Intellectual Property Rights & the HathiTrust Collection

Heather Christenson[1] and John P. Wilkin[2]
[1]*University of California, California Digital Library, heather.christenson@ucop.edu*
[2]*HathiTrust*

Abstract
Research libraries founded HathiTrust in 2008. This digital preservation and access collaboration of over 60 research libraries in the United States, Canada, and Europe utilizes a shared infrastructure to preserve digital copies of now over 10 million volumes digitized from print. HathiTrust's mission is "to contribute to the common good by collecting, organizing, preserving, communicating, and sharing the record of human knowledge." This paper introduces the goals of HathiTrust, describes the scale and scope of the HathiTrust collection and its significance, and discusses how the organization is providing services related to the digital collection, in light of changing conditions for maintenance of the participating libraries' print collections, and, in particular, in the context of the current environment for intellectual property rights.

Authors
Heather Christenson is Mass Digitization Project Manager at the California Digital Library (CDL), and Project Manager for the University of California's HathiTrust collaborative work. Ms. Christenson is responsible for coordinating operations across the UC Libraries' large scale digitization projects, and plays an integral role in planning for services surrounding digitized content. Prior to joining CDL, Ms. Christenson was involved in the development of commercial web search tools, and was a news librarian, law librarian, and cataloger. She received her M.L.I.S. from the University of California, Berkeley.

John P. Wilkin is the Associate University Librarian for Publishing and Technology at the University of Michigan, and is the Executive Director of HathiTrust. Mr. Wilkin provides operational and strategic leadership for HathiTrust, a growing international partnership committed to using digital collections to help libraries improve access and preservation strategies. In his role at the University of Michigan Library, Mr. Wilkin also leads publishing efforts, including the Library's pioneering digital publishing operations and the University of Michigan Press, and the Library's technology operations.

1. Introduction

Research libraries in the United States have been digitizing their materials for almost two decades, both individually and via collaborative projects. Digitization is expensive, and in the absence of an official national library program or long-term national funding, the libraries have accomplished the task of converting books to digital form in a number of ways: through partnerships such as collaboration with Google, grant funding, and self-funding. Unlike commercial enterprises, however, research libraries place a great deal of value on digital preservation, and in the provision of digital content for scholarly uses into the future.

Although the conversion of library materials from print to digital form has happened at a brisk pace, the law has been slow to evolve in terms of considering use of mass digitized library collections. Research libraries view as an imperative their traditional role as stewards of the record of human knowledge, regardless of format, so they must do their best in good faith to interpret existing laws, to act lawfully, and to act in the public interest.

In 2008 research libraries founded HathiTrust, a digital preservation and access collaboration of over 60 research libraries in the United States, Canada, and Europe. HathiTrust utilizes a shared infrastructure to preserve digital copies of now over 10 million volumes digitized from print. HathiTrust's mission is "to contribute to the common good by collecting, organizing, preserving, communicating, and sharing the record of human knowledge."[1] With partners that include libraries in major public and private colleges and universities, independent research libraries such as the Getty and the New York Public Library, and the Library of Congress, and a collection that is increasingly comprehensive, HathiTrust is rapidly becoming a central entity for preservation of and access to library collections. Any given HathiTrust partner library is likely to find more than 50% of its print collection online in HathiTrust. With such a large aggregate collection, the range of works in HathiTrust represents the full spectrum of research library collections, including the copyright status of materials in those collections. Consequently, the HathiTrust libraries must navigate current copyright law to practice responsible stewardship of library collections and to continue their service mission, in the digital realm.

This paper introduces the goals of HathiTrust, describes the scale and scope of the HathiTrust collection and its significance, and discusses how the organization is providing services related to the digital collection, in light of changing conditions for maintenance of the participating libraries' print collections, and, in particular, in the context of the current environment for intellectual property rights.

2. Goals of HathiTrust

Structurally, HathiTrust is not a "trust" in the legal sense of the word, nor is it a corporation or even a nonprofit organization. It is a collaborative enterprise of research libraries that depends on funding and in-kind contributions from its members.

