A Preservation Infrastructure Built to Last

Preservation, Community, and HathiTrust

Jeremy York
Project Librarian, HathiTrust

Abstract
This paper describes the strategies HathiTrust is taking to build a collaborative infrastructure capable of ensuring long-term access to digital collections at scale. HathiTrust's approach recognizes the deep interplay of social and technical factors that support our collections, and will determine their persistence and availability over time.

Author
Jeremy York has been the project librarian for HathiTrust since July 2008. His primary duties include project coordination among the partnership, maintenance of HathiTrust's informational web site, and activities surrounding new partners and partnership contracts. Jeremy received a bachelor's degree in history from Emory University in 2001 and a Master of Information Science from the University of Michigan in 2008, with a specialization in archives and records management.

1. Introduction

HathiTrust is a partnership of academic and research institutions that are pooling resources to collaboratively preserve and provide access to the cultural record. The core of the preservation strategy centers on a digital repository that is owned and operated by the partners to ensure the long-term preservation of digital materials owned by their institutions, and facilitate access to the greatest degree allowed by law or third party agreements. The repository was launched in 2008 and currently contains more than 10 million volumes, making it one of the largest research library collections in the world. This paper offers insights into the guiding principles and ideas that underlie the repository, and specific strategies the partners are employing to preserve and provide access to digital collections at such a scale.

2. Setting

In the last 10 years, the time in which HathiTrust was conceived and initiated, there has been an explosion in the amount of materials digitized and produced digitally by libraries and other cultural heritage institutions. This has resulted in an increased focus in the cultural heritage sector on issues of digital preservation. Libraries and other institutions have grown significantly in their knowledge of the specific components involved in digital preservation, such as formats, media, and management of digital objects over time. They have grown also in their understanding of, and tools for evaluating, attributes and characteristics of "trustworthy" initiatives for long-term preservation. The challenges of preserving our digital present and past have been increasingly well defined. However, questions remain about the best ways to meet these challenges, from preservation models to employ (e.g., distributed versus centralized architecture), to formats and technologies to use, to specifications and best practices to follow.

In seeking to address these challenges, and in building a preservation infrastructure designed to operate at tremendous scale, HathiTrust has taken an approach that recognizes preservation first and

Proceedings of The Memory of the World in the Digital Age: Digitization and Preservation. An international conference on permanent access to digital documentary heritage, 26-28 September 2012, Vancouver, British Columbia, Canada, edited by Luciana Duranti and Elizabeth Shaffer (UNESCO 2013).

foremost as a social and collaborative activity. This approach has led to technological, architectural, and procedural decisions that, while important in their own right, are subordinate to, and guided by, an overall aim to meet the needs of a targeted community, even as the needs of that community change over time. This paper walks through the specific strategies that HathiTrust is taking to address common challenges in digital preservation, including issues of authenticity, reliability, scalability, sustainability, and discovery and access, in light of two guiding principles: that it is we, collectively, who are responsible for ensuring the persistence and availability of our cultural record; and that we can do more together than we can do separately.

3. Community

> Viewed developmentally, the problem of preserving digital information for the future is not only, or even primarily, a problem of fine tuning a narrow set of technical variables. It is not a clearly defined problem like preserving the embrittled books that are self-destructing from the acid in the paper on which they were printed. Rather, it is a grander problem of organizing ourselves over time and as a society to maneuver effectively in a digital landscape. It is a problem of building—almost from scratch—the various systematic supports, or deep infrastructure, that will enable us to tame anxieties and move our cultural records naturally and confidently into the future.
>
> (Task Force on Archiving of Digital Information 1996, 7)

The quote above is taken from the 1996 report of the Task Force on Archiving of Digital Information—a report foundational to the establishment of criteria for certifying trustworthy digital repositories, and significant in the development of the framework for Open Archival Information Systems.[1] One of the primary focuses of the report was on the need to advance the establishment of trusted systems for preserving digital information (1996, 9-10). As the quote above demonstrates, the report also recognized the social framework and interplay of social factors that both support and benefit from trustworthy digital preservation.

The importance of both of these aspects, the technical and the social, are carried forward in the OAIS model and in Trustworthy Repositories Audit and Certification: Criteria and Checklist (TRAC), the culmination of many years work to establish criteria for certifying trustworthy digital repositories. It is significant, however, that social components in both of these models are articulated primarily in a service-consumer relationship. The OAIS model defines a repository's Designated Community as "an identified group of potential Consumers" of information, where Consumers are the "persons, or client systems, who interact with OAIS services to find preserved information of interest and to access that information in detail" (Consultative Committee for Space Data Systems 2002, 1-8). The Designated Community is specified separately from a Management function, which typically acts to provide funding, conduct reviews of the OAIS, determine pricing policies, and "provide support for the OAIS by establishing procedures that assure OAIS utilization within its sphere of influence" (2002, 2-8).

