HathiTrust Public Services Task Force Phase 1 Investigation

A Report Submitted to the User Services Council, Rutgers University Libraries

Rutgers University Libraries HathiTrust Public Services Task Force
(Prepared by the Phase 1 members and compiled by the Report Working Group)

May 21, 2015

HathiTrust Public Services Task Force Phase 1 members
Stephanie Bartz, Natalie Borisovets, Janet Brennan Croft, Bob Cupryk, Francesca Giannetti, Triveni Kuchi, Janice Pilch, Tibor Purger, Jane Sloan, Gracemary Smulewitz, Chris Sterback, Mary Beth Weber, Tao Yang (chair)

HathiTrust Report Working Group (a subgroup of the Task Force)
Janice Pilch (chair), Mary Beth Weber, Tao Yang
Introduction

The HathiTrust Public Services Task Force was formed in November 2014 by Melissa Just, Associate University Librarian for Research and Instructional Services (RIS), to investigate the potential public services impact of joining HathiTrust on the Rutgers University Libraries (the Libraries). The Task Force was charged to examine several issues including discovery, interlibrary loan, reserves, accessibility, and preservation, to explore areas of interest to our users, and to identify additional opportunities, with reference to the recent survey of HathiTrust service integration conducted by Brown University.

The Task Force operates under the auspices of the User Services Council (USC) with support from the AUL for Research and Instructional Services. The group will provide periodic reports to USC to help the Libraries prioritize the integration of HathiTrust resources into our public services. The Task Force is also charged with promoting HathiTrust membership benefits to library faculty and staff and library users. The original Task Force membership included: Stephanie Bartz, Natalie Borisovets, Janet Brennan Croft, Bob Cupryk, Francesca Giannetti, Triveni Kuchi, Janice Pilch, Tibor Purger, Jane Sloan, Gracemary Smulewitz, Mary Beth Weber, and Tao Yang who serves as chair. Chris Sterback was added to the Task Force in January 2015.

The Task Force held its kickoff meeting on December 16, 2014. Following that meeting, the group began investigating select topics and prepared a list of questions to address all areas of investigation in its charge. At the January 16, 2015 meeting, the questions were finalized and distributed to individuals or small groups; their responses to the questions were submitted a month later and became the basis of this report.

The Task Force also organized a visit from Jeremy York, Assistant Director of HathiTrust, to clarify questions and to officially launch HathiTrust at Rutgers with a public event. The meeting with York and the event, “Rutgers, please meet HathiTrust,” took place on February 24, 2015, both of which provided additional information for this report. The above activities are referred to as Phase 1 of the Task Force.

Based on the information gathered in Phase 1, the Task Force adopted an action plan for Phase 2 on March 18, 2015. The goals of Phase 2 are to complete the investigation and to implement the essential services related to HathiTrust. To implement Phase 2 activities, the Task Force divided into five working groups: Access Working Group, Catalog and Discovery Working Group, Preservation Working Group, Report Working Group, and Training, Documentation, and Promotion Working Group. Several new members were added to the Task Force in Phase 2. This report was compiled by the Report Working Group.

Questions for Investigation

The following questions were raised by Task Force members as the most essential issues to investigate in Phase 1:

Access Issues

1. What can be accessed in HathiTrust legally and contractually, and why?
2. What is the nature of the relationship between HathiTrust and the Internet Archive/Digital Public Library of America? How does it affect the HathiTrust content?
3. Is there a way for onsite guest users to download the full text of public domain works without logging in?
4. Is there a way to provide our visiting scholars with remote access? What about faculty emeriti?
5. What is the Service for Users with Print Disabilities? How do we meet the “Requirements for Access” specified by HathiTrust, particularly a defined process, a designated proxy, and a static IP address?
6. What is the potential impact of HathiTrust on reserve services?
7. For the databases A-Z list on the library website, what disciplines should consider HathiTrust a core, related, or supplementary resource?

