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, unbeholden to 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.[1] 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

Heather Christenson (heather.christenson@ucop.edu) is Project Manager, Mass Digitization, California Digital Library, University of California, Oakland.

Submitted August 4, 2010; tentatively accepted September 3, 2010, pending modest revision; revision submitted November 2, 2010, and accepted for publication.

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."[2] In 2008, Brantley, 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."[3] Leetaru made a case that the output of mass digitization is "access digitization" rather than "preservation digitization."[4] 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.[5] 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.[6]

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."[7] 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."[8] 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.[9] 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."[10]

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.[11] 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.[12] 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.[13]

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."[14] 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.[15] 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.[16] 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)[17]—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.[18] 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."[19]

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.[20] 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.[21]

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."[22] 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 Lavoi and Dempsey note that many of those fall into the orphan works category and actually may be in the public domain.[23] 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.[24] 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."[25]

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.[26] 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—is explicitly sanctioned in the settlement agreement, thus protecting this fundamental library-based effort from legal threats.[27]

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

to tap into expertise across the libraries. HathiTrust has a shared governance structure, with an executive committee that is the decision-making body, along with a strategic advisory board composed of university librarians and associate university librarians from the partner institutions. The strategic advisory board sets functional objectives, convenes task forces to address specific issues, and recommends policies, drawing on the array of experience and expertise of the members.

Within the past year, HathiTrust launched working groups on a wide range of topics, including communications, collection development and management, quality, ingest and error rates, collaborative development, resource discovery, faculty research, and storage expansion needs. In addition to these formal groups, HathiTrust has brought together technical talent from the participating institutions to develop and improve its operational processes and aims for more collaborative development in the future. Both the organization and individual participants are gaining experience in long-term collaboration on core infrastructure and services. In a stringent economic climate in which libraries are increasingly seeking to collaborate, the growing pool of expertise gained by participants becomes a valuable asset.

Preservation

Secure and long-term digital preservation of volumes in the repository is fundamental to the goals of the enterprise. The HathiTrust repository is sometimes compared to the Portico (www.portico.org) and CLOCKSS (Controlled Lots of Copies Keep Stuff Safe) (www.clocks.org) digital preservation services, but it differs from them in terms of the provenance of included content, archival philosophy, and underlying business and organizational structure. Both Portico and CLOCKSS focus on journal and e-book content originating from publishers, while HathiTrust has begun with content from the libraries' mass digitization projects. Both Portico and CLOCKSS are dark archives that make content available only when a trigger event (such as a publisher ceasing operation) occurs. Although a large amount of content within the HathiTrust repository is not viewable to end users by copyright law, all other content is available and the repository is technically a light archive. Portico and CLOCKSS are services of nonprofit ventures and partner with publishers, while HathiTrust is an organization composed solely of libraries.

HathiTrust is committed to preserving the intellectual content and, if reasonably possible, the exact appearance and layout of materials digitized for deposit and is committed to allowing the partners to make open and meaningful decisions about formats and quality. For example, upon joining, a partner institution may determine which image file format they want to use for their deposited content, and the decision process of each partner may be documented and shared to inform the others. Individual partner institutions may have varying positions on whether the digital copies of print books created via mass process are preservation-worthy copies, but HathiTrust is seeking to conduct research in this area and develop quality metrics. With funding from the Andrew W. Mellon Foundation, Paul Conway of the University of Michigan School of Information, in conjunction with HathiTrust, is investigating means of measuring quality and usefulness of digital objects and the feasibility of establishing a mechanism for branding the trustworthiness of deposited volumes for particular uses, such as reading, printing volumes on demand, and performing computational research.[28] The goal of this certification process is to "give assurance that content within a repository is worthy of preservation, and increase the value of that content in broader discussions about storage and management solutions for both digital and print collections."[29]