In TRAC, one of the three overarching areas comprising the evaluation is organizational infrastructure, including issues of governance, staffing, finances and sustainability, contracts and liability,

[1] The report was the source of the recommendation to institute a dialogue on the "standards, criteria and mechanisms needed to certify repositories of digital information as archives" (1996, iv), and is taken as a point of reference in the Trustworthy Repositories Audit and Certification: Criteria and Checklist (CRL and OCLC 2007, 1). It was also the basis for the Preservation Description Information in the OAIS model (Consultative Committee for Space Data Systems 2002, B-1).

and succession (CRL and OCLC 2007, 3). TRAC explicitly recognizes the social elements that underlie a trustworthy repository. TRAC borrows the definition of a Designated Community from OAIS, however, reasserting from an earlier 2002 report (RLG and OCLC 2002) that "the definition of a trusted digital repository must start with "a mission to provide reliable, long-term access to managed digital resources to its designated community, now and into the future" (CRL and OCLC 2007, 3). Here too, the envisioned relationship is weighted toward one where an organization (stakeholders fulfilling management functions) provides services to a separate body of end users.[2]

The notion of Designated Community does not exclude the idea or possibility that designated communities and stakeholder communities could be the same, but it is worth affirming this possibility explicitly, as it is precisely the model under which HathiTrust was established and operates. The partnership is a community of academic and research institutions that are collaborating to provide a shared digital preservation infrastructure that will enable them to better achieve their goals in provisioning the cultural record for the advancement of scholarship at their institutions. By doing this, the partnership serves immediate access needs of end users who are part of this community (those engaged in scholarship and research at the partnering institutions, including, as they may be, stakeholders, staff, students, and faculty, etc.). HathiTrust's Designated Community, then, encompasses both those who steward and manage the digital archive, and the immediate users of information contained within the archive. As the partners are committed to using their collaborative services to produce public goods (making materials in the digital repository as open and available to anyone in the world as possible), they are able to reach beyond their designated community to serve a broader worldwide audience of libraries and library users.

There are clear strengths to this arrangement, where there is a tight coupling between those who support and manage the archive, and those who use and benefit from it. Perhaps the most important of these are first, that it provides the basis for a deep, collaborative social infrastructure where institutions are able to leverage common interest, distributed expertise, and diverse resources to achieve common and institution-specific goals more effectively and efficiently. Second, it creates strong forces that favor long-term sustainability. The sustainability of the archive depends on the ability of shared management and governance to ensure the archive continues to benefit the investing partners, and on the continued interest of the community in general in supporting scholarship and research. Both of these are fundamentally social factors.

A key element in ensuring the archive's ability to benefit those supporting it, however, is the technology used and the archive's technological approach in general. As the 1996 Task Force report noted, "We can afford to continue and increase economic and social investments in digital information objects and in the repositories for them on the information superhighway if, and only if, we also create the archival means for the knowledge the objects and repositories contain to endure and redound to the benefit of future generations" (1996, 9-10).

The remainder of this paper describes the approaches HathiTrust has taken to address challenges in preserving and providing access to digital information at scale in light of the social factors that ultimately underlie its success and sustainability. The partners have striven to develop robust technological infrastructure that is designed above all to be responsive to community needs for preservation and access, and that prioritizes meeting these needs over the long-term, even as technologies and implementations change over time.

[2] Examples given in OAIS of Consumer interactions include "questions to a help desk, requests for literature, catalog searches, orders and order status requests" (Consultative Committee for Space Data Systems, 2-9, 2-10).

4. Overarching Considerations: Scale, Preservation and Access, Openness

There are three broad considerations that have had a significant impact on the design and implementation of the HathiTrust repository. All result from underlying goals to meet the needs of the designated community. These considerations are the exceptional scale of the repository, a philosophical belief that the value of preservation is gained through access—that there is no value to a community of preservation without access, and a strong commitment to openness. Ultimately, HathiTrust's need-based approach to the development of services and strategic directions is one of its strongest attributes.

4.1 Scale

In his testimony as part of the Google Settlement fairness hearing in 2010, Paul Courant provided a summary of the purpose of the libraries that were engaging in large-scale digitization of their collections in partnership with Google. The excerpt of his testimony below characterizes well the primary needs of HathiTrust's designated community:

> Without reliable access to the scholarly record, we cannot know what has been known, what has proved fruitful and fruitless in the past. The broad social benefit that derives from the progress of science and the useful arts depends on the ability to find, use, and reuse the scholarly record. Provision of the scholarly record for current and future generations is the primary mission of these research libraries. (Courant 2010)

This statement highlights the needs of researchers and scholars to discover, access, and cite the scholarly record over time, and of libraries to preserve the scholarly record and enable these activities. Something that immediately stands out about these needs is the expansive scope. The needs are not to preserve works from a particular country or time, by a particular author or set of authors, or in a particular format or medium (for instance print or analogue versus digital, or even born digital). The needs relate to all materials that can be used to further scholarship. HathiTrust acknowledges the scope of these needs in its broad mission "To contribute to the common good by collecting, organizing, preserving, and communicating the record of human knowledge" (HathiTrust n.d.a.). It acknowledges these needs also, and some particular points of strategy in addressing them, in its initial goals. One of most important points of strategy is that the effort to address the needs of HathiTrust's designated community should and must be collective: "co-owned and managed" as the goals state, by the institutions ultimately responsible for the digital archive. The goals are as follows:

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

- To create and sustain this "public good" in a way that mitigates the problem of free-riders.
- To create a technical framework that is simultaneously responsive to members through the centralized creation of functionality and sufficiently open to the creation of tools and services not created by the central organization. (HathiTrust n.d.a)

The initial goals center broadly around collections and collaboration: assembling a digital collection of materials, as comprehensive as possible; providing access to the materials; preserving them; and then using the materials in broader strategies that benefit first the partners, and by extension a larger worldwide community. Libraries' goals to improve access to materials (including discovery and use) require their willingness and ability to address a number of specific micro-level challenges such as proper identification, description, and rights determination of materials (gaining a knowledge of what we have in our collections). Understanding what we have in our collections, the relationships between individual items, and the items' rights statuses are precursors to the development of individual and collective strategies for macro-level challenges such as managing print and digital collections, expanding of lawful uses of in-copyright materials, and in general, improving our collective preservation infrastructure. The common characteristic of all of these challenges and strategies is that they are *big*. They lend themselves to, and can be best responded to by, collective action, at scale. Partnering Institutions support HathiTrust specifically as a platform to address their needs in these areas and facilitate this kind of collective action.

