HathiTrust Working Group “Green Paper”
Yale University Library

15 June 2011

(edited April 2012 for external audience)

Executive Summary

The HathiTrust (HT) working group was charged to explore where Yale University Library (YUL) is in the HT process and relationship and to begin to formulate some explicit strategies with respect to this partnership by studying HT's programs and potential offerings. The results are presented in this "Green Paper" containing recommendations and policies for discussion by YUL management and staff.

Recommendations

The working group recommends continued partnership in HT because it provides opportunities and benefits to a broad range of YUL services for students, faculty, researchers and staff in a wide spectrum of subject areas. Additionally, we recommend that YUL staff engage in the identification and development of HT planning and service enhancements. The group further recommends that this partnership be considered as an integral part of the Library's infrastructure and that funding for it be secured in such a way that annual contributions from individual libraries, departments, collections and curators are not required.

Benefits and Opportunities Summary

Continued partnership in HT provides:

- Preservation and access services for our own digital assets and access to a large and growing corpus of digitized book and serial content.
- Value through copyright work and related polices regarding access to deposited digitized content.
- A voice in shaping future development and services within this important community.
- Significant potential for the development of new services.

Report Organization

- **Introduction – HathiTrust Working Group**
  - Charge Summary
  - Working Group
  - Process
  - Background – What is HathiTrust
- **Summary Recommendation**
  - Partnership Recommendation
  - Participation Recommendation
  - Funding Recommendation
- **Analysis**
  - HathiTrust Partner Privileges
  - Costs
  - Immediate Benefits
    - Copyright

The HathiTrust Working Group would like to thank Ann Okerson, AUL for Collections and International Programs, for this opportunity to investigate HT and make recommendations about its future within YUL.

Introduction – HathiTrust Working Group

Charge Summary

In summer 2010, YUL joined the HathiTrust (HT) as a Partner Library. At that time, YUL paid associated fees through 2012 to cover the costs of loading, storage, access, and potential long-term preservation for the 29,000 digitized books digitized. Participating in HT provides YUL a path to launch strategic conversations about the benefit to Yale (and the larger library community) of participating in a growing mega-collection of digital books and serials (expanding beyond print in the future), in a large multi-library partnership. In joining, YUL believed this effort to have potentially large, positive impact for Yale students, faculty, researchers and staff.

The HT working group was charged to explore where YUL is in the HathiTrust process and relationship and to begin to formulate some explicit strategies with respect to this partnership by studying HT's programs and potential offerings. The results are presented in this "Green Paper" containing recommendations and policies for discussion by a YUL library management and staff.

Working Group

The working group was sponsored by Ann Okerson, AUL for Collections and International Programs. Members of the working group are:

John Gallagher, Medical Library
Kevin L. Glick, Manuscripts and Archives
Julie Linden, Social Science Library
Scott Matheson, Web, Desktop and Digital Services
Haruko Nakamura, East Asia Library
Audrey Novak, Library IT Office
Roberta Pilette, Preservation Department
Lidia Uziel, Humanities Collections and Research Education

Process

The HathiTrust Working Group met from March – June 2011. Our analysis process included:

- A review of HathiTrust policies, procedures, progress and services as documented on the organization's website (http://www.hathitrust.org/about).
- A telephone interview with Jeremy York, Project Librarian, HT.
- Comparison of Orbis content against the HT database using the HT APIs.
- Data collection and analysis by Preservation Dept. staff April 1- May 28th regarding available online, digital, full-text preservation replacement copies.

Background - What is HathiTrust

HathiTrust began in 2008 in order to build a digital archive that would provide access and preservation services for library materials digitized from its partners' collections by Google, the Internet Archive (IA) and Microsoft. The founding institutions include the thirteen universities of the Committee on Institutional Cooperation, the University of California system, and the University of Virginia.

Since 2008 HT has expanded to include digitized collections beyond the original print content digitized by Google, IA and Microsoft. In HT's first three years, the partnership has grown to a current total of 52 individual libraries. As of June 2011, the HT database had 4,794,000 book titles and 213,000 serial titles for a total of 8,793,500 volumes. Of those, 1,233,000 titles (2,398,500 volumes) or approximately 27% of the total were in the public domain.