The HathiTrust repository conforms to accepted standards and models for digital preservation, including the International Standards Organization's Open Archival Information System (ISO OAIS) reference model, the Metadata Encoding and Transmission Standard (METS), and the Preservation Metadata Implementation Strategies (PREMIS) Data Dictionary.[30] Digital objects are stored in formats that are documented, open, and standards-based with the intent of providing an effective means to migrate objects to successive preservation formats over time, as necessary. The repository utilizes robust technology and has the geographic redundancy of two mirror sites at the University of Michigan and Indiana University. In addition, each site has several layers of redundancy; a tape backup constitutes yet another copy. A cross-institutional working group reviewed the storage configuration, conducted a cost–benefit analysis regarding the need for more redundancy, and reported a "high level of confidence in the existing two-instance architecture."[31] The Center for Research Libraries is now reviewing HathiTrust for Trustworthy Repositories Audit and Certification (TRAC) compliance.[32] The TRAC review is an independent evaluation that gauges a repository's capability to reliably store, migrate, and provide access to collections, and it is sought after by preservation repositories as a community metric of confidence.

By virtue of its scale and its acceptance of varied content from many different sources, the repository is well suited for encountering and overcoming common challenges, specifically in areas of repository standards, best practices, and methods for certifying the quality of the deposited content. During the past year, the creation of new content-ingest streams has tested the original repository structure built by the University of Michigan, reinforcing some principles for homogeneity of file formats and metadata and

also identifying where the partners can make choices and where flexibility is required. A team of members from the University of Michigan and the California Digital Library (on behalf of the University of California (UC)) collectively tackled the creation of two new content-ingest streams: UC's Google-digitized volumes and UC's Internet Archive-digitized volumes. The group faced the technical challenges associated with allowing heterogeneity and ensuring the ability of the repository and its services to function. As each content stream was created, the team gave rigorous attention to choices about such elements as identifiers, image formats, individual files selected from digitization vendors' content packages, and specific tags and associated variables in the recording of the progression of transformative events upon ingest within preservation metadata. Effective management of objects in the repository must encompass digital preservation standards and uses within access services. Accommodating these dual purposes can present a technical challenge. For example, a choice may be made that PDF is not appropriate for preservation, although that format may be useful for end-user access. In that case, the object may be stored in a more preservation-appropriate format and, for access purposes, the PDF format may be derived from the preservation file format, requiring an extra process on the access end, a compromise that serves both purposes.

Discovery and Access

In addition to digital preservation and in concert with it, HathiTrust embraces access services as essential to its mission. The HathiTrust repository offers a number of end-user services, such as basic and advanced bibliographic search, full-text search based on extracted text, and a collection-builder tool (explained below). The bibliographic search uses an aggregation of records contributed by partner libraries and thus is based on rich descriptive metadata that is the output of decades of library cataloging. The bibliographic search is comprehensive across the full spectrum of the digital collection from in-copyright to public domain. Researchers have documented Google's metadata errors, primarily resulting from automated processes.[33] Since HathiTrust metadata originates from partner libraries, the libraries have a more direct opportunity to resolve errors, collectively explore how the original cataloging of print volumes can be enhanced and extended to digital volumes, and experiment with optimally integrating bibliographic metadata with full text for search purposes. The full-text search (also known as large-scale search) was built by developers at the University of Michigan, and further development is guided by the HathiTrust Discovery Interface working group. As of this writing, the repository is providing full-text search across more than 2.8 billion pages contained within 8 million volumes. A distinctive feature of this service is that libraries own both the search mechanism and the content on which it acts. This ownership is significant for several reasons. For end users, the selectiveness and ranking of search results are not influenced by commercial interests, and the material covered by the search is a known corpus of materials selected, cataloged, and curated by librarians with the interests of academic users in mind. For partner libraries, owning the full-text search and the content provides an opportunity to engineer end-user services that are configurable for scholarly uses as well as free from advertising, commercial bias, and censorship.