The name HathiTrust was chosen to reflect the values of the organization. Hathi (pronounced hah-tee) is the Hindi word for elephant, an animal that symbolically represents memory, wisdom and strength. In concert with its overarching mission, the initial goals set by the HathiTrust partners are:

- *To build a reliable and increasingly comprehensive digital archive of library materials converted from print that is co-owned and managed by a number of academic institutions.*
- *To dramatically improve access to these materials in ways that, first and foremost, meet the needs of the co-owning institutions.*
- *To help preserve these important human records by creating reliable and accessible electronic representations.*
- *To stimulate redoubled efforts to coordinate shared storage strategies among libraries, thus reducing long-term capital and operating costs of libraries associated with the storage and care of print collections.*
- *To create and sustain this "public good" in a way that mitigates the problem of free-riders.*

[1] http://www.hathitrust.org/mission_goals

- *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.* [2]

HathiTrust has values of openness and collaboration, and aims to be transparent in its governance and operations. In addition to digital preservation, the organization also aims to provide access to materials to the extent legally permissible.

3. The HathiTrust Collection

The HathiTrust collection has its origins in mass digitization projects conducted in partnership with Google and the Internet Archive, but incorporates much more. HathiTrust brings together collections from many major Google library partners, including the two largest, University of Michigan and University of California, and has the largest collection of items digitized by Google. HathiTrust has also gone a long way towards archiving digital volumes created during Microsoft's Live Search Books 2006-2008 project, volumes which were digitized by the Internet Archive and others. More recently, HathiTrust has focused on incorporating materials that have been locally digitized by the partner libraries. Although HathiTrust content primarily originates from libraries within the United States, the HathiTrust partnership includes international partners such as Biblioteca de la Universidad Complutense de Madrid in Spain, which has contributed a large amount of content, and McGill University in Canada. From its inception in 2008, HathiTrust has grown to include over 10.5 million volumes.

HathiTrust aspires to comprehensiveness and has used mass digitization to accomplish that goal. Consequently, HathiTrust does not have a collection development policy that requires the partners to adhere to any specific subject, language, or content criteria. The HathiTrust partner libraries believe that the value of HathiTrust is in the whole collection, and that this aggregation is reflective of research library collections selected for scholarly value and preserved over time in print by libraries. The aggregate collection also offers the opportunity for differentiation of specific digital collections from the whole, post-digitization, via a layer of services. For example, by aiming for comprehensiveness, HathiTrust is more easily able to offer up sub-collections like English language literature before 1800 or US federal government documents. The vast collection holds the potential to be curated and presented in a multitude of ways using tools available now or developed later. Like the research library collections encompassed within it, HathiTrust serves a broad constituency by incorporating works that did not top the best-seller lists but that serve the "long tail" activities of specialized research and scholarship.

The collection spans the gamut of languages in research libraries: more than 400 languages are currently represented in HathiTrust, of which the highest percentages of volumes are in English, German, French, Spanish, Chinese and Russian. Most languages are present in the collection in smaller percentages, but because the collection is so large, the percentages still represent large numbers of digital volumes, for example Indonesian (33,726), Norwegian (15,429) or Afrikaans (1,053).

4. Significance of the Collection

HathiTrust serves as shared infrastructure for partner libraries to use in managing their print collections. The significance of a *preserved and dependable* collection of this magnitude is beginning to be

[2] Ibid.

appreciated. In early 2011, an OCLC Research study by Malpas reported results of an analysis of the HathiTrust collection relative to the volumes held by US research libraries in print, and found HathiTrust to be increasingly representative of the physical collections in research libraries.[3] This holds a number of implications for the libraries in terms of greatly needed understanding of how much of what is held there has been digitized, and how much remains to be digitized; the costs to digitize the remainder cannot be determined unless we know the scale and scope of what is left.

If a given library can possess an understanding of how its particular collection maps to the digitized whole, and can rely on HathiTrust for preservation of digital versions of those books, the library can then make informed decisions about how and where to store its physical book collection, including which print books are essential to keep. When digital books are collaboratively made available, advantages can accrue through collaborative agreements for retention of the physical books, allowing libraries to reduce storage costs in the presence of widely available digital copies.

5. How HathiTrust Provides Services in the Context of the Current Environment for Intellectual Property Rights

The HathiTrust corpus includes millions of works, including both public domain and in-copyright books and serials. HathiTrust can store these works because US law places limitations on the exclusive rights of the rights holders, and those limitations support both fair use and preservation purposes. In order to provide access to the digital volumes in its collection, HathiTrust relies on US and international copyright law and rights determinations for the corresponding print volumes. For example, HathiTrust uses the publication dates and countries of publication in cataloguing records to identify large bodies of public domain works, and in many cases rights holders grant HathiTrust permission to provide open access to materials in the collection. The totality of the HathiTrust strategy can be characterized as a combination of automatic rights determinations, manual rights determinations, permissions and agreements, and legal interpretations.

