook for academic and research libraries in the 21st century. It promises to become a truly transformative force in the way libraries function and serve their users on a global scale. With nearly 10 million volumes digitized from partner libraries securely stored in a partner-supported, community-certified digital preservation repository, HathiTrust has quickly become home to one of the largest research library collections in the world. Sophisticated search and discovery mechanisms and forward-looking initiatives in areas of copyright, computational research, and community-wide collection development show early signs of what academic and research libraries can accomplish together to support scholarship in a changing and increasingly demanding educational environment. The advances libraries are making are critical for faculty and administrators to be aware of as they prepare their students, universities and colleges to become leaders in what promises to be a complex and challenging future.

Background

While HathiTrust itself is quite young, it is a direct outgrowth of long-standing efforts in libraries to aggregate, preserve, and provide access to our collected knowledge. Before Google considered digitizing and indexing the world's knowledge, leaders in research libraries were contemplating the possibilities of digitizing and cooperatively preserving materials from their library collections on a large scale. They understood the possibilities that such digitization and collaboration could have for making their collections more readily available to users, and for facilitating more efficient management of the original printed materials.

Mechanisms for large-scale digitization were not available in the early 2000s when these ideas were being formulated. Ideas turned into plans in 2004, however, when libraries began to collaborate with Google and others to digitize millions of volumes from their print collections. HathiTrust arose most directly out of a collaborative agreement among institutions of the Committee on Institutional Cooperation to build a shared digital repository to preserve and provide access to volumes scanned in their large-scale digitization programs. When the University of California system joined the initiative, it was expanded and rebranded to become HathiTrust. The partnership currently includes more than 60 academic and research libraries, with primary leadership from the University of Michigan, Indiana University, the rest of the CIC libraries, and the University of California.

Goals and Objectives

The mission of HathiTrust is "To contribute to the common good by collecting, organizing, preserving, and sharing the record of human knowledge." This mission recognizes explicitly HathiTrust's role as a public good to the community, as a curator and preserver of our collected knowledge, and as a proactive disseminator and communicator of that knowledge in the broader community.

Coming Soon —

The Library as a Recruitment Tool

The intentions of the partnership are as expansive as the mission indicates, and are not limited to specific formats such as digital or print. The founding partners set a course they believed would facilitate deep engagement by partners with fundamental challenges facing the library community today, and from which could be derived deep, shared, and lasting benefits. To focus the direction of HathiTrust in its early years, the partners articulated a set of six initial goals:

- 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 academic institutions.
- 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 redoubled efforts to coordinate shared storage 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 ensures the greatest availability of materials to the public, but offers additional value to members, mitigating the problem of "free riders".
- To create a technical framework for the repository that provides significant centralized functionality, such as full-text search, but also supports distributed development of tools and services to allow partners and non-partners to meet specific access needs of their user communities.

"Comprehensive," "co-owned," "preservation and access," "digital and print collections," "value to members," and "openness," are watchwords and concepts that have become hallmarks of the initiative in its first years, reflecting the accomplishments the partnership has made, and progress toward its initial goals.

Progress to Date

Comprehensive. HathiTrust aims to assemble a comprehensive representation of the cultural record in digital form, beginning first with published materials. In December 2011, HathiTrust held more than 9.9 million volumes, putting it on a par with the largest research libraries in the world. As the content derives from many of these libraries, the quality and depth of the collections are on a par as well.

Google-digitized volumes make up the largest portion of content in HathiTrust, but do not by far represent the scope of content digitized or produced digitally by libraries or publishers. Supporting submission of content from a variety of sources was thus an early and high priority for the partnership.

Successes with content digitized by the Internet Archive and Microsoft in its first two years, and experimentation with locally digitized materials from several partner institutions, led to the development of a framework that has enabled HathiTrust to accommodate deposits of content from large vendor or smaller in-house operations at scale. Submission of content in the latter category, frequently comprising institutions' more rare and special collections material, has increased significantly in the year since the framework has been available.

HathiTrust is also working with several university presses on the deposit of their back file publications, offering free archiving services in exchange for open access to the works in HathiTrust. By the end of fall 2011, content digitized through non-Google sources made up more than a quarter of a million volumes in HathiTrust, or 10 percent of those publicly available.

Looking beyond digitized book and journal content, HathiTrust is engaged in pilot projects to support submission and preservation of digital audio materials and digital images such as maps and photographs. By 2014, HathiTrust also plans to support publication of digitally created (as opposed to retrospectively digitized) open access journals directly from the repository. These initiatives round out HathiTrust's efforts to pursue a comprehensive archive of materials.

Co-owned. When formulating the mission for HathiTrust, the founding partners knew that the expansive effort could only succeed and reach its full potential with the deep collaboration of co-supporting and co-owning libraries from around the world. They also believed that the initiative, with its initial focus on preservation and accessibility of the published record, would attract supporters on its own merits, particularly if it were able to demonstrate cost efficiencies for member libraries.