Once discovered, digital volumes within the repository are accessible by various means depending on copyright status. Google-digitized public domain volumes are available in a full PDF download to authenticated users from partner institutions; public domain volumes digitized via Internet Archive and locally by partners are available in full PDF to all. All public domain volumes can be viewed on the web in a page-turner application. Volumes that are in copyright are discoverable via large-scale search, and users may view a list of pages on which their search term appears (snippets are not yet available). Most books are treated as in-copyright, but may be moved to an open status upon human-reviewed copyright determination (e.g., through the CRMS or on request from the rights holder). HathiTrust also offers services to print-disabled users who are located at the University of Michigan and plans to extend the service to other partners.[34] Printed versions of public domain books from some partners are now offered via a link within the HathiTrust Interface to print-on-demand service.

The Collection Builder functionality allows librarians and individual end users to create and share specific themed collections regardless of whether the end user is affiliated with a partner institution. The Collection Builder has great potential for integration within local services, such as online courses and themed collection portals built by local institutions. Once a collection is created, the full text of those volumes can be searched as a set. One can envision other future scholarly tools that can capitalize on a scoped, curated group of volumes by being able to manipulate and analyze them in various ways.

In keeping with its mission to enable local institutions to develop tools and services, the HathiTrust offers freely available data, open to any institution, that can be captured and incorporated in a local service. The data also is machine-accessible so that local services can be built using it. For example, the University of California uses data to provide direct links to the full text of HathiTrust public domain volumes via UC-eLinks (www.cdlib.org/services/d2d/ucelinks), its local link resolution service. A growing number of partner libraries provide links to HathiTrust resources within their online public access catalogs.

Supporting Research

Also emerging is support for scholarly computational research. During the past year, a working group convened to develop specifications for a research center for scholarly use. This action was taken in anticipation of the pending Google settlement, which includes terms that sanction the use of in-copyright works owned by HathiTrust institutions in "non-consumptive" computational research. Non-consumptive research is understood to describe "analysis of a form that does not require (and does not permit) reading access to in-copyright materials."[35] The terms of the settlement also provide for the establishment of up to two research centers that would enable this research across the entire body of Google-scanned content. HathiTrust is proposing a center that will support research capabilities across the HathiTrust corpus, which it defines as "the complete set of works in HathiTrust, including Public Domain, Google Public Domain, Open Access, and In-copyright Data."[36] The report states,

> The founding institutions of HathiTrust undertook the effort of building a repository of published content with the expectation that this content in addition to serving the needs of traditional reading and research would serve as an extraordinary foundation for many forms of computing-intensive research, particularly in the areas of language and literature.[37]

The working group characterized research types that a HathiTrust research center would need to support, including aggregation and distillation of subsets of data, development of tools, mechanisms for collaboration, and ability to preprocess and add data. Using this collectively defined framework, HathiTrust has begun to investigate the Software Environment for the Advancement of Scholarly Research (SEASR) (http://seasr.org) as a means to provide computational access to materials stored in the repository. SEASR, funded by the Andrew W. Mellon Foundation, is a research and development environment devoted to supporting digital humanities initiatives and fostering collaboration in a virtual environment.

Collection Development and Management

Underlying these services is the HathiTrust collection. The numbers do not tell the whole story of the depth and breadth of the collection; however, numbers give a frame of reference and starting point. At the time of this writing, within the more than 8 million total volumes (2 million volumes in the public domain), 4.5 million book titles and nearly 200,000 serial titles are represented.[38] The current HathiTrust collection spans several centuries and hundreds of languages. The top ten languages (English, German, French, Russian, Chinese, Spanish, Japanese, Italian, Arabic, and Polish) account for approximately 86 percent of the content, and the next forty languages account for another 13 percent.[39] U.C. Berkeley University Librarian Thomas Leonard has commented that we can view the HathiTrust collection in the same way astronomers look far out into the universe; like the images of stars that are light years away and thus ancient, the further back we go into the collection, the more we see a snapshot of what research libraries were collecting at the time.[40]