**Discovery Issues**

8. Can we use EDS as the discovery layer for HathiTrust? What are the implications?
9. Should ILL staff check HathiTrust before processing an internal ILL request for a Rutgers user? If yes, what should we do if an item is found in HathiTrust? According to the Brown survey, most libraries will return the request to the user, either asking if the print is still needed or providing a link to the HathiTrust version (and cancelling the request).
10. Are we allowed to fulfill external ILL requests with HathiTrust versions?
11. Do we want to load HathiTrust records into the library catalog? If not, why?
12. Do we recommend using the HathiTrust API in VALE Open Source ILS (VALID)?

**Preservation and Collection Management Issues**

13. What do we do if a lost/damaged item is available in HathiTrust? Do we still repair/replace the analog item or withdraw it?
14. What is the potential impact of HathiTrust on government documents collection and services?
15. How will HathiTrust affect our decisions to send items to the Annex?

**Training and Documentation Issues**

16. What do we know about the non-English materials in HathiTrust?
17. What is non-consumptive use? How does it work?
18. What is the HathiTrust Research Center? What opportunities available to the Rutgers community? How do we work with our users to take advantage of these opportunities?
19. How do we market HathiTrust to our users?

The following is a compilation of answers representing work done in the Phase 1 investigation.

**Access Issues**

1. **What can be accessed in HathiTrust legally and contractually, and why?**

HathiTrust makes available books, serial issues, and other works digitized from libraries around the world in accordance with copyright law (in defining public domain versus copyrighted works), contracts (such as third-party agreements signed by Google and its partner libraries), and permissions from rightsholders (such as through Creative Commons licenses).

**Public domain and uncopyrightable works.** Users affiliated with HathiTrust partner institutions can download full PDFs of all public domain works and print them. Users who are not affiliated with HathiTrust partner institutions can download single-page PDFs of all public domain works and full PDFs
of public domain works that are not subject to third-party agreements that restrict access. For example, restrictions set in the digitization agreements signed by Google and its partner libraries require HathiTrust to prevent systematic download and dissemination of Google-digitized materials. For this reason, HathiTrust restricts the general public from full access.

**Copyrighted works.** Both users affiliated with HathiTrust partner institutions and unaffiliated users can download full PDFs of copyrighted works made available under Creative Commons licenses or for which permissions otherwise have been granted, and to print them. Other copyrighted works are searchable but may not be read, downloaded, or printed.

Exceptions:

1) If the use falls under the specific area of HathiTrust Print Disabilities Access whereby eligible patrons can receive special access to materials that are held currently, or were held previously, in the partner institution’s library, as indicated through print holdings information submitted to HathiTrust. A detailed description of the HathiTrust Print Disabilities Access Service is available at [http://www.hathitrust.org/accessibility](http://www.hathitrust.org/accessibility).

2) If a partner institution’s work has been damaged or lost and the institution needs full access to a digital replacement copy.

**Definition of public domain.** Because of differences in national copyright laws, public domain is defined differently within each country. HathiTrust treats U.S. public domain as one category and international public domain as another category. Google and HathiTrust use the same practice for identifying the public domain for the published works they hold. They base the U.S. public domain on U.S. law, as works published before 1923, and they base the international public domain on an arbitrary standard, as a period of 140 years from the date of publication. The standard is a moving wall. In 2015, the dividing year for international public domain is 1875. Also included in the public domain are U.S. federal government documents. There is an explanation on the website: [http://www.hathitrust.org/copyright](http://www.hathitrust.org/copyright).

2. **What is the nature of the relationship between HathiTrust and the Internet Archive/ Digital Public Library of America? How does it affect the HathiTrust content?**

HathiTrust was formed in 2008 as a repository for books and other materials digitized by Google on the basis of agreements with Google partner libraries beginning in 2004. Many of the Google partner libraries agreed to submit their library digital copies to HathiTrust to form a single repository that complements Google’s undertaking that later became Google Books and serves as a public good. Both systems offer digitized books on the Internet, but HathiTrust provides additional content obtained from non-Google libraries, such as unique digital collections, works from institutional repositories, and native born-digital materials.