4.2 Preservation and Access

HathiTrust is a "light" archive and as such, strives to provide as much access as legally possible to all materials in the repository. Works that are determined to be in the public domain, or that rights holders have opened access to, are available to be read online as well as downloaded, subject to third party agreements.[3] HathiTrust recognizes legal constraints and contractual obligations on materials, but does not preserve materials that depositors would wish to be stored without access, when access might otherwise be lawfully granted. For works that are in-copyright, HathiTrust includes full-text OCR in its repository-wide full-text search index, so that even though they are not available for reading or download, in-copyright works can be searched to retrieve word or query frequencies that may assist in determining the relevancy of a work or in locating specific information in a hard-copy of the work.

HathiTrust's "light" orientation benefits users, as it provides access to a tremendous body of materials. It also has benefits for preservation, as the processes of retrieving and displaying data provide an additional check on the integrity of objects, and access in general gives the digital objects the best chance to be used and valued in the community, and therefore preserved into the future. The goal of providing access to preserved works manifests itself in numerous ways throughout HathiTrust's technological infrastructure.

4.3 Openness

The last of HathiTrust's goals speaks to a technical framework that provides significant centralized functionality, but is also open to distributed development of tools and services. This orientation and

[3] Full download of materials is available where no restrictions exist. In most cases Google-digitized materials, which make up the largest group of materials where restrictions exist, are only fully downloadable by members of HathiTrust partner institutions.

general strategy towards openness extends to all aspects of the repository, from content formats to hardware and software, to organizational structure. The general strategy is that the long-term sustainability of the repository is served to the degree to which it is possible for member institutions to make use of the collective assets and services of the partnership, and to contribute to and manage them as well. The impact of HathiTrust's strong commitment to openness will be discussed further below.

5. Technical Infrastructure, Social System

HathiTrust's technical infrastructure was designed to meet the needs of its designated community, which can be categorized broadly on one hand as reliable long-term access to materials, and on the other as more efficient management of materials and resources to this end. This needs-based approach has resulted in a step-wise, modular trajectory to repository development, where discrete components that fulfill the needs for preservation and access interoperate as an integrated whole.[4] It has also resulted in very practical decisions about these components that are fully cognizant of the concerns (including economic, technological, and sociological) of the designated community. The ways that HathiTrust has addressed core challenges in digital preservation, as articulated by the Task Force on Archiving of Digital Information's report in 1996 (which, as it has been noted, was a significant force in the development of TRAC and was used in the development of the OAIS model) is given below. It will be helpful before entering into a discussion of these elements to give a general description of the architecture and design of the repository.

5.1 Repository

The HathiTrust repository was developed according to the framework for Open Archival Information Systems and the Trustworthy Repository Audit and Certification criteria. The overall considerations for operation at scale, preservation and access, and openness have resulted in a strong drive for consistency and standardization across the repository. Consistency and standardization facilitate the operation of generalized processes across the repository for purposes of ingest and preservation (e.g., content auditing and reporting, replication, backup), as well as access (e.g., full-text search indexing, access to users through a variety of interfaces, collection-building capabilities). The major components of the infrastructure, shown in Figure 1, include:

- Ingest: processes check the fixity of objects received for deposit, transform them to HathiTrust specifications if needed, perform rigorous validation, package objects for ingest, and finally bring them into the repository.
- Archival Storage: HathiTrust storage consists of two geographically separated instances of the repository on spinning disk, with tape backup stored in a third location.
- Data Management: HathiTrust manages bibliographic and rights information about objects, as well as information about the print holdings of partner institutions that correspond with HathiTrust's digital holdings. The significance of managing partner print holdings will be discussed further below.

[4] See York 2010 for a detailed discussion of repository architecture.

