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. 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."  

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. 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." The Center for Research Libraries is now reviewing HathiTrust for Trustworthy Repositories Audit and Certification (TRAC) compliance. 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-digitzed 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. 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 are available in full PDF to all. All public domain volumes are freely viewable 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 in response to 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. 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.ucbib.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.”26 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.”26 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.27

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.28 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.29 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.30

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.”31 The HathiTrust website provides visualizations of the collection categorized by Library of Congress classification, language, and publication date.32 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.33 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 work might be emerging 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. 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. 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.

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. 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." 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." 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 re-envision 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 2006, Albanese described it as "the library community's most ambitious digital collaboration ever." 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 and 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


5. Ibid.


8. Ibid.


29. Hathitrust, Update on October 2009 Activities, www


36. Ibid.

37. Ibid.


44. Malpas, Cloud-Sourcing Research Collections.

45. Committee on Institutional Collaboration, “CIC-Go2ve Government Documents Project.”


49. Ibid.