The Internet Archive (IA) was founded in 1996 “to build an Internet library.” It was a competitor to Google as a member of the Open Content Alliance (OCA), a consortium formed to create a permanent, publicly accessible archive of digitized texts. Announced in October 2004, the OCA included Yahoo, IA, the University of California, the University of Toronto, plus Microsoft, which joined in October 2005 as part of its Live Book Search project but later withdrew in May 2008. The OCA sought to digitize copyrighted works only after asking and receiving permission from the copyright holder through an “opt-in” model. Since then, it has embarked on other types of digitization and book lending activity that go beyond an opt-in model. Some HathiTrust partners are also Internet Archive partners, and are depositing
their content digitized by the Internet Archive into HathiTrust. There is no known direct relationship between IA and HathiTrust at this time.

The Digital Public Library of America (DPLA) was also founded in reaction to Google activity, specifically to the Google Book Settlement, to create a digital library to serve the public. It has its origins in an October 2010 meeting in Cambridge, MA at which forty leaders from libraries, foundations, academia, and technology projects agreed to work together to create “an open, distributed network of comprehensive online resources that would draw on the nation’s living heritage from libraries, universities, archives, and museums in order to educate, inform, and empower everyone in current and future generations.” See “History,” Digital Public Library of America, http://dp.la/info/about/history/.

Officially launched on April 18, 2013, the DPLA is based at Harvard University’s Berkman Center for Internet & Society, and is being undertaken in collaboration with Europeana, a similar effort by the European Commission. Its structure is based on linking to existing library digitized materials rather than on serving as a repository, and is organized into six workstreams led by a Steering Committee. The distributed network of online resources takes the form of “hubs” constituting a national network of over forty state and regional digital libraries and large digital libraries that brings together digitized and born-digital content into a single access point for end users, and serves as an open platform.

There may be some overlap among HathiTrust, IA, and DPLA collections, but contributing libraries and archives make efforts to minimize overlap. DPLA aims to provide full access rather than restricted access to content it accepts, making it very different from HathiTrust.

3. **Is there a way for walk-in users to download the full text of public domain works without logging in?**

Full access to HathiTrust materials, including access to download all public domain works, is by login only. Walk-in users will be provided with a means to log in, with the “library walk-in option” provided by HathiTrust. There are two ways to implement the library walk-in option. The Libraries could create guest credentials with which public users log into Shibboleth. The other means would be determined by IP address that would supply credentials to HathiTrust via Shibboleth. The second approach is deemed to be more workable. The drawback of this approach is that if no credentials are passed to Shibboleth, the connection will not be personalized and the walk-in users will not be able to build their own collections. Please see the webpage about the Shibboleth Login (http://www.hathitrust.org/shibboleth) for more information.

4. **Is there a way to provide our visiting scholars with remote access? What about faculty emeriti?**

At Rutgers, visiting scholars are frequently, but not always, described as guests in the documentation of the departments who host them. The group of four Access Services supervisors in the Libraries has planned a meeting with RU Connection to clarify exactly who they designate as visiting scholars and guests and why. We want to ensure that the user categories they assign conform to the database services we are allowed to offer per our contract terms. We are working towards better and agreed-upon definitions that will serve us well not just with HathiTrust but with the upcoming move to OLE (Open Library Environment).

Currently, visiting scholars who are classified as guests and faculty emeriti can access library resources remotely by logging in with their library barcode and PIN. This works for everything except Shibboleth –
"member" and "alum" are the only accepted values for logging into HathiTrust through Shibboleth (http://www.hathitrust.org/shibboleth).

5. What is the Service for Users with Print Disabilities? How do we meet the “Requirements for Access” specified by HathiTrust, particularly a defined process, a designated proxy, and a static IP address?

This is a special service provided to eligible patrons, defined by HathiTrust as “an affiliated user at a Supporting Institution who has a print disability for which access to digital copies of library books is a reasonable and appropriate accommodation.” Print disability is defined by HathiTrust as “a disability – such as a visual impairment, learning disability physical disability, or other disability – that impedes a person’s ability to access print in the standard way.”