An analysis performed by Malpas on the subject distribution of titles, based on subject headings within bibliographic metadata, revealed "Language, Linguistics, and Literature" and "History and Auxiliary Sciences" to be the most populous subjects, followed by "Business and Economics," "Philosophy and Religion," and "Art and Architecture."[41] The HathiTrust website provides visualizations of the collection categorized by Library of Congress classification, language, and publication date.[42] Analysis of bibliographic metadata is only beginning to explore the types of collection analysis that might be possible via the full text search and specialized tools. Having bibliographic metadata, digital content, and management metadata in a common repository under library ownership likely will foster the development of analysis tools to answer questions that cross the boundaries of the data and depend on the synergy of the aggregation. For example, what has been collectively digitized, and what format is it in? What is the array of conditions that create a true duplicate, how much duplication is present, and what de-duplication strategies make sense?

HathiTrust formed a collections committee that will explore what additional tools and services may be needed to characterize the collection as it evolves. These may include analytical tools that examine subject, language, date, format, or other characteristics; extensions of the Collection Builder tool; and mechanisms that would be useful to describe the corpus to a potential user.

In concert with those activities, the HathiTrust corpus can be used as a basis for the development of comprehensive or distinctive digital collections in particular areas that build on participant strengths. The collections committee will tackle those opportunities as well. For example, the partners could develop a shared approach to government documents that capitalizes on the CIC's focused U.S. government documents digitization initiative.[43] Gap analysis and collection building will likely lead the partners to explore opportunities for digitization and collaboration with other initiatives.

Print Curation

Leveraging the HathiTrust corpus to manage print

collections both within and beyond the partner libraries is an active area of exploration. Driven by economics and space constraints, momentum is building toward putting ideas about collective print curation into practice. The mechanics of how aspects of this might work are beginning to emerge through recent research.

Malpas' 2010 Cloud Library research project explored the proposition that outsourcing management of portions of monographic print collections, based on replication in both shared digital and shared print storage, may be cost effective for libraries.[44] The study revealed marked overlap between the HathiTrust monographic collection and the holdings of major shared print repositories across the country, and thus a large potential library clientele for outsourced service. The study also found that until in-copyright works can be distributed digitally, the tipping point for cost-effectiveness would likely not be reached for most libraries. In addition, the libraries of the CIC universities have undertaken federal government documents digitization with an eye toward examining the relationship between print and digital copies to better position themselves for coordinated decisions about print retention.[45] Although small steps, these two examples, along with a trend toward shared print storage initiatives evidenced in discussions at the April 2010 meeting of the Association of Research Libraries, can be seen as early indicators of what is to come and of the economic incentives and collaborative structures that may be needed.[46]

HathiTrust has recently developed a cost model for participation to include libraries that may wish to leverage the digital collection for print collection management and other purposes.[47] The initial participation model has been that institutions pay infrastructure costs for the digital content they contribute. The second, newer participation model is aimed at institutions that do not necessarily have large collections (or any) of digital content to contribute but want to participate in the curation and management of the repository in return for specialized services. By paying a membership fee, these partners will contribute to sustaining a common resource, share in uses of relevant materials, and have a voice in future directions of HathiTrust. The second model also addresses the problem of "free riders," avoiding a situation where some partners would have access to an amount of content out of proportion to the amount of their monetary contribution. The newer participation model is based on partners' print holdings, and costs are calculated on a number of precise elements about cost and the "sharedness of the content," including costs to maintain public domain content.

Dempsey has used HathiTrust as an example of how "web scale" activity is "managed at the network level," and its "audience is potentially all web users."[48] Although all members pay, the network level of the HathiTrust infrastructure enables the libraries to pool their resources and reach more users more effectively at lower costs and to effectively "transfer resource[s] away from 'infrastructure' and towards user engagement."[49] The general public also benefits from this arrangement through access to public domain resources and discovery services.

Next Steps

Technical challenges are perhaps the easiest for pioneering organizations to overcome. Much more difficult are the challenges of achieving collaboration and political harmony, agreeing on policy, and implementing and building a new organizational culture within a group of geographically dispersed institutions with independent governance structures. In light of these challenges, HathiTrust has a plan for the next steps of its evolution. Following a formal review of the repository by partners, HathiTrust will convene its first Constitutional Convention in 2011. At this convention, the partners will have an opportunity to enhance, revise, or reenvision governance, partnership, and cost models.

