This Library Never Forgets: Preservation, Cooperation, and the Making of HathiTrust Digital Library

Jeremy York; University of Michigan; Ann Arbor, MI/USA

Abstract

HathiTrust Digital Library was launched in October 2008 as a joint undertaking by 25 research libraries to preserve and provide access to millions of volumes of their digitized holdings. Drawing on the collective experience of the founding members, which include the schools of the Committee on Institutional Cooperation (the Big Ten and the University of Chicago), the University of California system, and the University of Virginia, a robust and scalable infrastructure was created to house, manage, and provide access to the collections. This paper traces the development of the initiative from its origins, describing the challenges it has faced, its strategies for addressing them, and its vision and hopes for the future.

Introduction

In October 2008, HathiTrust, the largest digital preservation initiative in the history of libraries, was launched. Led by the University of Michigan and Indiana University, libraries of the Committee on Institutional Cooperation (including the schools of the Big Ten and the University of Chicago), the University of California system, and the University of Virginia came together in an unprecedented collaborative undertaking to digitally preserve and provide access to millions of volumes of their collective holdings.

Using the repository developed by the University of Michigan to house Google- and locally-scanned content as a model, the partners constructed a robust and scalable infrastructure for preserving digital content that is based on industry standards for preservation and long-standing models for Open Archival Information Systems (OAIS). The repository is designed to accommodate vast amounts of digital information in a framework that emerges from, and is infused with the traditional commitments of libraries to longevity, openness, and access within the bounds of law.

Although the standards and infrastructure on which the repository is based are not new to libraries, the scale of the initiative, both technically and operationally (in terms of governance and inter-institutional collaboration) has given rise to a number of issues and challenges that are. These include policies in areas such as governance, organizational financing, and inter-institutional authentication, and technical challenges such as ingest and storage of materials from multiple sources, full-text search, and integration of print-on-demand and local-access services, including services for users with print disabilities.

Many of these issues are important not only to the success of HathiTrust, but apply more broadly to the continued relevance and usefulness of libraries in the digital age. The effort is premised on the beliefs that the whole of the library community is greater than the sum of its parts, and that libraries can remain a valuable and indeed essential part of the global infosphere only if we take active responsibility for moving our content, services and values into the new information ecosystem. This paper traces the development of HathiTrust from its origins, the issues and challenges it has encountered, and specific strategies it has and is employing in its efforts to create a library that is truly universal: in its content, its contributors, and its constituents.

Origins and Goals

The fundamental purpose behind the establishment of HathiTrust is preservation. HathiTrust is a direct response to a need among libraries engaged in large-scale digitization to ensure the preservation of their content over the long term. Preservation of digital materials has been a concern for nearly two decades in the library community, but awareness of the challenges involved and actions to address them have vastly accelerated in the last several years. This is due first of all to a realization that born-digital materials and digital collections that were created in the early days of digital libraries would be lost if actions were not taken to maintain them, second to an acceptance of digital reformatting as a valid means of preserving analog collections, and third to a vast increase in the quantity of information that was being produced digitally or reformatted to be digital.

This last was the major concern for institutions preparing to engage in partnerships with Google to digitize large portions of their print collections in the mid 2000s. The University of Michigan was the first to sign an agreement with Google in 2004 and, recognizing the possibilities it would create for the future, inserted a statement in the agreement allowing the University to share the digital copy it received from Google with other libraries [1]. Two and a half years later, this statement became the basis for a shared digital repository formed by the schools of the Committee on Institutional Cooperation (CIC) to preserve content digitized through their collective partnership with Google [2]. A year later, in 2008, the University of California system and the University of Virginia joined the initiative, adding considerable holdings and expertise in digital libraries to the partnership. The shared repository was expanded and re-branded in October 2008 to become the HathiTrust Digital Library.

Because there was no model for the deep inter-institutional cooperation needed to produce and support a digital preservation infrastructure of this magnitude, the initiative has to a large degree been guided by the lessons and experiences of its founding members. The digital library program at University of Michigan brought a wealth of experience to the partnership in the creation and maintenance of large-scale production environments for digital materials. The repository originally created for its own large-scale program was expanded and replicated several years later to serve as the repository for the HathiTrust.