Through a designated proxy, eligible patrons at HathiTrust partner institutions may receive special access to in-copyright materials held in HathiTrust. The materials must be held currently or have been held previously by the institution’s library, as indicated through print holdings information submitted to HathiTrust. A proxy is “a representative designated by a HathiTrust Supporting Institution who accesses in-copyright content on behalf of an Eligible Patron.” The Libraries will need to determine a proxy and a workflow for providing the copies to Eligible Patrons.

To benefit from this service, HathiTrust requires institutions to have a defined institutional process for certifying eligible patrons. The Rutgers Office of Disability Services has a certification process in place, based on the Section 121 copyright exception and related laws and regulations, the Americans with Disabilities Act, and the Rehabilitation Act of 1973. The Libraries will rely on that office to confirm eligibility in each instance. HT requires The Libraries to confirm disability status but does not require us to provide them with a copy of official disability eligibility documentation.

HathiTrust’s emphasis on print disability stems in part from the significant international discussion on accessibility to copyrighted works that resulted in the Marrakesh Treaty to Facilitate Access to Published Works for Persons Who Are Blind, Visually Impaired, or Otherwise Print Disabled, adopted by the World Intellectual Property Organization (WIPO) on June 27, 2013 and signed by the U.S. on October 2, 2013 (http://www.wipo.int/dc2013/en/). The U.S. government is currently discussing its implementation of this treaty.


“HathiTrust Print Disabilities Access: Service Description and Requirements” are available at https://docs.google.com/document/d/1eU8n0IfBFdyxNhtDWTCqzJr0ucnA9karp-tviQDuawA/edit?pli=1

For the past two years, Rutgers has supplied HathiTrust with files of our print holdings and will continue to provide this information, which is required to “provide a foundation for the expansion of legal uses of materials in HathiTrust by partner institutions, including special access to materials under Section 108 of U.S. copyright law, and access to in-copyright works for users who have print disabilities.” The files we supplied contained complete listings of our print single and multi-part monographs that have OCLC numbers, including the current status of the items; and serial titles that have OCLC numbers, by title but not at the item level. For each holding we indicated one of these three statuses:

CH = Current Holding
6. **What is the potential impact of HathiTrust on reserve services?**

If a reserve book is a “full view” title in HathiTrust, there is no need to place the physical copy on course reserve nor should we place an e-copy on reserve. We should simply provide a link to the title in HathiTrust and instruct students to access the material directly using their own login. We will provide links in the same manner that we currently provide for e-books used for reserves. Both the instructor and the students must be provided with instructions on how to access and download titles. Rutgers students who have logged in from another country using their NetID will be able to access HathiTrust materials as if they were in a U.S. location.

Related to library support for instruction, some instructors may be interested in the Collection Builder feature in HathiTrust. Using it, they can create HathiTrust collections for their own classes or use existing public collections to teach.

7. **For the databases A-Z list on the library website, what disciplines should consider HathiTrust a core, related, or supplementary resource?**

Content in the HathiTrust Digital Library is relevant to a wide range of humanities and social sciences disciplines, especially English, History, and foreign language and area studies programs. Individual subject specialists must determine if HathiTrust is a core, related, or supplementary resource for their disciplines. Here is a summary of their decisions:

- African Studies: related
- East Asian Studies: related
- English Language and Literature: core
- French: related
- General Arts and Humanities: related
- General Social Sciences: related
- Government Publications: core
- History-British Isles: core
- History-North America: core
- Italian: related
- Latin America, Caribbean, Spanish & Portuguese Studies: related
- Linguistics: supplementary
- Political Science: supplementary
- South Asian Studies: related

**Discovery Issues**

8. **Can we use EDS as the discovery layer for HathiTrust? What are the implications?**