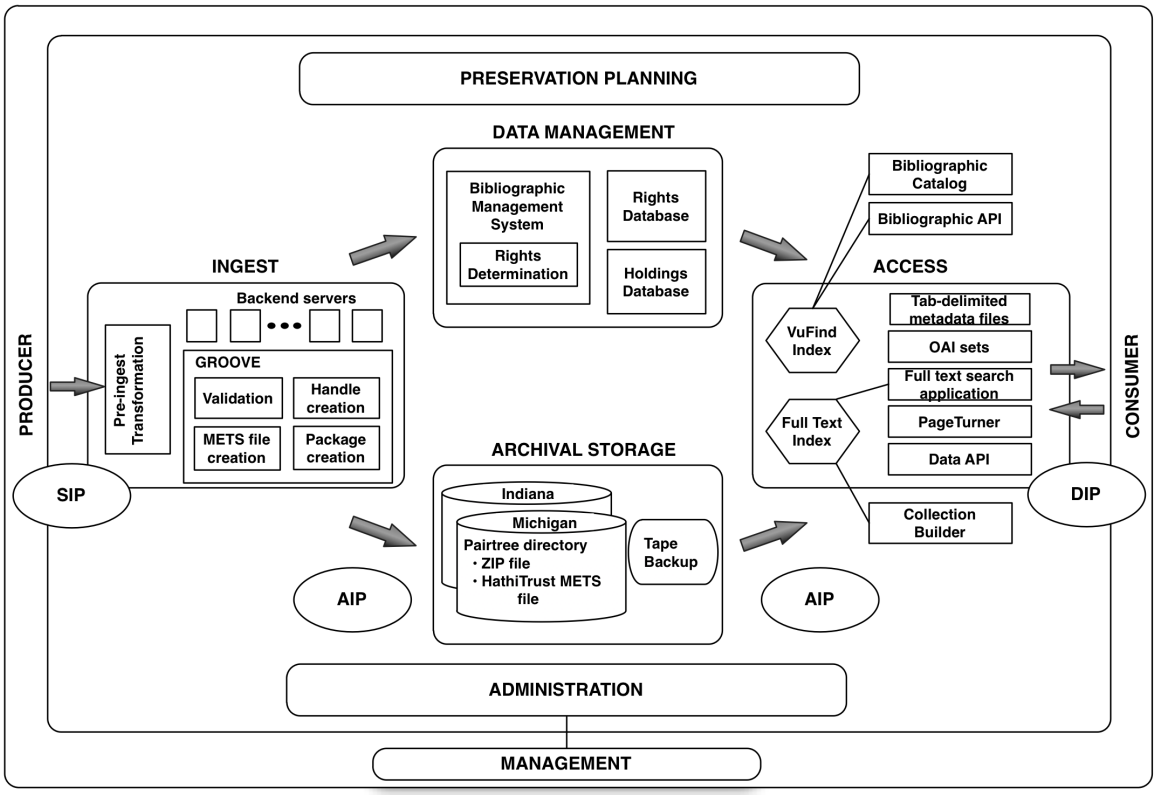

Figure 1. HathiTrust architecture according to the OAIS model (York 2010).

- Access: access services include
    - Bibliographic and full-text search of all materials
    - Reading and download capabilities for public domain and open access materials
    - The ability to assemble virtual collections of materials (i.e., "book bag" functionality)
    - A variety of APIs and data feeds for both bibliographic data and repository content (images and OCR text)

5.2 Preserving Digital Information

In describing the overall landscape of digital information and preservation, the Task Force on Archiving of Digital Information notes that:

> The process of preserving digital information will vary significantly with the different kinds of objects – textual, numeric, image, video, sound, multimedia, simulation, and so on – being preserved. Whatever preservation method is applied, however, the central goal must be to preserve information integrity; that is, to define and preserve those features of an information object that distinguish it as a whole and singular work. In the digital environment, the features that determine information integrity and deserve special attention for archival purposes include the following: content, fixity, reference, provenance, and context. (1996, 13)

Each of these elements, content, fixity, reference, provenance, and context, will be taken in turn.

5.2.1 Content

With regard to content, the Task Force states:

> The measure of integrity in the preservation process thus turns, at least in part, on informed and skillful judgments about the appropriate definition of the content of an digital information object—about the extent to which content depends on its configuration of bits, on the structure and format of its representation, and on the ideas it contains—and for what purposes. (1996, 13)

At the broadest level, considerations about content in HathiTrust begin with what is selected for digitization and preservation. The materials ingested by HathiTrust to-date have been those digitized and submitted individually by the partnering institutions (i.e., materials determined by those institutions to be of enduring value). HathiTrust formed a Collections Committee in 2010, however, whose charge includes making recommendations about content in HathiTrust (HathiTrust n.d.b), and partners recently approved a targeted initiative surrounding United States federal government documents (HathiTrust n.d.c). These initiatives underscore opportunities for collective decision-making about the addition of materials to the repository in the future.

Apart from selection of materials at an intellectual level, HathiTrust has defined parameters for the general types of materials as well as the content formats and specifications that are accepted. The general types of materials HathiTrust preserves at a production level currently are digitized books, journals, and book-like materials, such as codex manuscripts. Pilot projects involving image (e.g., maps and photographs), audio, and born-digital content are underway. For the books and book-like materials, there are only three formats in the repository that are primary targets of preservation: ITU G4 (bitonal) TIFF images, JP2 images, and Unicode text (HathiTrust volume packages include both plain text Optical Character Recognition (OCR) text and OCR with word coordinate location information). HathiTrust enforces adherence to these formats (including validity), minimum resolution standards, and internal image metadata specifications through rigorous validation processes on ingest. The specific types and numbers of formats in HathiTrust are not important in and of themselves (HathiTrust will undoubtedly support more formats and types of materials at a production scale over time), but are important the degree to which they satisfy a variety of community concerns. For example, ITU G4 TIFF, JP2 and Unicode are standard and open formats that meet community-accepted standards for digital preservation. They are also widely supported on a number of platforms and not dependent on particular hardware or software to render to users. These attributes of the formats inspire confidence in the community in their ability to be preserved and migrated forward to new preservation formats over time.

The formats HathiTrust accepts express its orientation toward openness, which in this case facilitates preservation of materials. The openness of formats facilitates access to materials as well, however, and access in particular at scale. The openness and flexibility of the formats allows them to be transformed on the fly to formats that can easily be downloaded or displayed to users on the Web. Management of files is thus simplified, as derivative images do not need to be stored in the repository, and repository systems do not need to be developed to maintain and disseminate them. The openness of the formats allows a uniformity of content across the repository that lowers the overhead of cross-repository management functions while offering a variety of access options to users.

It is relevant to note that HathiTrust has benefitted greatly from the uniformity of Google digitization with regard to content format and standards. The fact that millions of volumes have been digitized in the same way to the same specifications has greatly facilitated rapid growth of the repository (HathiTrust has

grown overall from 2.5 million volumes from its launch in 2008 to 10.5 million in 2012). HathiTrust has encountered challenges when seeking to accept content digitized from other sources, even when those sources use the same formats. This is due to the issues and work involved in transforming content to meet HathiTrust specifications, including assembling content and metadata into HathiTrust content packages. The collective work of the HathiTrust community has been key in addressing this challenge.