From its start, a goal of the HT was to reduce the long-term costs of these services through the implementation of a shared, large-scale storage infrastructure. Additionally, the archive was designed as an open, collaborative development environment for partners to improve existing, and develop new, access and collection services. HT's current and developing services include:

- A public discovery interface
- An application for reading, downloading, and interacting with (e.g., zooming and rotating) texts and images in HT
- Full-text search of the entire repository and the development of data mining tools for HT, and use by HT of other analysis tools from other sources
- Ingest mechanisms for digitized content submitted by partner libraries with format validation and error-checking
- APIs to access HT data and metadata and integrate them into local systems
- An accessible interface that uses strategies to facilitate navigation and use by users with print disabilities
- An application that permits individuals to create public and private collections
- Compliance with required elements in the Trustworthy Repositories Audit and Certification (TRAC)[1] criteria and checklist
- Support for formats beyond books and journals

In October 2011 representatives from the 52 partner institutions will participate in a constitutional convention that will define the second phase of HT governance and sustainability. HT will be asking institutions to commit to partnership for a five-year period.

History and Partnership - http://www.hathitrust.org/partnership
Mission and Goals - http://www.hathitrust.org/mission_goals
Functional Objectives - http://www.hathitrust.org/objectives

[1] TRAC http://www.crl.edu/sites/default/files/attachments/pages/trac_0.pdf defines the criteria, delineates the process, identifies the required documentation, and provides the tools and methodologies for assessing and determining whether a digital repository meets the requirements for sustainability and accountability that will certify it as trusted.

Constitutional Convention - http://ns.umich.edu/htdocs/releases/story.php?id=8121)

Summary Recommendations

The working group concludes from its analysis of HT content, services and objectives that continued partnership represents good value for the cost. A broad range of Library services for students, faculty, researchers and staff in a wide spectrum of subject areas and disciplines will benefit significantly from YUL's continued partnership in HT. We present three broad recommendations in this paper related to partnership, participation and funding.

Partnership Recommendation

Partnership in HT provides:

Preservation and access services for digitized book and serial content. HT has built a sustainable model for services based on the openness of its systems and services, and its reliance on broadly accepted standards and best practices for accessing, archiving and preserving digital content. The preservation services are TRAC compliant, and as such they provide a higher level of preservation protection than will be available to YUL for some time. In addition, HT's consortial approach to digital preservation provides a more sustainable model than individual institutional efforts, even given the size and resources of an institution like Yale. The HT database is a large and growing digital repository that provides full-text online access to close to 250,000 print titles held by YUL and almost a million public domain titles that are not part of YUL's collections. HT access services include a discovery interface, page turner, full-text search, data mining and robust APIs.

Value through HT's copyright work and related polices regarding access to deposited digitized content. HT has made very real progress in exercising "the limitations to exclusive rights" granted under sections §107 (Fair-use), §108 (Reproductions by libraries and archives), and §121 (Reproductions for blind or other people with disabilities) of the Copyright Act. The cost savings associated with making HT full-text content available as print replacement copies could offset the annual cost of partnership.

Potential for the development of new services. HT's open framework, well-structured hardware environment and critical mass of content and partners provide a platform for collaborative and strategic development of new and enhanced services that directly support students, faculty and researchers, as well as libraries.

Participation Recommendation

YUL should have a voice in shaping future development and services within this community. HT is currently considered a potential "digital library of the future" and a large number of academic libraries with strong, existing digitization programs have already joined the partnership. Having a voice within this community is important. The consolidation of efforts and talents can have significant benefits related to addressing issues of formats, quality, usability and copyright; resolving bibliographic

ambiguity (e.g., supplying missing dates or correcting records); and extending APIs to enable additional access services.

Broad participation across YUL will fully optimize our partnership. The benefits and opportunities provided by HT will affect a wide variety of staff across YUL requiring the development of new skills and new ways to approach tasks as well as require the development of new YUL internal partnerships.

Funding Recommendation

The working group believes that HT serves several fundamental library objectives and should be considered as an integral part of YUL's infrastructure. We recommend that funding for continued partnership in HT should be secured in such a way that individual libraries, departments, collections and curators are not required to calculate and evaluate on an annual basis the value of HT to their individual units.