The CIC has a long history of successful cooperation between the schools of the Big Ten and the University of Chicago. The voluntary nature of the consortium, and its view that building the strengths of individual institutions will increase the value of the resources available to all, are important components of HathiTrust's governance and partnership orientation.

The challenges the University of California has faced in coordinating its collections (both print and digital) among many distributed campuses have translated into numerous benefits for the library community, including HathiTrust. Reducing costs and redundancy through shared storage, cataloging, and collective development of library resources are key goals of the HathiTrust partnership. The development and implementation of community standards for sharing digital resources, in which UC has been instrumental, are fundamental ways of achieving them.

The digital library program at the University of Virginia, with the University of Michigan, is one of the oldest in the country. From the beginning it has distinguished itself in its attention to the needs of faculty and students at the University, particularly in the area of digital humanities. The University of Virginia's experience in this area has been a driving force behind efforts in HathiTrust to develop specific tools and services to serve the research community

It will be noted that the experiences of HathiTrust's founding partners go far beyond the realm of preservation. It is no mistake that they have become core components of what HathiTrust hopes to accomplish. Although the primary purpose of the repository is preservation, the partners were in agreement that preservation without access is of no value. HathiTrust was therefore conceived as a fully functioning digital library environment, with a mission not only to preserve content, but to "contribute to the common good by collecting, organizing, preserving, communicating, and sharing the record of human knowledge" [3].

The mission and goals HathiTrust has set for itself are far-reaching. They reflect the bold commitment of the founding partners to ensure the preservation, availability, and use of the knowledge they have collected and curated over centuries, well into the digital age. The remainder of the paper traces the challenges HathiTrust has experienced in its history thus far, the strategies it is employing to address them, and the future work it hopes to accomplish. This is done in the context of three main areas: Governance, the HathiTrust Repository, and the Services the repository provides.

HathiTrust Digital Library

Governance

The challenges to governance in a variety of collective environments (consortia, coalitions, etc.) are well known. In her article, "Come Together Right Now", Barbara McFadden Allen, current director of the CIC, describes two types of bad collaboration: collaboration where there is "goal drift", or movement away from originally stated aims, and collaboration where there is no buy-in from administrative bodies, resulting in insufficient funding and support [4, pg. 97]. She notes also two tensions that often arise in collaborative undertakings: a perception that collaboration will limit the independence of participants, and a fear that it will slow the decision-making process [4, pg. 98]. The solution she suggests to overcome these tensions is to enable collaborative activity that is flexible and voluntary, allowing participants to invest as warranted and allowing executive bodies to receive input from a broad variety of players but not be impeded in taking decisions. HathiTrust governance does just this, and has so far avoided the elements of bad collaboration McFadden Allen mentions. It is designed to be nimble but inclusive, voluntary yet stable. It has two parts: an Executive Committee and a Strategic Advisory Board.

Executive Committee

The Executive Committee is the decision-making body of HathiTrust. It is composed of a small number of senior library and information officers at partner institutions. Executive Committee members have the ability to influence financial and directional decisions at their respective institutions and are the nimble core of HathiTrust. These members currently include:

- Paul Courant, University Librarian and Dean of Libraries, University of Michigan
- Laine Farley, Executive Director, California Digital Library
- Paula Kaufman, University Librarian and Dean of Libraries, University of Illinois at Champaign-Urbana
- John King, Vice Provost for Academic Information, University of Michigan
- Brian Schottlaender, University Librarian, University of California, San Diego Libraries
- Patricia Steele, Dean of Libraries, Indiana University
- Brad Wheeler, Chief Information Officer, Indiana University
- John Wilkin, Executive Director of HathiTrust and Associate University Library, Library Information Technology, University of Michigan

Strategic Advisory Board

The Strategic Advisory Board is the guiding hand of HathiTrust. Its role is to guide HathiTrust development efforts, convene task forces to address issues such as cross-institutional development and de-duplication, and develop policies for HathiTrust and its partners. As of the writing of this paper, the Strategic Advisory Board is still being convened. Once it is fully constituted, it will be comprised, at least, of four members from the institutions of the CIC and three from the University of California system.