HathiTrust partners have worked closely together to define specifications and process for transforming content from large-scale digitization sources such as the Internet Archive, and develop tools, available from the HathiTrust website (HathiTrust n.d.d), that allow partners to transform, validate, and package content that is digitized on a smaller scale to HathiTrust specifications prior to submission. Working collaboratively, HathiTrust institutions have developed a framework that allows institutions to participate deeply in the preservation of their content, while lowering the overall costs to the repository of staging and transforming content, some of the highest costs associated with digital preservation repositories.

5.2.2 Fixity

Fixity in the Task Force report refers to the way content is "fixed as a discrete object," with the concern that objects might be changed or corrupted without notice (1996, 14). The concept of fixity relates closely to the concept of authenticity, as articulated by Luciana Duranti (Duranti 1995) and authenticity and integrity, as discussed by Clifford Lynch (2000). These relationships will be explored more closely below.

HathiTrust verifies the fixity of objects internally at several levels. The first is through verification, when possible,[5] of checksums for content as a part of the ingest process (calculating a message digest for content in the Submission Information Package and comparing it with the digest provided with the content). The second is by periodically re-calculating the checksums of objects in the repository and comparing them with checksums generated prior to ingest.[6] The third is through data integrity mechanisms internal to the storage itself, which use checksums to ensure that data transferred from one storage site to another are not corrupted, and to detect and automatically repair errors, including those caused by "bit rot" phenomena such as misdirected or torn writes.

HathiTrust communicates fixity, to a degree, to users as well, through the use of watermarks on images displayed in or downloaded from HathiTrust Web interfaces. The watermarks are not actually inscribed into the images themselves; they are overlaid on derivative images when the derivatives are created from the master files. It is thus possible to tell from the Web and printed copies that the images came from HathiTrust, although the watermarks do not have meaning for internal tracking.

These represent some of the mechanisms HathiTrust has in place. As Clifford Lynch has discussed, however, issues of authenticity and integrity at their base are largely functions of trust and context (2000). In discussing checks of internal consistency using checksums that are calculated for objects, he notes that when such a checksum or digest is used, "our confidence in the integrity of the object is only as good as our confidence in the authenticity and integrity of the digest." In such a situation, the link between the claim that a message digest is correct and the claim that an object maintains its integrity "is done by association and context—by keeping the claim bound with the object, perhaps within the scope of a

[5] It is possible that materials desired for ingest are not accompanied by valid checksum information.
[6] Checksums are recorded in metadata that is stored with objects in the repository.

trusted processing system such as an object repository." Put in other words, it is within a trusted environment that claims of authenticity and integrity have their meaning.

HathiTrust uses automated checks on integrity to detect random or accidental corruption of objects in the repository, but these mechanisms would likely not be sufficient to ensure integrity in the event of a successful intentional attempt to corrupt content. HathiTrust's multiple levels of redundancy (multiple storage locations and backup) could be used to restore any or all objects in the repository following such an act. A key point regarding fixity, however, is that it is broader mechanisms of system security (to prevent malicious forces from outside) and trust in staff (to ensure security from inside) that ensure the integrity of the overall environment, and give validity to further internal checks that are performed. As the fixity and integrity of objects depends to a significant degree on trust in the people and social system (the libraries) operating the repository, it is essential for the libraries participating in HathiTrust to maintain the trust of the community in this social system, as well as and including the technical systems it has in place.

5.2.3 Reference

With regard to reference, the Task Force states: "For an object to maintain its integrity, its wholeness and singularity, one must be able to locate it definitively and reliably over time among other objects" (1996, 15). HathiTrust addresses issues of reference in several ways. The first is the way items in the repository are identified. When an object enters the repository it is assigned an identifier that is composed of the identifier for the object prior to when it entered the repository, if available, and a namespace. HathiTrust prefers to use identifiers for objects that are in use by the depositor (in the case of digitized books this is often the barcode of the physical volume) if they have good identifier qualities, including guaranteed uniqueness (HathiTrust n.d.e). This is to avoid maintenance that would be involved in mapping and updating HathiTrust-generated identifiers, and to facilitate references by institutions to representations of their materials in HathiTrust. Namespaces are selected by the depositor and are used to identify the depositing source, as well as distinct identifier schemes of submitted objects. If items from a depositor have more than one identifier scheme, more than one namespace is used (HathiTrust n.d.e). As an example, the University of California uses the namespaces "uc1" and "uc2" to distinguish volumes digitized by Google and by the Internet Archive, each of which have distinct identifiers schemes. An example of an identifier in each group is given below:

uc1.b3543486 (Google-digitized)
uc2.ark:/13960/t26973133 (Internet Archive-digitized)[7]

HathiTrust thus takes great care to ensure the unique identification of items in the repository and enable references to original items where possible.

HathiTrust further enables reference through the structure of the repository. The objects in HathiTrust are stored in directories in one large file system. The repository uses a Pairtree structure, which maps identifier strings to directory paths for digital objects pair-wise, with the name of the final directory being the object identifier (Kunze et al. 2008). For example, the path on the repository file system to the directory of the item "uc1.b3543486" is ../uc1/pairtree_root/b3/54/34/86. The files of the object itself are located in this directory and named b3454386.zip and b3454386.mets.xml. The zip file contains the content files of an object—the images and OCR for digitized books—as well as additional

[7] These are the same examples used in York 2010.

content metadata. The XML file contains a variety of technical, administrative, and structural metadata encoded in the Metadata Transmission and Encoding Standard (METS) (Library of Congress n.d.a), that serve purposes both of preservation and access.[8] There are several benefits to using the Pairtree structure, including that it a) ensures that objects are uniformly accessible to repository systems and access services; b) makes it easy for content to be imported, understood, and used in new storage system without the system knowing anything about the nature or contents of the stored objects; and c) allows object operations such as backup and restore, to be performed using native operating system tools, facilitating disaster recovery (Kunze et al. 2008). These benefits have clear advantages for reference, as objects can be located and operated on through automatic processes. They also inherently facilitate operation at scale, preservation and access, and openness—openness to the degree that HathiTrust objects are not tied to the specific infrastructure they are stored on and could either be operated on by tools independent of the software used in current storage, or moved to totally different storage and operated on immediately.

