

# The HathiTrust

## A Report for the ALA Office For Information Technology Policy

February 27<sup>th</sup>, 2009

Greg Grossmeier<sup>1</sup>

### Overview

The HathiTrust (pronounced hah-tee) is a collaboration started by a 13 member consortium including the Committee on Institutional Cooperation and the University of California and open to any library working on large-scale digitization projects. The goal of the collaboration is to create a massive digital repository where the member universities can archive and share their collections with an emphasis on long-term preservation.

The project is currently hosting a very large collection of scanned works. At the time of writing the HathiTrust holds 2,625,451 volumes of which 393,793 (~15%) are in the public domain<sup>2</sup>. In contrast, the Internet Archive is hosting 1,064,822 full-text/public domain volumes through its OpenLibrary service<sup>3</sup>.

### Relation to Google Book Search

While the HathiTrust is a collection of materials scanned by the partner libraries it is primarily made up of scanned volumes by the Google Books Library Project (GBLP) at the University of Michigan. This is due to the extreme pace at which Google is able to scan books compared to the past and current scanning projects at these libraries. The vast quantity of these books scanned by Google in the HathiTrust means that the agreement between the University of Michigan and Google influences the activities of the HathiTrust. For instance, due to the agreement, Michigan must have various technical restrictions on its digital copy of the book to prevent automated downloading by third parties. The specific restrictions in place are throttling of web traffic and “chunking”<sup>4</sup> of the books into individual pages<sup>5</sup>.

Some of the agreements between Google and GBLP participating libraries seem to not allow a partnership such as the HathiTrust. For instance, the agreement between the University of California and Google states that the UC Library is not allowed to “share, provide, license, distribute, or sell” their Google-supplied copies “to any entity in any manner.” It can, according to the agreement, only provide up to 10% of a digital copy to other library institutions for scholarly functions.<sup>6</sup> Google has, however, given “informal permission”<sup>7</sup> to the libraries participating in the HathiTrust to share their library-owned digital copies of the scans. While this is just informal permission at this time, the members of the HathiTrust have worked to ensure that “revisions to agreements that will necessarily come in the wake

---

1 greg@grossmeier.net – This article is licensed under a Creative Commons Attribution-ShareAlike license:  
<http://creativecommons.org/licenses/by-sa/3.0/>

2 “Welcome to the Shared Digital Future | [www.hathitrust.org](http://www.hathitrust.org)” Available at <http://www.hathitrust.org>

3 “OpenLibrary.” Available at <http://openlibrary.org>

4 Personal communication with John Wilkin, Executive Director, HathiTrust. February 24<sup>th</sup>, 2009.

5 The HathiTrust also implements the use of a robots.txt file that tells well-behaving web archivers to not download certain sections of the HathiTrust website. A robots.txt file does not stop the downloading of files by malevolent entities.

6 Cooperative Agreement between Google and the University of California. Available online at [http://www.cdlib.org/news/ucgoogle\\_cooperative\\_agreement.pdf](http://www.cdlib.org/news/ucgoogle_cooperative_agreement.pdf)

7 Personal communication with John Wilkin, Executive Director, HathiTrust. February 24<sup>th</sup>, 2009.

“[T]hough the UC agreement would not appear to, Google gave their permission because they believe there’s value in having institutions collaborate to do the types of things in HathiTrust. That is, they have given informal permission to each of the participating institutions and have not required a separate or revised agreement.”

of the [Google Book Search] settlement will treat this explicitly.”<sup>8</sup>

As the Google Book Search settlement has not yet been finalized by the court there are additional issues which should be addressed to ensure libraries are satisfactorily represented. One aspect is Google's copyright-like restrictions on public domain materials. The disclaimer on the first page of a downloaded PDF of a public domain book includes a description of what you can and can not do with the PDF. The first point is not to make commercial use of the book. As these works are in the public domain, and any faithful and accurate representation of a public domain work does not create any new copyright interest, the digital scans provided by Google are themselves public domain<sup>9</sup>. Thus, any person is allowed to make use of these scans in any fashion they desire, including commercial endeavors. However, agreeing to the usage guidelines provided by Google artificially limits that right.