HathiTrust is still quite young. The achievements of the collaboration in a very short time have been remarkable, but there is much further still to go. Future developments will be influenced by the decisions of the partners (via the Strategic Advisory Board and Executive Committee). The current plan for HathiTrust is that after a set period of time (a year-and-a-half or two years), when more institutions have had the opportunity to join, something akin to a constitutional convention will take place among the partners to determine a governance scheme that will best serve the needs of all. It is counted as one of the strengths of the current model that such adaptation and response is possible.

Another of the adaptive characteristics of HathiTrust is that membership in the organization is voluntary. This is a thin line to walk for an organization that is premised on long-term preservation, but the partnership model is meant to account for both stability and change. All members join HathiTrust for an initial five-year period, and according the contractual agreement

with HathiTrust, at the end of those five years a member may leave the partnership with a full return of the content they have deposited. The partnership model for HathiTrust is discussed further in the section on Finances below.

Finances

The financial structure of the organization is closely tied to its governance. As such, it is also designed to accommodate flexibility that may occur in partnership while ensuring a high degree of stability. HathiTrust has been funded by its founding partners for an initial five-year period. The funds that these partners contribute come directly out of the annual budgets given to the partner libraries by their sponsoring institutions (with the exception of Indiana University, where the effort is supported directly through the office of the CIO). It is conceivable that funding of HathiTrust by these institutions could end in this time, but as long as it is a partner library's decision to remain in the organization, it is just as likely that an institution would stop funding the library to buy books. In other words, the preservation activities HathiTrust is engaged in are central to the mission of the libraries involved, and funding for them is as stable as funding for those libraries themselves.

The University of Michigan is the current host of the HathiTrust infrastructure, with a mirror site at Indiana. As the host, it is contributing more to the upkeep and maintenance of the repository than the other partners. If it happened that all other partners decided to remove their content after the five-year period, Michigan would need to scale back the size of the operation, perhaps even to one storage site and backup tape, but its own permanent investment in the repository would ensure its survival. Any institution that becomes the host of the repository in the future would likely need to assume the same responsibility (these terms will be made clearer through an ongoing review process). To prevent the possibility of a sudden dissolution of the partnership after five years, a formal evaluation will be conducted by the partners in the third year to assess future goals and strategies.

An additional way that HathiTrust seeks to ensure stability in the enterprise is the funding model for partners. In the current model, when a partner joins for a five year period, it makes an estimate of the number and size of volumes that will be deposited over the course of those five years. In order to avoid financial fluctuations that would come from more content being deposited in one year than others, partners pay an average cost per year, calculated by dividing the total cost of deposit over each of the five years. The final cost to partners is a one-time fee in the first year that is 25% of the average annual cost. This fee is used to build a pool of funds among the partners that can be used in case of a one time resource-intensive need such as migrating all content or moving to a different storage platform.

Repository

A variety of challenges relating to the repository of HathiTrust have arisen in its short history, but it has already had significant success in addressing them. The first of these has been to become widely recognized as a trusted environment for digital preservation, which it has sought to do through international certification processes such as the Trustworthy Repositories Audit and Certification (TRAC) [5] and the Digital Repository Audit Method Based On Risk Assessment (DRAMBORA) [6]. Although these certification processes did not exist when early development of the repository was taking place, the criteria released in the 2002 RLG-OCLC report on Trusted Digital Repositories have been a guide post for the initiative from its earliest days [7].

The repository constructed by the University of Michigan, initially for its own purposes, was built on practical principles. Its purpose was to be adequate for preservation, cost-effective, and quickly implemented. It employed metadata standards such as METS and PREMIS, format standards such as TIFF ITU G4 and JPEG2000, and industry best practices for quality control, data storage and backup, as well as for data integrity and format validation. These decisions have served the repository well, and the infrastructure was expanded to be the basis for HathiTrust when the partnership was launched.

With broad support from multiple leading research institutions and the added preservation expertise and resources those institutions provide, it is hoped that HathiTrust will in the near future be certified as a trustworthy repository. It has already received positive reviews in a recent DRAMBORA evaluation, and will undergo a TRAC assessment by CRL later in 2009. Documentation of HathiTrust's current compliance with TRAC is available on the HathiTrust website [8].