HathiTrust is available in EBSCO Discovery Service (EDS) Central Index. However, there are two components of the resource defined in the EDS Wiki, HathiTrust Open Access and HathiTrust restricted. After surveying the EDS users’ listserv for information on other institutions’ implementations, we have
found that libraries have slightly different interpretations of the restricted component. EBSCO’S liaison who is assigned to the list offered the following explanation:

“the ‘HathiTrust – Restricted Content’ CustomLink includes works that are currently still in copyright, while the Open Access collection contains materials that are out of copyright. The following is taken from the HathiTrust website:

‘The initial focus of the partnership has been on preserving and providing access to digitized book and journal content from the partner library collections. This includes both in copyright and public domain materials digitized by Google, the Internet Archive, and Microsoft, as well as through in-house initiatives.’”

After conducting several searches in EDS and limiting them to HathiTrust content, we have not had consistent search results. In some cases the full-text access was not available when it was expected. HathiTrust can be represented in EDS, but it is unclear which avenue we will pursue. We need further clarification as to what is discoverable and a better understanding of the construct for custom links in EDS for HathiTrust.

Additionally, we have not promoted EDS for monograph discovery. Patrons may find some monograph collections in EDS because the provider included the monographs in its platform for journals.

9. Should ILL staff check HathiTrust before processing an internal ILL request for a Rutgers user? If yes, what should we do if an item is found in HathiTrust? According to the Brown survey, most libraries will return the request to the user, either asking if the print is still needed or providing a link to the HT version (and cancelling the request).

Yes. If a record is found, we have a standard text that is used when an e-book is available, cancelling the request. If the patron indicates that the e-book item does not meet his or her needs, we will request the item in a physical format.

10. Are we allowed to fulfill external ILL requests with HathiTrust versions?

No, we may not lend HathiTrust volumes to fulfill external ILL requests. The access level that HathiTrust provides to partner institutions is available only to authorized users. A certain level of basic access to HathiTrust materials is available to the general public. Rutgers access is limited to our users. For example, HathiTrust licenses the Google-digitized volumes under specific terms to which member institutions agree and we may not violate the terms of these licenses.

11. Do we want to load HathiTrust records into the library catalog? If not, why?

This question is under discussion. We most likely will not load the records. The reasons for this are:

- There is not really a way to “load” records. We may do better by providing access via the API.
- The HathiTrust is a dynamic system. We would need to engage in constant updates to our records to keep the catalog accurate.
- Access levels vary by work, and may change with the legal landscape and with revisions to HathiTrust records.
• If we decided to load the records, it would make sense only to load records for works that are available for reading and downloading (public domain works and works available by permission of the rightsholder).

12. Do we recommend using the HathiTrust API in VuFind and VALE Open Source ILS (VALID)?

This is still under discussion. The Libraries will implement VuFind in summer 2015 and the HathiTrust API has been installed on the test implementation of VuFind. There is a longer timeline to implement VALID. Whether to use the HathiTrust API is a question that the VALE consortium needs to consider, and is not limited to Rutgers. Currently Rutgers and Princeton are the only VALE members who are HathiTrust partners. Users from other institutions may only access public domain resources, which they can also do by going directly to HathiTrust on their own.

Preservation and Collection Management Issues

13. What do we do if a lost/damaged item is available in HathiTrust? Do we still repair/replace the analog item or withdraw it?

It is lawful for the HathiTrust to provide a digital replacement copy of an item in our collections that is damaged, deteriorating, lost, stolen, or in an obsolete format, provided that we make a reasonable effort to determine that an unused replacement cannot be obtained at a fair price, relying on the HathiTrust In-print Status Review procedures and the Libraries’ guidelines for determining a fair price by discipline (to be provided). We must also ensure that the digital replacement copy is not made available to the public in that format outside the premises of the library, following Section 108 of the copyright law, under which HathiTrust operates under the terms of a settlement agreement from January 2015. HathiTrust confirms that we own a copy of the item through print holdings information submitted to them.