To give HathiTrust a start, the founding partners established a preliminary governance system composed of an Executive Committee and a Strategic Advisory Board. The Executive Committee agreed to charter HathiTrust for an initial five-year period, with a formal review of governance and sustainability to be conducted in the third year. The Executive Committee planned for a "constitutional convention" of partners to be held in the third year to coincide with this formal review. Such a convention would allow new institutions and consortia (those that were members of HathiTrust by October 31, 2010) to participate in establishing the governance model for HathiTrust beyond the first five-year period, and set future directions for the partnership.

The HathiTrust Constitutional Convention took place in October 2011 with the participation of 52 institutions – more than double the number of institutions that launched the initiative – including both large research and smaller academic institutions. Representatives of 8 other institutions that joined HathiTrust in 2011 were also present. The outcomes of the convention are available on the HathiTrust website. (www.hathitrust.org)

Preservation and Access. The development path of HathiTrust's repository underscores its fundamental purpose to ensure the preservation of deposited materials, and the belief of partners that preservation without access is of no value. The dual goals of preserving items, and providing as much access as legally possible, has guided every aspect of the repository's design and development.

The repository infrastructure is built on systems developed by the University of Michigan to ensure preservation and access for its own digitized book and journal collections. When HathiTrust was launched, these systems consisted of one active storage site in Michigan with tape backup, interfaces to view and build collections of volumes from the repository, and mechanisms to determine and store copyright information for each individual volume.

Within 3 months, with contributions from Indiana University, HathiTrust had established a fully redundant storage site in Indianapolis. In six months, a temporary bibliographic catalog was operational, and in just over a year from the time of its launch, full-text search was enabled over the entire corpus of material.

In April 2011, HathiTrust was certified by the Center for Research Libraries as a "Trustworthy Digital Repository" according to the Trustworthy Repository Audit and Certification: Criteria and Checklist (TRAC). Only two digital repositories currently bear this distinction. HathiTrust is the only one owned and operated by libraries.

Digital and Print Collections. In libraries' efforts to offer the finest educational resources to students and faculty, academic and research libraries, particularly in America, have amassed general collections of tremendous similarity. The collections are so similar that in June 2010, when HathiTrust held 6.1 million volumes, OCLC Research estimated a median overlap of 31 percent between HathiTrust and ARL institution collections. Overlap with smaller academic libraries, such as those in the Oberlin Group, was nearly 10 percentage points higher. Trends observed by OCLC would place the median overlap at closer to 50 percent today – a number that will only increase as HathiTrust continues to grow.

Libraries are keen in today's economic climate to reduce unnecessary redundancy and duplication in their collections. Because of the high degree of overlap in HathiTrust, the repository is poised to play a major role in decisions libraries make about storing and retaining portions of their printed collections. The scale of HathiTrust, and emerging initiatives by partner and non-partner institutions to coordinate print storage at regional and national levels, put libraries on a trajectory that may prefigure the way they operate and provide services in the near future.

What might the library landscape look like in 10 or 15 years, for example, if libraries as a community pooled resources to sustain a generalized, shared collection of print and digital materials? The impact such collaboration could have in reducing costs to institutions, freeing funding to pursue specialized collection and research interests, and improving services and access to collections, would be transformative. HathiTrust partners believe that taking broad, and even radical, steps toward this level of collaboration offers the best chance for libraries to ensure they remain relevant to their institutions and communities as primary sources of information. We must continue to do what we have always done, but we must do it in a new environment with new tools and a new vision for how we manage resources and deliver them to users.

Value to Members. In such a shared environment, what is the value to the institutions that support the effort? What prevents an institution that is not a member from piggybacking on the efforts of others and gaining the same ends? Any institution may, and certainly will, use HathiTrust as a resource for its students and staff, as a collection development tool, or in other ways it finds useful. The difference between members and nonmembers lies in the benefits of active participation, which come in forms of legal participation, and allocation of in-kind resources.

More than 70 percent of the Hathi Trust corpus is composed of in-copyright materials. While these are searchable in HathiTrust, they are not generally available for reading purposes. They may be available for reading, however, and other uses, under fair use and lawful reproduction provisions in U.S. copyright law and similar laws in other countries. In the United States, these provisions allow HathiTrust to offer access to in-copyright works

1. to users at partner institutions who have print disabilities (persons whose disabilities make it difficult to use printed books)
2. where the works fall under Section 108 provisions in copyright law and
3. where the works are determined to be copyright-orphaned. Orphan works are works that are in copyright, but for which the rights holder is unknown or cannot be contacted for permission to use the work.

As several recent lawsuits have made clear, including a suit brought by the Authors Guild and others against HathiTrust and five partner institutions, there is legal liability associated with these uses. Institutions must be willing to accept this liability to allow their users to make lawful uses of library materials. This is done through a formal agreement with HathiTrust. As champions of scholarly and educational uses of materials of all kinds, partner institutions have found this liability and the costs of partnership to be low in relation to the preservation benefits and the value gained from expanded access.