There are three further ways that HathiTrust facilitates reference. The first is by embedding the identifier of objects in the metadata of images that make up digital volumes themselves.[9] The second is through the creation of unique and permanent identifiers for objects using the Handle System (Corporation for National Research Initiatives, n.d.). Permanent identifiers comprise a Handle namespace and the HathiTrust identifier, which are combined together to form a permanent URL where the object can be located on the Web. HathiTrust also provides the date of the most recent version of the volume in HathiTrust, facilitating citation (versioning is discussed in the section on Provenance below).

As in other areas, the mechanisms that HathiTrust uses to facilitate reference depend on broader social factors—for instance, the selection and use of identifiers by depositing institutions; the reliability of the Handle service. By taking a stance that is sensitive to these factors (e.g., favoring the use of existing identifiers, using a uniform scheme of structure and reference in the repository), HathiTrust positions itself to be responsive to them and as needs for and applications of reference capabilities change over time.

5.2.4 Provenance

The Task Force report highlights two ways that establishing the provenance of objects serves to preserve their integrity:

> First, a tracing of chain of custody from the point of creation helps to create the presumption that an object is authentic, that it is what it purports to be and that its content, however defined, has not been manipulated, altered or falsified (Duranti 1995: 7-8). The second effect of establishing provenance through a chain of custody is to document, at least in part, the particular uses of the object by the custodians. (1996, 17)

HathiTrust traces the provenance of digital objects by recording the original source of the material represented in HathiTrust, the agent of digitization, and a variety of administrative (including provenance and preservation) metadata about objects, where this metadata is available.[10] HathiTrust uses the

[8] Details about HathiTrust content packages, and the metadata contained in the XML file are available at (HathiTrust n.d.f). See also York 2010.
[9] In the DocumentName element of TIFF files and the dc:source element of JP2 files.
[10] At a minimum, HathiTrust requires information about the digital capture of items. An explanation is needed if this information is not available. HathiTrust accepts other information relevant to provenance and preservation of materials prior to their entry into HathiTrust, but does not require it.

Preservation Metadata: Implementation Strategies (PREMIS) standard (Library of Congress, n.d.b) to record preservation metadata, including the time and date of digital capture, transformations that may have been performed, fixity checks, validation, quality review, and other events that may occur either prior to HathiTrust taking responsibility of a digital object or subsequently.

HathiTrust does not keep multiple versions of objects in the repository.[11] If a new version of an object is available for ingest, as is often the case in particular with Google-digitized volumes,[12] the new version overwrites the existing version. A new series of preservation events (e.g., fixity check, digest calculation, validation, and ingest) is written into the PREMIS metadata for the re-ingested object, and all previous events are retained, providing a means to determine whether and how many times an object has been ingested. HathiTrust uses this practice primarily because of the immense scale of the repository and the frequency with which volumes from Google are reprocessed and made newly available. The cost of preserving multiple versions of Google-digitized volumes would drastically increase the cost of preservation to partners, and the value of retaining these versions is not clear. For instance, it is important for users to be able to reliably cite the version of an item that they used (a version date is provided for this purpose as mentioned above). It is also important for HathiTrust to record changes that are made to volumes where possible.[13] Whether or not it is important to record Google's or another entity's attempts to create a reliable representation of a known object is a separate question. The Task Force report acknowledges the relationship between documenting provenance and the concepts of fixity and authenticity, and factors affecting and complicating notions of authenticity have been discussed in the section on Fixity above. It is worth exploring these factors a little more deeply, however, particularly in relation to the concept of reliability, as it relates to issues of versioning and is raised in the article by Duranti that the Task Force cites.

A primary point that Duranti makes in the article is that the "concepts of authenticity and reliability must be kept intellectually separate"[14] A danger in conflating the two, or of focusing primarily on the integrity and authenticity of records, is that records may be completely authentic, but this does not mean that they are reliable (1995, 7). Duranti states that, "A record is considered reliable when it can be treated as a fact in and of itself, that is, as the entity of which it is evidence" (1995, 6). According to Duranti, the elements that provide a record with reliability are its form and its and procedure of creation ("the body of rules according to which acts or portions of them are recorded")" (1995, 6). Duranti notes, "A record is regarded as reliable when its form is complete, that is, when it possesses all the elements that are required by the socio-juridical system in which the record is created for it to be able to generate consequences recognized by the system itself." Some elements that commonly contribute to form are signature and date

[11] Versions here refer to multiple different copies of a distinctly identified object (for instance, one version of an object that is missing a page and a corrected version that is not). HathiTrust does preserve multiple editions of the same work, as well as multiple copies of the same work that are in the repository under different identifiers (i.e., duplicate volumes).

[12] Google is constantly improving the algorithms it uses to process the raw images it captures of library volumes. When new versions of volumes from any institution are available, if they pass a certain quality threshold, they are re-ingested into HathiTrust.

[13] The partners have observed a trend in higher quality scans returned from Google over time, but an automated mechanism to determine whether or how much the quality of a given volume has changed following reprocessing by Google (and correspondingly, what specifically has changed) does not exist. This can only be determined through manual inspection, though many minute changes would likely not be detectable.