As previously noted under question 5, Rutgers has supplied HathiTrust with files of our print holdings for the past two years and will continue to provide this information, which is required to use HathiTrust for preservation purposes. We would need to withdraw the original item from circulation but not discard it.

Obtaining a digital replacement copy of a copyrighted work would enable us to provide our users with the fullest possible access and use of the digital replacement copy on the library premises (full-text reading and printing capability, such as with a print book). The Libraries are discussing a pilot project to test the In-print Status Review procedures. The proposed workflow is:

- Conduct In-print Status Review on each work, using
  - Using HathiTrust documentation, https://docs.google.com/document/d/1YimevfZlhLKGunWvVnV4Tk21EjNnU-hbwjhRU7f71GM/edit?pli=1
  - The Libraries’ guidelines for determining fair price by discipline
- Withdraw the items from the catalog
  - Correct the catalog entries
- Send a text list of about 100 items that will be part of the pilot to Jeremy York, providing HathiTrust identifiers (HathiTrust volume IDs)
- Verify that the works are in need of preservation and how they are damaged, deteriorating, etc.
- Add a notice on the file that it is a preservation copy as an instruction to users (to be confirmed)
• Provide access on premises only and/or NETID only (downloading access to be confirmed)

The number of simultaneous users who can access a partner institution’s digital replacement copy is determined by the number of print copies held (or previously held) in the library system. If a library system only has one print copy, only one user at a time will be able to access the digital copy.

14. What is the potential impact of HathiTrust on government documents collection and services?

Many of the benefits of HathiTrust with respect to government documents are the same as those for the general collections. Aside from the potential for full-text access to materials not otherwise available electronically, or in some cases, not available in any format, HathiTrust has the potential to be a valuable tool for weeding, preservation, and associated activities. The more unique benefit for government documents is the HathiTrust’s potential to serve as a catalog for much of the uncataloged material in the Libraries collections.

Although the concentration of documents in HathiTrust seems to be largely U.S. federal, a variety of other government entities are also included. A recent search for material authored by the League of Nations on the topic of refugees, for example, identified more than 80 titles. While many have a “Limited (search-only)” status, the result set can serve as a finding aid and partial catalog for the uncataloged League of Nations material held at Rutgers. The same holds true for documents from other international organizations, states, foreign governments, and other government entities. Since a large percentage of the government documents collections at the Libraries are uncataloged, HathiTrust will aid in searching and identifying documents that have been cataloged in other institutions.

The ongoing development of the HathiTrust Government Documents Registry, intended “to include metadata for the comprehensive corpus of U.S. federal documents,” has additional implications for Rutgers collections. The registry, which is expected to include 1.5 to 2 million items, can inform digitization, deaccessioning, and other collection development and management projects such as shared print initiatives. See “Creating a Registry of US Federal Government Documents,” HathiTrust Digital Library, http://www.hathitrust.org/usgovdocs_registry.

With more than 600,000 U.S. documents available as full-text in the HathiTrust repository as of January 2015, there is extensive overlap between the electronic collection and the tangible collections at Rutgers. See “Help – HathiTrust US Federal Government Documents Initiative,” HathiTrust Digital Library, http://www.hathitrust.org/help_usgovdocs. The depository materials split between Alexander Library and the Library of Science & Medicine represent a selection rate of more than 95% of the material made available through the Federal Depository Library Program. The bulk of the material selected has been retained. The Libraries also include additional depositories at the Dana and Robeson libraries, plus the Camden and Newark law libraries, albeit with lower selection rates.

Although U.S. Federal documents are generally classified as public domain or “public domain-Google digitized,” there is a subset of materials that are restricted by copyright. Careful attention will need to be paid to public accessibility of material before consideration is given to withdrawing damaged or duplicate material while adhering to regulations applicable to the depository collections.

In addition to providing a much needed catalog of a large percentage of the extensive uncataloged U.S. documents collections, the Registry may also present opportunities for librarians and staff at Rutgers to engage with CIC colleagues. A recent report from the Government Documents Initiative Planning and Advisory Working Group includes recommendations such as conducting additional analysis on the
existing corpus of documents in the HathiTrust and identifying additional cataloged content not yet in the repository (see http://www.hathitrust.org/documents/HathiTrustGDIPAWGwhitepaper.pdf). These and other recommendations represent potential areas of involvement for Rutgers documents librarians and staff.