HT benefits the library in numerous significant and specific ways. For instance, it now provides a centralized storage solution for much of the book and serial content the library digitizes, as well as digital preservation and access solutions for these files. It makes available to Yale patrons digital access to content valuable for a broad spectrum of disciplines and subject areas. Searchable full-text content makes HT a powerful discovery tool for users. Over time, partnership in HT will influence collection development and management decisions the Library makes about its print collections, particularly related to print replacements for preservation and transfer to the Library Shelving Facility (LSF). The quantity of interlibrary loan requests made for Yale patrons may decrease when HT content is integrated into our local discovery environment. HT may also present the library with opportunities to make decisions that will ultimately result in efficiencies and savings as described in this paper.

Analysis

HathiTrust Partnership Privileges

Anyone can view public domain and open content volumes. Anyone can download full PDFs for the public domain/open content volumes that were not contributed by the Google Books digitization project (i.e., about 10% of the public domain content).

As a partner, YUL and its patrons would have the following additional capabilities:

- Login using Yale's Central Authentication Service (CAS)
- Full-text download of the public domain content from Google Books which represents 90% of unrestricted content in HT
- Collection Builder, which allows our users and staff to create and share public and private collections of HT content
- Ongoing full use of APIs. Currently the APIs are available to everyone, but in the future HT anticipates adding functionality that is available only to members. For example, use of the Data API to look at copyright restricted content in order to determine if it is an appropriate preservation replacement would be a function available to member libraries only

- Access to all content, public domain and copyright restricted, for persons with print disabilities
- Access to copyright materials for Section 108 uses (e.g., print replacement copies)
- The ability to add a Print-on-Demand link in HT to our content in Amazon

Costs

HT's current cost model is based solely on storage costs for content deposited with HT. The 2010 charge to partner libraries is $3.86 per GB per year. This model greatly benefits libraries like YUL with little deposited content.

A new cost model will take effect in 2013. The new model is benefits-based. It charges libraries for the shared cost of storing volumes that are currently or once were in their collections, regardless of which library digitized and deposited the volume in HT. Under the new model, fees for libraries with large deposits will decrease because other libraries who also hold the title will share the storage costs. Fees for libraries with collections that significantly overlap with the HT content and are not significant contributors will increase. HT's estimate of YUL's cost under the new model is $70,000/year assuming no significant increase in deposited content. Annual costs, however, will vary in future years depending on the overall size of the HT database, YUL's contributions, the amount of public domain content, the number of partners, and the overlap percentages between YUL's print holdings and the HT content. See Appendix A for projected annual costs given variable collection size and overlap percentages.

The new cost model depends on the successful implementation of a print-holdings database that contains information about the monographic and serial holdings in each partner library. This database will be used to determine an institution's overlap with HT digital holdings. It will also enable the use of HT content for preservation print replacements and services for people with print disabilities. Partner libraries will be required to refresh their information in the HT print-holdings database on a schedule to be determined. If HT is unsuccessful in establishing and maintaining the print-holdings database the new cost model will be revised.

Effective participation in HT will also require an investment of staff time and some local resources. Staff will need time to work with and provide input to HT committees and the Library will need to be represented at HT partner meetings. Training and workflow adjustments will also be required as HT is incorporated into the Library's existing processes.

Details about HathiTrust's current and proposed cost model are available here: (http://www.hathitrust.org/documents/hathitrust-cost-rationale-2013.pdf, http://www.hathitrust.org/help_new_cost_model)

Immediate Benefits

Copyright

HT allows YUL to exercise existing rights to give some of our patrons access to online, full-text works they will not have otherwise. It provides a platform for continuously expanding that access. There are

nominally increased access benefits for partners that are not available to non-partner libraries, but access to the rights management platform is a community good to which YUL should contribute.

HT has very clearly defined policies on copyright and is committed to ensuring that access to copyrighted content is provided only "where permitted by law or by the rights holder." The HT Rights Database is a critical tool maintained to ensure appropriate access to all items in its library. The rights information about each item in HT is stored, tracked and updated within this database.

The rights database enables a partner library to provide greater access to post-1923 digitized content than was previously practical. HT systems recognize traditional §107 fair use rights by displaying snippets and by creating a full-text searchable index. The partnership further allows YUL to exercise rights under §108 to provide copies of lost, damaged or fragile works in our collection (or previously in our collection) to our patrons. HT also allows for use of in-copyright materials for those with print disabilities as allowed by §121. This is an expansion of access and takes advantage of long-standing legislative provisions. The §108 and §121 benefits are only available to HT partners.