[14] (Duranti 1995, 8). Duranti's concern is primarily with electronic records, but many of the same concerns apply to digitized volumes.

of creation. Procedures of creation might include appropriate responsibility for signing, recording of facts by multiple persons, or distribution to multiple addresses (1995, 6). Duranti states that the same elements, completeness of form and procedure of creation, determine the reliability of copies that are made of originals, and acknowledges that there are different degrees of reliability, ranging from a simple copy made "without the dating and attestation of the copying person," to copies that might be more reliable than the originals themselves (1995, 7).

The digitized volumes found in HathiTrust are intended to be copies of the original physical volumes.[15] However, while HathiTrust records the date images of original volumes were created, in most cases (except when files are received directly from the publisher) there is no official entity that verifies the reliability of the copies with respect to the way they represent the originals. In the socio-juridical context of library digitization, the reliability of digital copies, their ability to stand for the items they represent and to generate consequences (i.e., be cited and used as surrogates for the physical volumes), has generally been established by libraries through quality assurance. Whether libraries conduct the digitization of items themselves or contract with a vendor, there is an underlying assumption, because of the trust society places in libraries and libraries' living up to this trust over time, that appropriate steps have been taken to ensure that digitized copies accurately represent the original items, or that steps can be taken to address problems that are encountered.

In this context, where libraries have primary responsibility to their communities for ensuring the reliability of digitized items they make available, the question of retaining versions of items, and broad questions of reliability in general, become questions that depend on the needs and resources of the social system. The question in this case becomes not whether to retain multiple versions of items, but how to best meet community needs for preservation, including reliability and authenticity, in ways that are sustainable and responsibly manage community resources. Recording the provenance of digitized items is crucially important, but does not in and of itself speak to issues of reliability, and can be completely separate if a digitized item has quality problems and is not able to stand as a faithful copy of the original it is intended to represent. Issues of reliability, particularly in relation to information quality and in light of community needs, are a current area of study for the partnership.[16]

5.2.5 Context

Content in the Task Force report refers to "the ways in which [digital information objects] interact with elements in the wider digital environment" (1996, 18). The report points to three dimensions of context that have to do with technical aspects (hardware and software dependencies), linkages among digital objects (to the degree to which the integrity of an object lies in the network of linkages), and communication medium (the extent to which the way materials are distributed—for instance, bandwidth or security constraints or attributes—account for characteristics of the digital objects) (1996, 18-19).[17] The

[15] HathiTrust's Digital Preservation Policy notes that "HathiTrust is committed to preserving the intellectual content and in many cases the exact appearance of materials that have been digitized for deposit." This includes "Digital representations (images) of content as the content appeared in its original form, with the same layout and colour (e.g., for illustrations and artwork), and in the same order." (HathiTrust n.d.g).

[16] See, for example, the work of Paul Conway (Conway 2011). HathiTrust's policy on quality is available at http://www.hathitrust.org/quality.

[17] With regard to communication medium, the report gives the example of increasing bandwidth resulting in the production and distribution of high-bandwidth products such as "full-motion video."

report points also to a broader social environment and the contextual role that policies and implementation details regarding bandwidth, security, and other network qualities can have on information integrity.

HathiTrust's use of open formats and practice of transforming master images for access in different contexts address many of these contextual concerns about information integrity (the integrity of objects in HathiTrust does not depend on hardware or software, and due to the flexibility of formats, access considerations are separate to some degree from preservation concerns). HathiTrust objects exist entirely within the repository, so the issue of linkages as it is relayed does not apply. There are some significant elements of context in HathiTrust, however, that are relevant beyond the explicitly digital environment in which they exist. The first of these is the relation of objects in HathiTrust to their print counterparts, and the second has to do with discovery and use.

Relation to Print. HathiTrust's fourth stated goal is "To stimulate redoubled efforts to coordinate shared storage strategies among libraries, thus reducing long-term capital and operating costs of libraries associated with the storage and care of print collections." Understanding the relation of the digital objects in HathiTrust to the print items owned by libraries (whether the item in HathiTrust was digitized by their library or not) has had profound implications for the development of HathiTrust. HathiTrust began with a pricing model based on a per-gigabyte fee, covering the infrastructure costs of the content institutions deposited. As HathiTrust has grown (from 2.5 million volumes when it was launched to nearly 10.5 million today), the overlap with North American academic and research libraries has become so significant—likely more than 50% on average with Association of Research Library libraries[18]—that it has shifted from being a strategy for institutions to preserve their digital volumes, to being a strategy to preserve their print volumes as well (with digital "backups").

In recognition of this fact, HathiTrust developed a pricing model, which will be in effect in 2013, that is based on the overlap of partnering institutions' print holdings with the digital holdings in HathiTrust. The pricing model is supported by a holdings database (represented in the Data Management component in Figure 1) that maps institutional print holdings to holdings in HathiTrust. This holdings database represents a new contextualization of the digital objects, and the physical objects as well, in a broader information environment. In addition to helping libraries to understand the relationships between their collections of print and digital objects, the holdings database will support the expansion of lawful uses of in-copyright materials in HathiTrust that are owned in print by the partnering libraries.[19] Contextualizing the holdings of HathiTrust revolutionizes the way libraries conceive of their collections and provides a basis for a "deep infrastructure" on which libraries can collaboratively move their content and services into the future.