Additional considerations:

- Once the API and user interface have been developed, will the HathiTrust Government Documents Registry be publicly available as a searchable resource separate from the existing HathiTrust interface?
  
  Yes. The Government Documents Registry is a separate resource from the HathiTrust catalog.

- Will Rutgers be asked to contribute content in the form of scanned government documents or government documents to be scanned?
  
  No, Rutgers is not required to contribute content.

**15. How will HathiTrust affect our decisions to send items to the Annex?**

This question is under discussion. We need to consult the Library Resources Council and selectors to develop policies and guidelines.

---

**Training, Documentation, and Promotion Issues**

**16. What do we know about the non-English materials in HathiTrust?**

The HathiTrust Digital Library includes materials from all over the world in several hundred languages. The top 15 languages are English, German, French, Spanish, Russian, Chinese, Japanese, Italian, Arabic, Latin, Portuguese, Polish, Dutch, Hebrew, and Hindi and about half of all HathiTrust titles are in English (for more detailed, up-to-date information, see http://www.hathitrust.org/visualizations_languages). The language breakdowns of public domain titles are slightly different (see http://www.hathitrust.org/visualizations_languages_pd). On both webpages, users can click on the language name to browse the titles in that language, which obviously is a useful feature for the languages with smaller amount of books in HathiTrust. There are two ways to search for non-English materials:

1. Use the advanced catalog search in HathiTrust and limit *by Language* (search terms not required).
2. Use the advanced full-text search and limit *by Language* (requires search terms).

**17. What is non-consumptive use? How does it work?**

Non-consumptive use means data mining. The term came from the original Google Book Settlement dated October 28, 2008 (that was subsequently rejected). The settlement agreement defined “Non-Consumptive Research” as “research in which computational analysis is performed on one or more Books, but not research in which a researcher reads or displays substantial portions of a Book to understand the intellectual content presented within the Book” and offered several categories.
Under that agreement, non-consumptive research was a service available to fully participating libraries under certain conditions outlined in the settlement. The so-called Research Corpus used for the non-consumptive research was the basis for forming the HathiTrust. In the wake of the rejected settlement, the permissibility of data mining the Research Corpus was resolved through a fair use determination by the U.S. District Court for the Southern District of New York on October 10, 2012, affirmed by an appeals court on June 10, 2014.

The HathiTrust Research Center developed the framework of the Data Capsule to support the analytical investigation of the copyrighted corpus in HathiTrust while prohibiting data (i.e. the text of the original works) from leaving the capsule. The HTRC Data Capsule provides researchers with their own virtual machines that run within the HTRC domain. Researcher can configure the VM as they would their own desktop with tools of their choice. After the researcher is done, the VM switches into a “secure” mode, in which network and other data channels are restricted in exchange for access to the data being protected. Results are emailed to the researcher.

18. What is the HathiTrust Research Center? What are the opportunities available to the Rutgers community? How do we work with our users to take advantage of these opportunities?

The HathiTrust Research Center (HTRC) is a collaboration between Indiana University and the University of Illinois at Urbana-Champaign, together with the HathiTrust Digital Library. It was established in 2011 to enable computational research across the HathiTrust’s vast collection of published works. The systems and tools available through the HTRC are free and open, although they do require the creation of user account(s). From the HTRC portal (https://htrc2.ntl.indiana.edu/), it is possible to run analytic algorithms against the HathiTrust volumes that are in the public domain (~2.5 million).

The Workset Builder is a search interface for the HathiTrust public domain corpus; search results can be saved as a “workset”: a collection of volumes on which the text mining algorithms are performed. Users can build their own workset and make it public or private, or reuse another that has been made public.