Additionally, the new usage guidelines provided by Google expressly do not allow users to “engage in large scale redistribution or rehosting of the files”<sup>10</sup>. This explicitly denies the possibility of such projects as the Internet Archive from hosting the public domain works which is, again, a copyright-like restriction on public domain material. It would be in the interests of libraries and library users to ensure that adequate changes are made to the settlement to forbid Google from making these restrictive assertions in the future<sup>11</sup>. And any decision by a university library to join the HathiTrust should also acknowledge the implications of the Google Book Search Settlement and be fully aware of any limitations or secondary requirements that it entails.

### **Copyright Determination**

The University of Michigan has recently received a grant from the Institute of Museum and Library Services to develop a Copyright Review Management System (CRMS) “to increase the reliability of copyright status determinations of books published in the United States from 1923 to 1963”<sup>12</sup>. The goal of this system is to increase the number of books determined to be in the public domain from the period of time when that is not an easy question to answer.

To determine the copyright status of these books requires two important pieces of information: if the work had a copyright symbol (©) displayed on the work and if the work's copyright was renewed at the United States Copyright Office. Using the digital scans of the books' copyright page and available copyright renewal database records, the CRMS will develop a workflow that will efficiently and accurately determine the copyright status of the large number of books available in the HathiTrust<sup>13</sup>. The limitation of this program is that only books available the CRMS via the HathiTrust will be examined. After a public domain determination is made by the CRMS, not only will the book be available in full-text view in the HathiTrust, the Google Book Search service could accept this determination and provide the same level of access there.

### **Comparison to Institutional Repositories**

The HathiTrust is essentially a Multi-Institution Repository. It provides a very large amount of server space to host the materials: 190 Terabytes total with about 100 Terabytes in use<sup>14</sup>. The complete system is mirrored at both Michigan and Indiana. One of the benefits of this system is that it is uniformly managed at both mirrors. From this uniformity the ability to share knowledge between

---

8 Ibid.

9 See *Bridgeman Art Library v. Corel Corp.*, 36 F. Supp. 2d 191 (S.D.N.Y. 1999) and *The Antithesis of Originality: Bridgeman, Image Licensors, and the Public Domain*, 30 *Hastings Comm. & Ent. L. J.* 257, 267 (2008)

10 “What can I do with the PDFs I download?” Available at <http://books.google.com/support/bin/answer.py?answer=44667>

11 For a discussion of recommended changes to the settlement, see: *How to Improve the Google Book Search Settlement* (forthcoming) by James Grimmelmann.

12 “IMLS – News and Events.” Available at [http://www.ims.gov/news/2008/091008a\\_list.shtm#MI](http://www.ims.gov/news/2008/091008a_list.shtm#MI)

13 “UM receives grant for Copyright Review Management System” Available at <http://mollykleinman.com/2008/09/11/um-copyright-review-management-system/>

14 “Update on October 2008 Activities.” Available at [http://www.hathitrust.org/updates\\_october2008](http://www.hathitrust.org/updates_october2008)

participants is maximized thus lowering training and maintenance costs. Another benefit is the geographical redundancy that the mirroring provides; if one location is damaged due to a natural disaster or similar event the other will have a full working copy.

Additionally, the HathiTrust, like many Institutional Repositories, provides a persistent resource locator for all the materials within it. This is done using the handle.net service<sup>15</sup> which provides unique persistent URLs for any object available online. This is the same service that the University of Michigan already uses for projects like the Scholarly Publishing Office<sup>16</sup> which publishes open access journals and monographs. The HathiTrust is also working with the OCLC to provide more avenues of discovery, specifically adding records to the WorldCat search database for HathiTrust materials<sup>17</sup>.

The HathiTrust differs from a standard Institutional Repository by its collection, not the technical aspects of the system. The collection in the HathiTrust is not focused solely on the scholarly output of one specific institution. It is instead focused on preserving and providing access to the collections of the libraries as a whole. However, including the content of the participating members' institutional repositories (and similar services) is a possibility for the HathiTrust but "it's more a theoretical possibility than an impending one"<sup>18</sup>.