Certification as a trusted repository is not an end in itself for HathiTrust. Although it would be a significant achievement and validation of what HathiTrust has set out to accomplish, the partners recognize the additional value that bringing such a corpus of materials – a collection of collections, essentially – out of their combined holdings, can have.

One of the most significant of these is cooperative collection management and development. For a variety of reasons, collaborative development of print (and electronic) collections among universities has been slow to develop despite the ability to share and aggregate holdings information in unified catalogs and resources. HathiTrust is seen as a strategy for addressing this, and for enabling universities to reduce redundancy and duplication in their combined collections.

There are several parts to this, and several challenges. The first part is discovering redundancy through digitization and ingest of materials into the repository. Google already has a process through which it attempts to avoid digitizing copies of the same work that come from different institutions. Additional mechanisms will be needed, first to ingest content from sources other than Google (such as the Open Content Alliance and local digitization programs), and second to identify duplicates from those sources. This may become more complicated as additional partners join. Policies will also need to be developed to handle instances where the same item is held in the repository in different formats (once as JPEG2000 and once as a hybrid of JPEG2000 and TIFF, for example).

The second part is understanding and responding to the implications of having a unified digital repository of research materials drawn from libraries around the world. It is hoped that at some point it will be possible to certify individual volumes in the repository as permanent preservation representatives of their print counterparts. Under these circumstances, it may be possible for some libraries to stop collecting, or even to begin de-accessioning volumes in their own collections that are held as certified copies in the common collection.

Such thinking is not without precedent. As Daniel Greenstein and Suzanne Thorin pointed out in 2003, this has already begun to some degree in libraries with regard to scholarly journals held by JSTOR [9]. Universities such as the University of California have dramatically reduced costs and improved services to individual campuses and libraries by creating shared repositories for their print collections. It is not difficult to see how these benefits could be transferred to the digital realm. Libraries, especially small libraries, that were relieved of the need to spend large portions of their budgets accessioning, maintaining, and preserving a core collection of materials, would be able focus their collecting efforts in more targeted areas and special collections. A view of libraries similar to the model of the CIC emerges, where individual institutions are able to build more freely on their strengths, contributing at the same time to the benefit of the entire community.

Google Settlement

A note here should be made about the Google Settlement because it has great bearing on the repository in this regard, and in others to be discussed below. Significant debate is occurring in library, publishing, and legal communities surrounding the terms of the Settlement: what it means, what it will make available, what it will restrict. There is no doubt that the Settlement will have a deep and lasting impact on the nature of scholarly research. The HathiTrust partners believe that the Settlement will have an enormous positive bearing on the initiative, allowing it to make more materials available with better tools for accessing and using them than would otherwise be possible. Some of ways this will occur are described below. Specifics of the Settlement are not discussed here, but a wealth of information, including positions of HathiTrust partners and others, can be found on the Internet.

Services

Returning to the benefits of a single architecture and infrastructure for shared collections discussed in the previous section, an additional value that is gained through this arrangement is the range of services that are enabled. In HathiTrust, these have come in three different levels: basic access, search, and extended capabilities. As in other areas of HathiTrust, some of the services below have been implemented and others are yet to come.

Basic Access

Some may argue that search and access should not be separated in this way, but what is meant by basic access here is access where there was none before – specifically, to users with print disabilities. Certified users at partner libraries will be able to "check out" electronic copies of in-copyright books owned by the partner library for use with screen readers and digital Braille devices. The HathiTrust partners see this as one of the most exciting services the shared repository will enable, and it is one that will be explicitly sanctioned under the terms of the Settlement. An additional key goal of the repository in the area of basic access is to ensure compliance with accessibility standards and best practices for users with disabilities more broadly, making the content in HathiTrust as open to as many users as possible.