In addition to the main HTRC services (Portal and Workset Builder), there is a sandbox stack with the same tools. The sandbox runs against non-Google scanned content (about 260,000 volumes). The advantage of the sandbox is that the index and Data API can be accessed directly, permitting the use of the researcher’s algorithms.

Bookworm - http://sandbox.htrc.illinois.edu/bookworm/
Of the HTRC tools, Bookworm is the most user-friendly. It visualizes language usage trends as a simple time plot.

Data Capsule - https://wiki.htrc.illinois.edu/display/COM/HTRC+Data+Capsule
The HTRC Data Capsule permits “non-consumptive” research on HT’s use protected texts (works in copyright).

Single Interface for Music Score Searching and Analysis project (SIMSSA) - http://simssa.ca/
SIMSSA researchers use Optical Music Recognition (OMR) software to transform digital images of music into searchable representations of music notation.

This partnership seeks to produce the first large-scale cross-cultural study of the novel using quantitative methods.
Workset Creation for Scholarly Analysis - http://worksets.htrc.illinois.edu/worksets/

An immediate objective for HTRC is to allow scholars to collect items for computational analysis. But the often rudimentary, print-based metadata of the HT does not adequately support the creation of scholarly worksets. A goal of the WCSA is to enrich the metadata in HT and augment it with URIs to leverage discovery and sharing through external services.


The HTRC has been holding an annual UnCamp since 2013. "The UnCamp is part hands-on coding and demonstration, part inspirational use-cases, part community building, and a part informational, all structured in the dynamic setting of an un-conference programming format.”

The HTRC Wiki is very useful for wrapping one's head around their activity: https://wiki.htrc.illinois.edu/display/COM/HathiTrust+Research+Community+Pages.

The HTRC’s computational tools will be of particular relevance to Rutgers DH practitioners. For example, students in a Literary Data course (http://rci.rutgers.edu/~ag978/litdata/syllabus.pdf), developed a basic competence in analyzing data using R (statistical programming language) as they engaged in critical writing and discussion on debates in humanistic computational methods. A tool like the Data Capsule provides an opportunity for these students to perfect the skills learned in class.

Other means to build awareness of the HTRC among our users include the development of an HTRC libguide, web tutorials, and in person workshops and/or hackathons. The HTRC UnCamp is a highly intensive opportunity for involvement that should be promoted in the Rutgers DH community.

19. How do we market HathiTrust to our users?

While promoting the HathiTrust event in February and the HathiTrust workshops in April, we identified and used the following library and university channels:

- Library listservs –rul_everyone and rul_faculty
- Library liaisons sent the information to teaching faculty and students
- News story on the Library website (for the February event only)
- University Faculty and Staff Bulletin
- Rutgers Events Calendar
- Research Communications from the Office of Research and Economic Development (for the April workshops only)
- Slides being run on library electronic displays to promote workshops and HathiTrust itself

The promotional efforts paid off since faculty and students in School of Arts and Sciences (six departments), School of Information and Communication, and Graduate School of Education in New Brunswick attended the February event and about half of the 20 plus participants at the April workshops were from outside of the Libraries. We are also considering sponsoring a HathiTrust Week and advertising HathiTrust in the Libraries’ communications to donors.
Additional Questions Raised in Phase 1:

- **Which group in the Libraries will be responsible for monitoring the ongoing developments at HathiTrust and share the information widely with staff and users?** Issues on the horizon include collecting open access university press books and non-text formats in HT.
  
  This issue is under discussion.

- **Why are Rutgers users identified as "member" rather than by name (or something unique) since we log in with our NetIDs?** It says "Hi member" in the bar at the top and when any Rutgers member creates a collection, it's identified with "Owner: member (Rutgers University)." Other institutions have more specific identifications (e.g., Owner: Willis Goth Regier (University of Illinois at Urbana-Champaign)).
  
  This issue is under discussion.

- **Will the Libraries consider contributing content to HathiTrust?**

  The Libraries are not required to contribute locally digitized content to remain members of HathiTrust but will consider contributing in the future.