Direct Financial Benefits

The working group estimated real dollar values for the direct benefits of HT partnership. With assistance from Library Information Technology Office (LITO)[2], we derived statistics by comparing local content with HT content using the HT APIs.

Benefit	Analysis	Estimated Value
Print-replacement copies (historic)	Orbis has 85,617 items marked as lost, missing, withdrawn or damaged representing all formats. Of these 35,138 matches were found in HT. 6,802 matches are public domain content. Thus, 28,336 (33% of our historic lost, missing, withdrawn or damaged content) are restricted content that would become available as online full-text through HT.	$2,125,200 - $4,250,400 Low = 28,336 * $75 (circulation replacement cost) High = 28,336 * $150 (preservation reformatting cost)
Print-replacement copies (ongoing)	40% of volumes searched have copy in HT (based on a three-month survey)	$19,000/year is the conservative savings estimate in preservation reformatting

[2] The working group would like to thank Eric James and Roy Lechich of LITO for their work comparing Orbis records to the HathiTrust content.

ILL requests (outgoing)	5,022 outgoing requests were made in the past 12 months. Of these 684 matches were found in HT. 73 of the 684 matches are public domain content.	$3,650 (73 * $50/transaction)
"Extra-Orbis" full-text content (public domain, public domain in the US and titles that have been manually opened up through HT's copyright work)	HT contains 1,233,000 titles that are unrestricted. Of these 996,000 are not owned by YUL, and therefore not represented in Orbis.	Cannot be estimated

Management, Preservation, and Access to Digitized Content

Partnership in HT would allow YUL to more effectively and efficiently manage, preserve, and provide access to its digitized book, serial, and codex manuscript materials.[3]

Management: Currently, management of the products of YUL digitization is decentralized, not well organized, and handled differently from one project to another. There are several different siloed access systems for book-like digitized content, including CONTENTdm,[4] Greenstone,[5] and Fedora.[6] YUL has not adopted a single solution for all of this type of content. More importantly, there is currently no solution for managed storage that connects the master files to the use copies. Each unit is responsible for seeking its own solutions and managing its own materials, regardless of the resources it has to do so. This is a particular problem for units that have already undertaken much of this digitization -- special collection units like Beinecke Rare Book & Manuscript Library, Manuscripts & Archives, Lewis Walpole Library, Yale Divinity Library, Yale Medical Library, as well as International and Area Studies.

Preservation: There is currently no solution at Yale for the long-term preservation of digitized book, serial, and codex manuscript materials.[7] While there are significant efforts by YUL and Yale, working towards potential future solutions, HT has already made great strides in this area. HT has developed a

[3] Management and preservation of digitized content are different from the simpler concept of storage of that content. There are currently several solutions (Isilon, YUL virtual tape, etc.) for effective storage of YUL digitized content. Both *management* and *preservation* assume that the digital content is stored appropriately. However, *management* implies a higher level of maintenance activities necessary in order to protect and maintain accessibility of authentic copies of digital content. This includes actions like metadata creation, clear allocation of responsibilities for maintenance over time, transfer of data to new storage media on a regular basis, redundancy and geographic separation, and disaster planning. *Preservation* assumes that in addition to the full management of the digital content, a fuller set of technical metadata is created and maintained, while extra long-term preservation actions are performed on the content to make it accessible over the long-term, like migration, emulation, or encapsulation.

[4] Yale Digital Collections, http://digital.library.yale.edu/cdm/browse.php.

[5] The Yale Medicine Thesis Digital Library Project, http://ymtdl.med.yale.edu/.

[6] Arabic and Middle Eastern Electronic Library, http://sesame.library.yale.edu:8080/fedoragsearch/ameeltreeresult.