5.1 Automatic Rights Determinations

HathiTrust's automated rights determination processes identify materials that we can reliably characterize as being in the public domain, either based on US law or common attributes of non-US copyright law. By analysing a number of fixed and free fields in the MARC record, we make a first pass at identifying public domain works, characterizing the remainder as presumptively in copyright. Although we are not able to exhaustively detail the criteria that we consider in making these determinations, several key examples will help illustrate the process. Most US works published in the United States before 1923 are in the public domain worldwide. US law defines the majority of US federal government publications as public domain. US law also treats non-US works published before 1923 as public domain in the United States; consequently, HathiTrust provides open access to these publications for users coming from network addresses within the United States. To provide access to non-US works for users outside of the United States, HathiTrust generally relies on the Berne Convention[4] as a framework for decision-making.

[3] Constance Malpas, *Cloud-Sourcing Research Collections: Managing Print in the Mass-Digitized Library Environment* (Dublin, Ohio: OCLC Research, 2011). http://www.oclc.org/research/publication/library/2011/2011-01.pdf

[4] http://www.wipo.int/treaties/en/ip/berne/trtdocs_wo001.html

In many countries, non-US works are only in the public domain 70 years after the death of the author, and because author death date information cannot be reliably inferred in the cataloguing record, we have created a "rolling wall" of 140 years before the current year. (This approach provides us with some protection against assuming a public domain status for a work where the author published a work at a young age and lived for an exceptionally long time). HathiTrust's automated routines for determining the public domain status of works is published online on the HathiTrust website.[5] Although the current decision-making framework is focused primarily on US copyright law, Canadian HathiTrust partner libraries have begun discussions to define specific exceptions found in Canadian copyright law.

5.2 Manual Rights Determinations

HathiTrust also uses a carefully defined set of procedures, systems and legal guidance to make manual rights determinations. With generous assistance from the Institute of Museum and Library Services, HathiTrust implemented a Copyright Review Management System (CRMS) in 2008. The design of that system was guided by and continues to be refined by legal scholars. It incorporates strategies such as double-blind review in order to increase the reliability of determinations. The reliability of determinations has been and will continue to be tested against benchmark data (e.g., record analysis by the US Copyright Office). The first CRMS work was focused on books published in the United States between 1923 and 1963; current work is focused on non-US books published in English-language speaking countries and Spain. The Copyright Review Management System is documented online.[6] Additionally, legal experts may flag individual works for review. All manual decisions override automated decisions, and both sets of decisions are registered in a HathiTrust-maintained Rights Database. The architecture and decision-making related to the Rights Database is also documented online on the HathiTrust website.[7] More than 100,000 works have been opened using manual determinations; roughly 55% of the works reviewed have been found to have a public domain status.

6. Permissions

Recognizing that many rights holders believe that open, online access to their publications is either part of their mission or in their best interests, HathiTrust supports several strategies for individuals and organizations to open access to their publications. HathiTrust makes a form available online so that the rights holder may convey perpetual and non-exclusive permission for access; this form supports many methods for access, including simple permission without changing the copyright status of work, and application of a Creative Commons license.[8] Additionally, HathiTrust has negotiated agreements with some rights holders (e.g., Duke University Press) where entire lists of titles are opened with a Creative Commons license, and, in return, HathiTrust provides files and updates to the rights holder. More than 7,000 works have been opened through explicit rights holder permissions.

Using this combination of strategies, HathiTrust has opened more than 3 million works for users at US network addresses, and more than one million works for users worldwide. This represents roughly 30% of the HathiTrust collection. The remainder of the HathiTrust collection, approximately 70% of the

[5] http://www.hathitrust.org/bib_rights_determination
[6] http://www.lib.umich.edu/grants/crms/
[7] http://www.hathitrust.org/rights_database
[8] http://www.hathitrust.org/permissions_agreement

works, has either been determined to be in copyright through the CRMS process or is assumed to be in copyright, pending further investigation.

7. Other Lawful Uses of Digital Materials

HathiTrust has used legal guidance to undertake other strategies to provide access to works in the HathiTrust corpus. Rubrics such as fair use in US copyright law (or fair dealing in other regimes) provide a framework for some uses, and HathiTrust supports some of these. In addition, constituencies like the blind or other persons with print disabilities may be served under legal regimes like that in the United States. Similarly, some provisions of US law support HathiTrust's preservation mandate.