Looking toward the future, the membership will need to continue to think boldly. Could the HathiTrust's mantle of stewardship and the values it embraces enable it to evolve into a broader role as a de facto national research library? Might even commercial agents (such as Google) come to view HathiTrust as a solution to the problem of long-term digital preservation? Even if these entities do realize that HathiTrust can fulfill the need for digital preservation, how do the public good aspects of HathiTrust's mission intersect with the commercial interests of for-profit enterprises? What new partnerships can be formed to advance the scholarly agenda of the HathiTrust partners? When research libraries collectively hold digital copies of significant portions of their collections, how comfortable will they be with collectively pushing the boundaries of legal use of the digital copies, and how effectively can they advocate for copyright reform?

Conclusion

In naming the founding of HathiTrust as one of *Library Journal*'s top academic library stories of 2008, Albanese described it as "the library community's most ambitious digital collaboration ever."[50] Two years later, the HathiTrust partners are making progress on issues such as cost-effective digital preservation of very large collections of digital volumes, access mechanisms for such a collection, including openly available metadata, and support for computational research. HathiTrust represents a growing digital aggregation of research library content at a scale with the potential to support collection management decisions as research libraries face financial pressures and weigh the relative value

of print and digital volumes. The widespread collaboration, aggregated expertise, and pooled digital collections of HathiTrust seem to be resulting in beneficial progress for both the library community and end users.