[7] There are several digital preservation activities that have been undertaken by Yale units, including YUL and Office of Digital Assets & Infrastructure (ODAI). However, none so far have focused on digitized book and book-like content. The preservation of this content, perhaps through HT, is just a part of the larger digital preservation solution at Yale.

set of preservation policies, established a strong infrastructure, and is committed to standards for both content and metadata. HT already produces regular checks on the integrity of stored content in the repository and has plans to include more preservation services in the future. The HT has already been certified to be a Trusted Digital Repository through an audit conducted by CRL under its *Trustworthy Repositories Audit and Certification Checklist* (TRAC) process.[8] Yale has not even begun to discuss TRAC certification for any of its own repositories. Therefore, participation in HT provides us with immediate access to a preservation repository for digital files of our scanned volumes.[9] In addition, the consortial approach to digital preservation development taken by HT provides a more sustainable model than individual institutional efforts, even given the size and resources of an institution like Yale.

Access: A key service provided by HT is access to a large corpus of digitized content. HT has developed a series of access services to that content that include a discovery interface, page turner, full-text search and data mining, and robust APIs.

Management, Preservation and Access Action Items

- Given the existence of content in HT and our continued partnership in the organization, guidelines for the selection of local content for future digitization projects should be established.
- HT should routinely be used as a preservation repository for our digitized print content.
- YUL should define criteria for the delivery of digitized print and identify when HT meets those requirements and when it is necessary to create and maintain local repositories, search interfaces, navigation and page turning applications. Local delivery through CONTENTdm, Greenstone and Fedora should be evaluated periodically as HT functionality expands.
- Making HT content easily discoverable by our students, faculty, researchers and staff should be a priority effort at YUL.[10]
- About 50% content in HT will also be available as print-on-demand (POD) from Amazon. We should implement a direct link to the Amazon content from HT.

---

[8] The audit was conducted November 2009 to December 2010 with reference to generally accepted best practices in the management of digital systems. HT was rated in the categories of: Organizational Infrastructure; Digital Object Management; and Technologies, Technical Infrastructure, Security. The resulting ratings for the most part reflect robust systems and sound processes.

[9] CRL HT Audit report: http://www.crl.edu/archiving-preservation/digital-archives/certification-and-assessment-digital-repositories/hathi.

[10] We have several options for making this content locally discoverable. The simplest way is to direct patrons to the HT user interface, but this method of discovery is not integrated into our current environment in any way. We can also embed HT search widgets in YUL websites to execute HT catalog and full-text searches. A more seamless discovery approach for the public domain content in HT that is also held by YUL is to add an HT option to the newOrbis record similar to the Google Books option currently displayed in newOrbis. A more demanding approach to providing links from our MARC records to the HT content is inserting 856 tags in our records. We can add 856 tags for all content that is common to HT and newOrbis, or only for the content that we deposit in HT. For the extra-Orbis content, i.e., the content that is in HT and not in Orbis, we can load MARC records to newOrbis, or we can depend on the extensible searching capabilities of next-gen OPAC and web-scale discovery interfaces once such applications are fully adopted by YUL

- Additionally, HT content could be printed for patrons on-demand (with or without payment, depending on policy) using a local POD book machine.

Services for Persons with Print Disabilities[11]

The primary mission of Yale's Resource Office on Disabilities is to facilitate individual accommodations for all students with disabilities, and by so doing, work to remove physical and attitudinal barriers, which may prevent their full participation in the University community. Of the approximately 450 students registered last year with this office, 165 can be considered "print disabled" and includes students with visual impairments and dyslexia. The numbers of students continue to increase each year, due to the awareness of and prominence of diagnoses of dyslexia. This office offers to all students who are print disabled the option of support by way of digital texts required in their course work. There is an inherent time delay in producing a digital copy though our current resources. There are occasions when this limits the student's choice of reading. Access to the HT will ensure a prompt and improved service for these students and offer a more equal choice and timeline when compared to other students.

Opportunities Created Through Partnership

The opportunities for long term benefits of HT partnership are considerable: the scale of the project and depth of the collaboration involving major U.S academic libraries will shape future library standards, specifically in the areas of collection development and management, repository standards and best practices, methods for certifying the quality of deposited content, improved access and new public services. The following are some potential benefits resulting from opportunities created through the HT partnership that the working group identified.