### **The Scope of the HathiTrust**

The HathiTrust is focused primarily on preserving text materials including books, monographs, and scholarly journals. The scholarly journals which it is archiving are the bound editions available in the libraries of the participating libraries. These were scanned by either Google or by individual libraries own digitization efforts. Because it is only scans of bound journals, the collection of scholarly journals may not be the complete collection that the libraries have access to digitally. Also, in the future the HathiTrust plans to expand to a more encompassing collection of scholarly material, not just the material that was scanned from the shelves.

However, given the current goals of the HathiTrust, it is not going to start including other types of media such as video or audio. There are other projects underway that are performing those functions such as the Internet Archive with public domain material and the National Film Registry<sup>19</sup> for many works still under copyright. The scope of the HathiTrust is text and expanding that to include other materials would not provide many benefits. Instead, there should be an extensive Application Programming Interface (API) for the HathiTrust so that the material can be used and displayed in new and imaginative ways. This API is already in development and is allowing others to develop their own methods of searching the HathiTrust<sup>20</sup>.

### **Where it Fits with Other Projects**

There are projects performing a similar function in other areas of cultural preservation. I will outline these services based on the licensing restrictions of the content they host.

#### *Public Domain or Openly Licensed Content*

I will take as an example one of the original and largest preservation projects of public domain or openly licensed content: the Internet Archive. The Internet Archive is not just a project to archive internet webpages, but also an archive of public domain text, video, and audio. Along with the collected public domain content the Internet Archive also allows uploading of newly created content as long as it is available under permissive licenses such as a Creative Commons license or similar.

The HathiTrust provides a unique and complementary service to the Internet Archive. The

---

15 "The Handle System." Available at <http://www.handle.net/>

16 "Scholarly Publishing Office." Available at <http://www.lib.umich.edu/spo/>

17 "HathiTrust to work with OCLC." Available at <http://www.oclc.org/news/releases/20097.htm>

18 Personal communication with John Wilkin, Executive Director, HathiTrust. February 24<sup>th</sup>, 2009.

19 "National Film Registry." Available at <http://www.loc.gov/film/filmnfr.html>

20 "HathiTrust API." Available at [http://www.hathitrust.org/hathitrust\\_api](http://www.hathitrust.org/hathitrust_api)

Internet Archive hosts a very general collection of public domain texts while the HathiTrust focuses more narrowly, and in a deeper fashion, on content that is useful for researchers and their output.

#### *Media Restricted By Copyright*

There are a few projects that are actively preserving and mirroring content that is still restricted by copyright. The most well-known of these projects is the LOCKSS (Lots of Copies Keeps Stuff Safe) and CLOCKSS (Controlled LOCKSS) project from Stanford University. LOCKSS/CLOCKSS focuses primarily on scholarly journals that libraries subscribe to digitally. It acts as both a geographically distributed backup mirror and as a guarantee that the libraries still have access to the content in the event that it is no longer provided by the publisher.

While the HathiTrust will be preserving the scanned version of many scholarly journals the overlap will not be 100%. This is due to the fact that libraries are moving away from purchasing the print version of scholarly journals and are instead opting use the electronic-only access method. This saves the libraries both money and shelving space but also moves the burden of preservation to the publisher of the journals. Additionally, the HathiTrust will not be able to immediately begin displaying, in an Open Access method, the content of a publication if that publication is no longer available due to fundamental copyright restrictions. The LOCKSS/CLOCKSS program has made those agreements with the publishers and will be providing that access instead. Thus, the HathiTrust and LOCKSS/CLOCKSS programs are complimentary in their mission, not redundant.

#### **Moving Forward**

Because it provides a complimentary and effective service to the world of research libraries the HathiTrust is an important and critical part of the research library community's preservation effort. And with the unknown future of the Google Book Search program – if they will provide access indefinitely – this provides another alternative to accessing those materials. The obvious limitation of the HathiTrust collection is that it is only as good as the participating members; to improve it more libraries with digitized special collections must join the cooperative. I argue that any library in the process of digitizing their materials, whether internally or through a cooperative agreement with a third-party, should join the HathiTrust. Additionally, by having more books in the HathiTrust, and thus available for examination by the Copyright Review Management System, we will have more books with a more certain copyright status determination. The CRMS is only a partial solution to the Orphan Books issue, but it is an important step in the right direction. By joining the HathiTrust, participating libraries will vastly improve the wider academic community's access to knowledge and ability to perform extensive and ground-breaking research.