References and Notes

1. Jeffrey Toobin, "Google's Moon Shot: The Quest for the Universal Library," *New Yorker*, Feb. 5, 2007, www.newyorker.com/reporting/2007/02/05/070205fa_fact_toobin (accessed Oct. 8, 2010); Sergey Brin, "A Library to Last Forever," *New York Times*, Oct. 9, 2009, www.nytimes.com/2009/10/09/opinion/09brin.html (accessed Oct. 31, 2010).
2. Trudi Bellardo Hahn, "Mass Digitization: Implications for Preserving the Scholarly Record," *Library Resources & Technical Services* 52, no. 1 (Jan. 2008): 18, 24.
3. Peter Brantley, "Book Search Will Not Work Like Web Search," online posting, Jan. 1, 2008, Peter Brantley's Thoughts and Speculations, http://blogs.lib.berkeley.edu/shimenawa.php/2008/01/02/trade_for_our_own_account (accessed Oct. 1, 2010).
4. Kalev Leetaru, "Mass Digitization: The Deeper Story of Google Books and the Open Content Alliance," *First Monday* 13, no. 10 (Oct. 6, 2008), www.uic.edu/htbin/cgiwrap/bin/ojs/index.php/fm/article/viewArticle/2101/2037 (accessed Oct. 31, 2010).
5. Ibid.
6. William C. Dougherty, "The Google Books Project: Will It Make Libraries Obsolete?" *Journal of Academic Librarianship* 36, no. 1 (Jan. 2010): 86–89.
7. Council on Library and Information Resources, *Many More than a Million: Building the Digital Environment for the Age of Abundance. Report of a One-Day Seminar on Promoting Digital Scholarship Sponsored by the Council on Library and Information Resources (Nov. 28, 2008): Final Report* (Mar. 1, 2008), www.clir.org/activities/digitalscholar/nov28final.pdf (accessed Oct. 1, 2010).
8. Ibid.
9. Oya Y. Rieger, *Preservation in the Age of Large-Scale Digitization: A White Paper*, CLIR Publication 141 (Washington, D.C.: Council on Library and Information Resources, 2008): vi, www.clir.org/pubs/reports/pub141/pub141.pdf (accessed Nov. 11, 2010).
10. Mike Furlough, "What We Talk About When We Talk About Repositories," *Reference & Users Services Quarterly* 49, no. 1 (2009): 22.
11. Mark Sandler, "Collection Development in the ~~Age~~ Day of Google," *Library Resources & Technical Services* 50, no. 4 (Nov. 30, 2005): 241.
12. Charles Henry, "Epilogue," *The Idea of Order: Transforming Research Collections for 21st Century Scholarship*, CLIR Publication 147 (Washington D.C.: Council on Library and Information Resources, 2010): 121–23, www.clir.org/pubs/reports/pub147/pub147.pdf (accessed Sept. 2, 2010).
13. Constance Malpas, *Cloud-Sourcing Research Collections: Managing Print in the Mass-Digitized Library Environment* (Dublin, Ohio: OCLC Research, 2011), www.oclc.org/research/publication/library/2011/2011-01.pdf (accessed Jan. 19, 2011).
14. Google Books, Google Books Library Project—An Enhanced Card Catalog of the World's Books, books.google.com/googlebooks/library.html (accessed Sept. 3, 2010).
15. Microsoft, "Microsoft Live Search Fact Sheet," www.microsoft.com/presspass/newsroom/factsheet/LiveSearchFS.mspx (accessed Oct. 8, 2010).
16. Open Content Alliance, Global Consortium Forms Open Content Alliance to Bring Additional Content Online and Make It Searchable, web.archive.org/web/20051007010920/http://www.opencontentalliance.org/OCARelease.pdf (accessed Oct. 28, 2010).
17. Committee on Institutional Collaboration, About CIC, www.cic.net/Home/AboutCIC.aspx (accessed Oct. 31, 2010).
18. HathiTrust, Major Library Partners Launch HathiTrust Shared Digital Repository, www.hathitrust.org/press_10-13-2008 (accessed June 30, 2010).
19. HathiTrust, Mission and Goals, www.hathitrust.org/mission_goals (accessed Aug. 1, 2010).
20. HathiTrust, Currently Digitized, www.hathitrust.org/ (accessed Feb. 10, 2011).
21. HathiTrust, Mission and Goals, www.hathitrust.org/mission_goals (accessed Aug. 1, 2010).
22. Roger C. Schonfeld, "Conclusion," in *The Idea of Order: Transforming Research Collections for 21st Century Scholarship*, CLIR Publication no. 147 (Washington D.C.: Council on Library and Information Resources, 2010): 116–20, www.clir.org/pubs/reports/pub147/pub147.pdf (accessed Sept. 2, 2010); HathiTrust, "FAQ", www.hathitrust.org/faq (accessed July 21, 2010).
23. Brian Lavoie and Lorcan Dempsey, "Beyond 1923: Characteristics of Potentially In-Copyright Print Books in Library Collections," *D-Lib Magazine* 15, no. 2/1 (Nov./Dec. 2009), www.dlib.org/dlib/november09/lavoie/11lavoie.html (accessed Oct. 31,2010).
24. University of Michigan Library, Copyright Review Management System—IMLS National Leadership Grant, Copyright Review Management System, www.lib.umich.edu/imls-national-leadership-grant-crms (accessed Aug. 30, 2010).
25. HathiTrust, Hathifiles Metadata, www.hathitrust.org/hathifiles_metadata (accessed Aug. 3, 2010).
26. Jonathan Band, *A Guide for the Perplexed Part III: The Amended Settlement Agreement* (Washington, D.C.: ALA, Association of College and Research Libraries, and Association of Research Libraries, 2009), www.arl.org/bm~doc/guide_for_the_perplexed_part3_final.pdf (accessed Oct. 29, 2010).
27. John Wilkin, "BackTalk: HathiTrust and the Google Deal," *Library Journal* (Dec. 23, 2008), www.libraryjournal.com/article/CA6624782.html (accessed Oct. 31, 2010).
28. University of Michigan School of Information, "Mellon Grant Aids Research Criteria for Digital Libraries," online posting, Sept. 28, 2009, SI Informant, blog.si.umich.edu/2009/09/28/mellon-grant-aids-researching-criteria-for-digital-libraries (accessed Aug. 30, 2010).
29. HathiTrust, Update on October 2009 Activities, www