Opportunities in the area of collaborative collection development

YUL's selectors' decisions related to collection development activities will be influenced by HT collection activities. The following benefits could influence several YUL's collecting policies such as purchase of retrospective materials and databases, withdrawals and replacements:

- An opportunity to build a platform and infrastructure that will allow development of a nearly *comprehensive* archive of published literature in the future. The HT content is constantly growing and on average, the number of unique titles in the database is increasing by about 6% each month. This growth represents an average increase of nearly 150,000 new titles each month. By 2013, the HT collection may be equal in size to Harvard University Libraries (cf. 16 million volumes). Within a decade, it could cross the threshold of 30 million volumes, making it larger than the U.S. Library of Congress is today.[12]

[11] This section was provided by Judy York, Director, Resource Office on Disabilities, Yale University.

[12] Constance Malpas, *Cloud-sourcing Research Collections: Managing Print in the Mass-digitized Library Environment*, http://www.oclc.org/research/publications/library/2011/2011-01.pdf

- The development of a *unique* repository of digital content that aggregates the content of large-scale digitization projects (Google, Internet Archive, Microsoft) with content drawn from the collections of HT partner libraries and external content from third party institutions.[13] In fact, *"attempting to attract and aggregate additional public domain content"* [14] is one of the future priorities of the Trust.

- *"Developing particular types of collections within the HT corpus, such as comprehensive or distinctive collections in particular areas that build on participant strengths."*[15] This is particularly important in the context of current budgetary constraints and pressure for libraries to reevaluate their efforts to create comprehensive collections when the idea of being a "library of record" is no longer sustainable.

- Capitalizing on the collective expertise of its partners in order to "*explore opportunities for digitization and collaboration with other initiatives."*[16]

- Developing specialized collaborative approaches for particular type of materials, for example: "*developing a shared approach to government documents that capitalizes on the work undertaken by CIC."*[17]

Opportunities related to collection management and access

- The convenience of the online environment that the HT offers can benefit users by broadening resource choice, providing ease of access/use, and saving them time.[18] Given the fact that academic and everyday-life information seeking behaviors of users change in different situations, the HT may or may not have an impact on the perceived value of print to our users. However, the convenience factor could potentially have a significant impact on print storage decisions in the future.

- Curating print and digital collections: The HT is a platform with a robust and *unique* infrastructure to efficiently store, manage, and preserve their collections of digital content, as well as to curate print more efficiently. In fact, the platform will allow leveraging *"the HathiTrust corpus to manage print collections both amongst and beyond the HT partner libraries, including extramural partnerships with third party organizations.*[19]

---

[13] Cf. Library vendors: for example, H.W. Wilson's *Essay & General Literature Retrospective*, *Short Story Index Retrospective*, and *Book Review Digest Retrospective* databases already feature links to the full text of public-domain books and other materials available through the HathiTrust digital library.
[14] The HathiTrust Collections Committee Charge: http://www.hathitrust.org/wg_collections_charge
[15] Ibid.
[16] Ibid.
[17] Ibid.
[18] Lynn Silipigni Connaway, *If It Is Too Inconvenient, I'm Not Going After It:" Convenience as a Critical Factor in Information-seeking Behaviors*, http://www.oclc.org/research/publications/library/2011/connaway-lisr.pdf
[19] The HathiTrust Collections Committee Charge: http://www.hathitrust.org/wg_collections_charge

- Changing policies and procedures: The partnership will bring increased operational efficiencies for institution members. For example:

    1. Library Shelving Facility (LSF) transfer: HT has the great potential to influence LSF transfer decisions, given the ability to link to digitized content in HT from newOrbis and the existing faceted-search discovery applications. The availability of publicly accessible content in HT could potentially change internal policies and workflows related to LSF transfer decisions.
    2. HT could change interlibrary loan workflows and enable better international sharing of resources. With a constantly growing collection, its impact on interlibrary loan can be significant. By investigating how preservation issues and lending policies are complicated by intellectual property rights at the international level, HT has great potential to advance digital resource collaboration worldwide and define what the next steps might be for the interlibrary loan community.

Opportunities related to new public services

HT provides an opportunity for breaking new ground in the areas of text mining and data-extracting research by allowing scholars to fully utilize constantly growing published works in the public domain (as well as limited access to works under copyright) that are stored within it.

The HathiTrust Research Center (HTRC), a new collaborative research center, was recently launched with the objective to help to develop cutting-edge software tools and cyber-infrastructure to enable advanced computational access to the growing digital record of human knowledge. This development presents a great opportunity for libraries to develop various public services such as:

- Training for students and scholars providing them with all necessary knowledge and techniques on how to conduct full text-mining research in HT.