Search

As of the time of writing, there is no single interface for searching the HathiTrust collections. Although a temporary public beta discovery system is planned for release in April, access to the repository is currently gained through library catalogs and independently created prototypes [10] that make use of tools such as the HathiTrust OAI feed, metadata files that are available for download, or a specialized API to obtain bibliographic data for HathiTrust items [11]. The last two of these are mechanisms created specifically for this purpose, and are examples of the services HathiTrust will provide. Libraries anywhere in the world are currently able to add HathiTrust records to their online catalogs, expanding the holdings available to their users.

Full text search of repository materials is an ongoing challenge for HathiTrust. In September 2008, a large-scale search benchmarking process was begun, exploring the use of Solr [12] to provide keyword and phrase searching capabilities across the entire corpus. As with governance, there were no comparable models for how to proceed in an undertaking of this magnitude. A report containing results of the first two stages of benchmarking was published on the HathiTrust website in December [13] and testing of further stages is ongoing. It is not clear currently if Solr's capabilities will scale to the needs of full-text searching in HathiTrust, but testing so far has been encouraging. A publicly available search beta was released in November 2008 to demonstrate the progress that has been made [14]. More hardware is currently being purchased to continue the search benchmarking.

Extended Capabilities

In "What Is A Digital Library Anyway?" Lagoze, Krafft, Payette, and Jesuroga from Cornell University describe a model for digital libraries "that intentionally moves "beyond search and access", without ignoring those basic functions, and facilitates the creation of collaborative and contextual knowledge environments." [15] This is precisely what HathiTrust aims to do. In addition to the mechanisms for sharing bibliographic information described above, a data API will soon be available that will allow the creation of tools and services on top of the repository. This may include integration of openly accessible volumes with course reserves software, integration with primary source collections, or anything else that partner and other libraries can imagine. Current examples of applications built using this model are the HathiTrust PageTurner [16] and Collection Builder [17].

A component closely related to building tools and services on top of the repository is making volumes inside the repository available for non-consumptive research. This is another activity that will be explicitly sanctioned and enabled through the Google Settlement. In addition to providing valuable linguistic data, the ability to run targeted routines on the repository will be invaluable to researchers in the digital humanities. HathiTrust is currently exploring the prospects and feasibility of using the SEASR [18] framework to provide these capabilities.

An additional service that is envisioned for HathiTrust at this time is print-on-demand of repository materials. The Scholarly Publishing Office at the University of Michigan Library has been a leading force in this regard, and is in active negotiations with Amazon and Hewlitt Packard (its distributing partners) to develop a workflow for making University of Michigan materials in HathiTrust available for print-on-demand. The Scholarly

Publishing Office hopes, in cooperation with HathiTrust, to design a set of print-on-demand services that will be of benefit to all the partners.

Some of the main challenges to offering services on top of, or with, repository content are inter-institutional authentication and security. Services that have the potential to be most useful to users are those that allow customized views of the repository based on identity (the ability to save items into collections, view recommended items of interest, etc.) and that integrate with local institutional services. Models and policies for how to provide these capabilities are still in formative stages.

Security is also a large concern. The digital library's focus on preservation, concerns about copyright, and partner agreements with Google create many levels of responsibility that must be acknowledged and upheld when providing access to, and distributing content. The desire to make resources as open and available as possible must be balanced with the need to ensure their integrity in the long-term, and honor the agreements and terms by which they have entered the repository.

In all of these efforts, through the repository it has created and services it provides, HathiTrust's overarching goal is to create a community-owned, community-driven resource of unprecedented size and flexibility, that is able to provide the next generation of tools and services to serve and support the scholarly community.

Conclusion

HathiTrust is still quite young, but the changes and adjustments it is experiencing should not be seen as the fits and starts of a fledgling operation. They are rather the first bold steps of a digital library movement with roots deep in the histories and traditions of libraries that is finally beginning to hit its stride. In a 2003 essay entitled "Keeping Libraries At The Center Of The University", Daniel Atkins expresses significant concern that academic libraries are not doing all they could or should do to remain active players in an increasingly federated, collaborative, open, and digital information environment. He asserts the need for academic libraries to "envision and plan for non-linear change and act more radically in order to stay the same" – that is, in order "to continue to add significant value to the academic community in achieving intellectual, physical, and long-term access to data, information, and knowledge." [19]

HathiTrust is that radical action. It is a bold move by leading research libraries to move their collections and traditional values into a new digital era. The governance and repository structures of HathiTrust have been engineered to create a library that never forgets; the services it offers will ensure it is a library that is never forgotten. HathiTrust can be the universal library, but it is something we will all need to be a part of, and something it will take all of our institutions to create.