Under US copyright law, including the fair use provisions, HathiTrust has developed and provides a powerful discovery mechanism for the entirety of the corpus. Every word and phrase (in hundreds of languages and many character sets) in HathiTrust is indexed and searchable by users worldwide. Where HathiTrust has determined that a work may be made accessible to a given user, the search results provide a significant amount of context, and links are provided to the full text, which can then be read online. In other cases, either in the limited number of instances where we know the work to be in copyright, or where we treat the work as being in copyright in the absence of more reliable information, HathiTrust reports the page numbers and the number of hits per page to the user who conducts a search. This powerful search capability has been extremely helpful to many scholars, as it serves as a master index to a corpus of billions of pages.

Many legal regimes support use of in-copyright works for users with print disabilities. For example, the Chaffee Amendment to US copyright law, Section 121, allows an authorized entity to provide access to works that are protected by copyright to certain users.[9] In addition, certain uses of in-copyright works to make them available to the blind have been determined to be fair use under US copyright law. The mechanisms HathiTrust has put in place are preliminary, pending the resolution of the legal challenges facing HathiTrust. Currently, using this framework, HathiTrust provides access to millions of works for University of Michigan users certified to have print disabilities. In each case, HathiTrust provides the authenticated user access to the underlying text through a special interface so that the user may use the text with a digital braille or other reading device. Only digital copies of works that have been determined to be part of Michigan's print collections are included in the service.

Preservation-related provisions in law support other lawful uses of works that are in copyright. In US copyright law, Section 108 supports limited services when an in-copyright work is not available on the market in an unused copy at a reasonable price, and where the library's copy is damaged, deteriorating, lost or stolen.[10] As with services for the print-disabled, the mechanisms HathiTrust has put in place are preliminary, pending the resolution of the legal challenges facing HathiTrust. Currently, at the University of Michigan, HathiTrust provides access to certain damaged, deteriorating, lost or stolen works under this interpretation of US law, only to University of Michigan users. These University of Michigan users are not able to download the work in its entirety (i.e., they are currently only able to read the work continuously on the screen or to download one page at a time). No more than one simultaneous user per copy owned may view the work. Each work is clearly marked as being in copyright and the user is notified that access is supported under this interpretation of US copyright law.

[9] For example, see http://www.loc.gov/nls/reference/factsheets/copyright.html.
[10] See, for example, http://www.law.cornell.edu/uscode/text/17/108.

As librarians, we must navigate a complex intellectual property rights landscape. Because of the importance and complexity of our work, our University counsels and other legal scholars guide us in making these decisions. Recent work by Peter Jaszi, Jennifer Urban, Pam Samuelson, and other American legal scholars has been helpful, but practical decisions for an organization like HathiTrust are largely untested. We hope, through the processes documented here, to build responsible foundations upon which other uses can be defined.

8. Conclusion

As a digital research library collection unprecedented in size and scope, HathiTrust serves an increasingly pivotal role. HathiTrust has become a vehicle to support end user access to the record of human knowledge and to support the preservation of library collections. Libraries have existed for hundreds of years, each building its distinctive collection with more or less complementarity to other collections. Now, through aggregation, libraries are using HathiTrust to explore questions of bibliographic identification, of collection management, and of copyright determination. Through this collectivity, libraries have begun to make strides in facing the economic and legal challenges inherent in the management and use of digitized library collections.

References

Berne Convention for the Protection of Literary and Artistic Works
http://www.wipo.int/treaties/en/ip/berne/trtdocs_wo001.html

Cornell University Law School, Legal Information Institute. "17 USC § 108 - Limitations on exclusive rights: Reproduction by libraries and archives." http://www.law.cornell.edu/uscode/text/17/108

HathiTrust. "Mission and Goals."
http://www.hathitrust.org/mission_goals

----------."Bibliographic Rights Determination"
http://www.hathitrust.org/bib_rights_determination

----------."Permissions Agreement"
http://www.hathitrust.org/permissions_agreement

----------."Rights Database"
http://www.hathitrust.org/rights_database

Library of Congress, "NLS Factsheets: Copyright Law Amendment, 1996."
http://www.loc.gov/nls/reference/factsheets/copyright.html.

Malpas, Constance. 2011. *Cloud-Sourcing Research Collections: Managing Print in the Mass-Digitized Library Environment*. OCLC Research. http://www.oclc.org/research/publication/library/2011/2011-01.pdf

University of Michigan Library. "Copyright Review Management System - IMLS National Leadership

Grant." http://www.lib.umich.edu/grants/crm