The second form of benefit to partners derives from the allocation of time and in-kind resources to HathiTrust initiatives. Partner institutions may participate in operational and strategic working groups and governing bodies that "do the work" of the partnership in planning and conducting day-to-day business. Partners may also devote resources to shared initiatives that, while providing benefits to the broad community, offer specific advantages for HathiTrust partners. Some examples of these include

- Review of volumes in HathiTrust for compliance with copyright formalities and identification of orphan works
- Creation of a database of partner print holdings
- Creation of a research center to support computational analysis of HathiTrust collections
- Development of a new system for managing HathiTrust bibliographic data.

- An effort to investigate and certify the quality of volumes in HathiTrust

Openness. HathiTrust partners have made significant progress in establishing a centralized infrastructure that is capable of ensuring long-term preservation of materials while offering a suite of access services. These services include bibliographic and full-text search across the entire repository and mechanisms for reading and browsing volumes, including from mobile devices. Centralized infrastructure allows these services to be offered efficiently and at scale.

One size does not fit all, however, and a key goal of HathiTrust has been to enable users, whether from partner institutions or not, to create custom views of content in the repository and use the content in services built outside the central infrastructure. HathiTrust's Collections application is a step towards fulfilling this first purpose, allowing users to create public or private collections of HathiTrust materials that can be searched independently of the larger corpus.

The ability to develop external services is supported through several mechanisms that allow partners and non-partners to obtain bibliographic information about volumes in HathiTrust, and a data API that allows programmatic retrieval of page images, pages of text, or entire volumes from the repository.

Being "open" in this way is an explicit goal of HathiTrust with respect to technology, but the partnership's commitment to openness extends through all facets of its operations and activities. This includes public posting of governance committee meeting minutes, detailed monthly reporting on activities and plans, use of open source technologies and contributions back to open source communities, participation in internal and external audits, and public access to materials to the greatest extent allowed by law and third-party agreements. HathiTrust strives for openness and transparency in all these areas as matters of service and accountability to its partners, and in the interest of providing public goods to the community.

Benefits to Scholars / Students

The greatest benefits HathiTrust provides to scholars and students will always be the permanence and accessibility of its collections. Users of HathiTrust can expect that the volumes they use will be right where they left them today and 50 or 100 years from now. Libraries are the entity best entrusted with this responsibility, and no other entity has a better track record in fulfilling it.

Special uses of materials are a second benefit. Apart from uses of in-copyright materials that are explicitly allowed by copyright law, libraries are particularly well placed to take the lead in making "fair uses" of in-copyright works, for instance providing access to users who have print disabilities and offering access to orphan works. Expanded access to materials is a significant benefit to users from HathiTrust partner institutions.

The ability of HathiTrust to tailor content and search interfaces to scholarly needs is an additional benefit. Tailoring may take the form of targeted content acquisition or the adjustment of search results ranking to better support research. HathiTrust is different from vended services that libraries subscribe to because the direction the partnership takes and the services it offers are under libraries' control and governed by libraries' interests in supporting teaching, education, and research at their institutions and in the broader community.

Ubiquitous discovery of materials in HathiTrust is the fourth benefit. HathiTrust is committed to developing its own discovery services and interfaces, and to disseminating the indexes and metadata behind these services to third parties to ensure the greatest discovery of its collections. HathiTrust records are harvested for discovery in OCLC's Worldcat, and the HathiTrust full-text index is being incorporated into services offered by Serials Solutions, EBSCO and other vendors. Discovery serves the immediate access needs of scholars and researchers, and is fundamental to HathiTrust's strategy to ensure that materials are valued and preserved over the long-term.

Conclusion

Today's environment is one of rapid change in higher education. Administrators are continually challenged to ensure their students and faculty have access to the best and most comprehensive resources available. Technological, economic, and social trends are changing the face and, in some cases, the very nature of research that our institutions perform. In the midst of these changes, libraries are responding with deep collaboration and technological innovation to ensure that our collections, values, and expertise remain central to the teaching and research activities at our institutions.

Doing what we do better, at lower cost, will give faculty and administrators greater flexibility to adapt to changing needs and research possibilities at their institutions, and ensure their institutions are competitive and more successful than ever in educating the leaders of the future.

Jeremy York is the Project Librarian for HathiTrust at the University of Michigan.

MOUNTAINSIDE
PUBLISHING, INC.

Library Issues: Briefings for Faculty and Administrators (ISSN 0734-3035) is published bimonthly beginning September 1980 by Mountainside Publishing Co., Inc., 321 S. Main St., #213, Ann Arbor, MI 48104; (734) 662-3925. **Library Issues**, Vol. 32, no.3 ©2012 by Mountainside Publishing Co., Inc. Subscriptions: $84/one year; $144/two years. Additional subscriptions to same address $26 each/year. Address all correspondence to **Library Issues, P.O. Box 8330, Ann Arbor, MI 48107**. (Fax: 734-662-4450; E-mail: sales@libraryissues.com) Subscribers have permission to photocopy articles free of charge for distribution on their own campus. Library Issues is available online with a password or IP access at <http://www.LibraryIssues.com>