References

[1] University of Michigan, Google Inc. Cooperative Agreement (December 14, 2004), 4.4.2. Available at: http://www.lib.umich.edu/mdp/um-google-cooperative-agreement.pdf (accessed March 2009).

[2] Committee on Institutional Cooperation, Google Inc. Cooperative Agreement (June 6, 2007) 4.13. Available at http://www.cic.net/Libraries/Library/CIC-GoogleAgreement.sflb (accessed March 2009).

[3] HathiTrust Mission and Goals. Available at: http://www.hathitrust.org/mission_goals (accessed March 2009).

[4] Barbara McFadden Allen, Come Together Right Now in Digital Libraries: A Vision for the 21st Century (University of Michigan Library Scholarly Publishing Office, Ann Arbor, 2003). Available at: http://quod.lib.umich.edu/cgi/t/text/text-idx?c=spobooks;;idno=bbv9812 (accessed March 2009)

[5] Trustworthy Repositories Audit and Certification (TRAC): Criteria and Checklist (OCLC and CRL, 2007). Available at http://www.crl.edu/PDF/trac.pdf (accessed March 2009).

[6] DRAMBORA. http://www.repositoryaudit.eu/about/ (accessed March, 2009).

[7] RLG-OCLC, Trusted Digital Repositories: Attributes and Responsibilities (RLG, Mountain View, CA., 2002). Available at: http://www.oclc.org/programs/ourwork/past/trustedrep/repositories.pdf (accessed March 2009).

[8] HathiTrust Review of Compliance With Trustworthy Repositories Audit & Certification (TRAC). Available at: http://www.hathitrust.org/technical_reports/trac.pdf (accessed March 2009).

[9] Daniel Greenstein and Suzanne E. Thorin, The Digital Library: A Biography. (Digital Library Federation, Council on Library and Information Resources ,Washington, D.C., 2003) pg. 25. Available at: http://www.clir.org/pubs/reports/pub109/pub109.pdf (accessed March 2009).

[10] Access to HathiTrust. http://www.hathitrust.org/access (accessed March 2009).

[11] HathiTrust Bibliographic Data Distribution. http://www.hathitrust.org/bibliographic_data_distribution (accessed March 2009).

[12] Solr. http://lucene.apache.org/solr/ (accessed March 2009).

[13] Large Scale Search. http://www.hathitrust.org/large_scale_search (accessed March 2009).

[14] Beta Full-Text Search. http://babel.hathitrust.org/cgi/ls (accessed March 2009).

[15] Carl Lagoze, Dean B. Krafft, Sandy Payette, Susan Jesuroga, "What Is A Digital Library Anyway? Beyond Search and Access in the NSDL," Jour. Computing and Information Science, 11, 11 (2005). Available at: http://www.dlib.org/dlib/november05/lagoze/11lagoze.html (accessed March 2009).

[16] HathiTrust PageTurner (example). http://babel.hathitrust.org/cgi/pt?id=mdp.39015000673742;page=root;view=image;size=100;seq=9;num=vii (accessed March 2009).

[17] HathiTrust Collection Builder. http://babel.hathitrust.org/ (accessed March 2009).

[18] SEASR. http://seasr.org (accessed March 2009).

[19] Daniel E. Atkins, Keeping Academic Libraries At The Center Of The University in Digital Libraries: A Vision for the 21st Century (University of Michigan Library Scholarly Publishing Office, Ann Arbor, 2003). Available at: http://quod.lib.umich.edu/cgi/t/text/text-idx?c=spobooks;;idno=bbv9812 (accessed March 2009).

Author Biography

Jeremy York is a project librarian for HathiTrust Digital Library. He graduated from Emory University in 2001 with a B. A. in History and received a Master of Information Science from the University of Michigan in 2008. He has more than ten years experience in libraries, working in areas of course reserves, archives and special collections, and information technology.