.hathitrust.org/updates_october2009 (accessed July 30, 2010).
30. ISO Archiving Standards—Reference Model Papers, http://nssdc.gsfc.nasa.gov/nost/isoas/ref_model.html (accessed Oct. 31, 2010); Library of Congress, Metadata Encoding and Transmission Standard, www.loc.gov/standards/mets (accessed Oct. 28, 2010); Library of Congress, PREMIS Data Dictionary for Preservation Metadata Version 2.0, www.loc.gov/standards/premis, (accessed Oct. 28, 2010).
31. HathiTrust, Recommendations Regarding a Third HathiTrust Instance, www.hathitrust.org/documents/hathitrust-3rd-instance-recommendations.pdf (accessed July 30, 2010).
32. OCLC and Center for Research Libraries, *Trustworthy Repositories Audit & Certification: Criteria and Checklist Version 1* (Chicago: CRL; Dublin, Ohio: OCLC, 2007), www.crl.edu/sites/default/files/attachments/pages/trac_0.pdf (accessed Aug. 1, 2010).
33. See, for example, Geoffrey Nunberg, "Google Books: A Metadata Train Wreck," online posting, Aug. 29, 2009, Language Log, languagelog.ldc.upenn.edu/nll/?p=1701 (accessed July 29, 2010).
34. Paul Courant, "Digitization and Accessibility," online posting, Nov. 2, 2009, Au Courant, paulcourant.net/2009/11/02/digitization-and-accessibility (accessed Aug. 1, 2010).
35. HathiTrust, Call for Proposal to Develop a HathiTrust Research Center, www.hathitrust.org/documents/hathitrust-research-center-rfp.pdf (accessed July 30, 2010).
36. Ibid.
37. Ibid.
38. HathiTrust, Statistics Information, www.hathitrust.org/statistics_info (accessed Feb. 10, 2011).
39. HathiTrust, HathiTrust Languages, www.hathitrust.org/visualizations_languages (accessed Aug. 31, 2010).
40. Thomas Leonard, "From Print to Digital: Visions of 21st Century Collections" (presentation, Pacific Rim Digital Alliance 2010 Meeting, Shanghai, China, October 21–22, 2010).
41. Constance Malpas, OCLC Research, "Subject Distribution of Titles in the Hathi Repository June 2010" (slideshow presentation at the American Library Association Annual Conference, Washington, D.C., June 2010), www.slideshare.net/oclcr/june-2010-subject-snapshots (accessed July 30, 2010).
42. HathiTrust, Visualizations, www.hathitrust.org/visualizations_callnumbers (accessed Aug. 1, 2010).
43. Committee on Institutional Collaboration, "CIC–Google Government Documents Project," www.cic.net/Home/Projects/Library/BookSearch/Govdocs.aspx (accessed Aug. 1, 2010).
44. Malpas, *Cloud-Sourcing Research Collections.*
45. Committee on Institutional Collaboration, "CIC-Google Government Documents Project."
46. Lizanne Payne, "Models for Shared Print Archives: WEST and CRL," (slideshow presentation at the 156th Association of Research Libraries Membership Meeting, Seattle, Washington, Apr. 28–30, 2010), www.arl.org/bm~doc/mm10sp-payne.pdf (accessed Nov. 11, 2010).
47. John Wilkin, memo, Feb. 12, 2010, www.hathitrust.org/documents/hathitrust-cost-rationale-2013.pdf (accessed Oct. 31, 2010).
48. Lorcan Dempsey, "Sourcing and Scaling," online posting, Feb. 21, 2010, Lorcan Dempsey's Weblog, orweblog.oclc.org/archives/002058.html (accessed July 29, 2010).
49. Ibid.
50. Andrew Albanese, "The Library Journal Academic Newswire Year in Review, the Top Academic Library Stories of 2008," *Library Journal* (Jan. 7, 2009), www.libraryjournal.com/article/CA6626579.html?nid=3603 (accessed Oct. 31, 2010).