- Consultations for researchers focused on identifying existing collections in HT for text-mining projects as well as on developing collections for text-mining projects that will include the selection of materials to be digitized that are not in HT and that correspond to researcher needs.

- Digitization and ingest of identified collections into HT.

- Solutions to problems with copyright issues.

YUL Involvement with HathiTrust

The working group identified three areas of potential collaboration with HathiTrust.

Improving Access

Our analysis identified the need for institutional or departmental access to HT in addition to the individual access that is already provided to Yale students, faculty, affiliated researchers and staff.

The Resource Office for Disabilities frequently acts as proxy to secure digital texts for the students. HT's model for providing access granted under §121 (Reproductions for blind or other people with disabilities) of the Copyright Act is based on individual authorization. Working with HT to develop proxy capabilities for partner institution departments, such as the Office of Disabilities, would benefit our own and the larger community.

Quality of digitization

The quality of digitized volumes varies tremendously. Quality of the digital copy is important if we are to consider a digital copy for a preservation replacement of Yale's copy. The University of Michigan with the support of the HathiTrust Advisory Board is in the midst of a three-year IMLS grant, *Validating Quality in Large-Scale Digitization: Metrics, Measurement, and Use Cases*. The grant will be looking at errors in the areas of data, including OCR; page images, what is obscured, blurred, etc; as well as the whole volume with regards to page order, missing pages, etc. At this time there is no systematic way to judge quality of volumes in HT. Currently, for volumes being considered for format conversion due to brittle paper, YUL Preservation Department is looking at page images and the completeness of volumes for volumes with digital copy held in HT. Based on information gathered over three months, 40% of volumes with digital copy in HT are considered good or usable copies; that is, they are complete with all pages, fold outs and/or plates and the content on all pages is legible. If invited, YUL should participate in phase II of the grant where the developed metrics will be applied to measure the extent of error in HathiTrust content.

API extensions

The available APIs work well and through using them for the content comparisons documented in this report, we see immediate opportunities for enhanced functionality. For example the Data API should be enhanced to enable limited access to partner libraries specifically to make possible the evaluation of the quality of restricted content as a preservation replacement for print. YUL staff should work with the HT partner community to identify enhancements to the current APIs.

Appendix A – Cost Examples[20]

Example 1 (2013):	10+ million volumes
$70,000	3 million public domain
	20% of Yale's collection represented in the in-copyright corpus
	Average Yale in-copyright book is owned by 5 other libraries
Example 2 (2015):	13 million volumes
$131,000	3.5 million public domain
	30% of Yale's collection represented in the in-copyright corpus
	Average Yale in-copyright books is owned by 4 other libraries
Example 3 (2017):	15.5 million volumes
$200,000	4.2 million public domain
	One third of Yale's collection represented in the in-copyright corpus
	Average Yale in-copyright books is owned by 3 other libraries
Example 4 (2017):	15.5 million volumes
$66,000	4.2 million public domain
	15% of Yale's collection represented in the in-copyright corpus
	Average Yale in-copyright books is owned by 5 other libraries
	Storage/management cost per book decreases 5% per year

ASSUMPTIONS AND CAVEATS - SEE NEXT SHEET TO USE DIFFERENT ASSUMPTIONS

These projections assume a 10% collection growth rate, balanced between PD and in-copyright portions. This is not likely to be sustainable, as the PD portion will eventually stagnate, resulting in Yale's costs increasing more quickly.

These projections (except 4) assume a 3% cost-per-book deflator based on declining storage costs. If storage costs do not decline at faster than 3%, increased energy and management costs may negate these savings. If the average file size of an item increases dramatically, this net cost could increase.

These projections assume a constant 1.5 cost multiplier. If increased development is needed to accommodate new formats, e.g., video, audio, this multiplier may increase. Conversely, if development slows, costs could decline.

[20] The spreadsheet used to generate these examples is available here: https://collaborate.library.yale.edu/hathi/Shared%20Documents/hathi-costs.xlsx

These projections assume fixed library participation. If libraries drop out of HT, Yale's cost will increase. If many smaller libraries join, Yale's cost will decrease modestly. If many large libraries join, Yale's cost will decrease more markedly. As the number of libraries duplicating Yale's in-copyright holdings increases, costs decrease dramatically.