Discovery and Use. HathiTrust offers centralized access services, including bibliographic and full-text search, as well as reading, downloading, and collection-building capabilities. In addition to these services, and in support of the last of its stated goals related to a centralized yet open technical framework, HathiTrust offers several APIs and data feeds that allow partner and non-partner institutions to contextualize their own collections in relation to HathiTrust. For instance, partners can use information from HathiTrust APIs to add links or entire records for HathiTrust items to local discovery mechanisms.[20]

[18] This figure is extrapolated based on analysis of trends observed in (Malpas 2011) and HathiTrust repository growth since that time.
[19] A description of the specific uses and circumstances of access to in-copyright works that HathiTrust has targeted is given at http://www.hathitrust.org/authors_guild_lawsuit_information#Details.
[20] Information on how this can be done is available at (HathiTrust, n.d.h).

The collection-building capability allows users to create and reference canonical bibliographies of materials such as the English Short Title Catalog, and other sets of materials.[21]

In each of these ways, by contextualizing HathiTrust materials in their broader environment, and allowing others to contextualize local collections, HathiTrust creates pathways for meaningful connections between collections and items to be made. This simultaneously improves discovery and use of materials, and uncovers new opportunities for libraries and cultural heritage institutions to engage in collective action to address shared challenges.

6. Conclusion

This paper has highlighted the ways that HathiTrust's understanding of preservation as a social and collaborative activity has influenced specific approaches it has taken to preserving digital information—both in its technical infrastructure, and in relation to issues of content formats, fixity, reference, provenance, and context. By focusing on community needs and social factors in concert with technical considerations, HathiTrust has been able to gain a broad base of support for its activities, and take decisions that strengthen its ability to meet community needs over the long-term. In the end, it is we, collaboratively, who have responsibility for moving our collections into the future, and strategies that bring our efforts closer in concert with one another are the most likely to succeed.

Acknowledgements

This paper has only been possible through deep consultation and collaboration over a period of years with staff at the University of Michigan Library that designed and built HathiTrust's repository infrastructure, and staff at numerous institutions as the partnership has grown. I would like to extend special thanks to Cory Snavely, Aaron Elkiss and Sebastien Korner for their input and clarification of repository processes, and John Wilkin for reviewing and commenting on the paper.

References

Consultative Committee for Space Data Systems. Reference Model for an Open Archival Information System (OAIS). Washington, D.C: CCSDS Secretariat, 2002; Program Integration Division (Code M-3), National Aeronautics and Space Administration.

Conway, Paul. "Archival Quality and Long-term Preservation: a Research Framework for Validating the Usefulness of Digital Surrogates." *Archival Science* 11, no. 3 (2011): 293–309. doi:10.1007/s10502-011-9155-0. http://www.springerlink.com/content/9322j77m6327h123/abstract/.

Corporation for National Research Initiatives. "Handle System." http://handle.net/.

Courant, Paul. "Testimony of Dean Paul Courant." Fairness Hearing on Proposed Settlement, February 18, 2010. http://www.lib.umich.edu/michigan-digitization-project/fairness-hearing-testimony-of-dean-paul-courant.

[21] The collection of English Short Title Catalog materials is available at http://babel.hathitrust.org/cgi/mb?a=listis;c=247770968. See also HathiTrust's featured collections and the full list of user-created public collections: http://babel.hathitrust.org/cgi/mb?a=listcs#all.

CRL, and OCLC. Trustworthy Repositories Audit and Certification: Criteria and Checklist. 2007.http://www.crl.edu/sites/default/files/attachments/pages/trac_0.pdf.

Duranti, Luciana. “Reliability and Authenticity: The Concepts and Their Implications.” *Archivaria* 1, no. 39 (1995): 5-10. http://journals.sfu.ca/archivar/index.php/archivaria/article/view/12063/13035.

HathiTrust. n.d.a. “Mission and Goals | HathiTrust Digital Library.” http://www.hathitrust.org/mission_goals.

———. n.d.b. “Collections Committee Charge | HathiTrust Digital Library.” http://www.hathitrust.org/wg_collections_charge.

———. n.d.c “Constitutional Convention Ballot Proposals | HathiTrust Digital Library.” http://www.hathitrust.org/constitutional_convention2011_ballot_proposals#proposal4.

———. n.d.d. “Ingest Tools | HathiTrust Digital Library.” http://www.hathitrust.org/ingest_tools.

———. n.d.e. “Guidelines for Digital Object Deposit | HathiTrust Digital Library.” http://www.hathitrust.org/deposit_guidelines#pdi.

———. n.d.f. “Digital Object Specifications (METS and PREMIS) | HathiTrust Digital Library.” http://www.hathitrust.org/digital_object_specifications.

———. n.d.g. “Digital Preservation Policy | HathiTrust Digital Library.” http://www.hathitrust.org/preservation.

———. n.d.h. “Data Availability and APIs | HathiTrust Digital Library.” http://www.hathitrust.org/data.

Kunze, John, Martin Haye, Erik Hetzner, Mark Reyes, and Cory Snavely. “Pairtrees for Object Storage.” 2008. https://confluence.ucop.edu/download/attachments/14254128/PairtreeSpec.pdf?version=1.

Library of Congress. n.d.a. “Metadata Encoding and Transmission Standard.” http://www.loc.gov/standards/mets/.

———. n.d.b. “PREMIS: Preservation Metadata Maintenance Activity.” http://www.loc.gov/standards/premis/.

Lynch, Clifford A. 2000. “Authenticity and Integrity in the Digital Environment: An Exploratory Analysis of the Central Role of Trust.” In *Authenticity in a Digital Environment*. CLIR Papers. Washington, D.C: Council on Library and Information Resources. http://www.clir.org/pubs/reports/pub92/lynch.html.

Malpas, Constance. “Cloud-sourcing Research Collections: Managing Print in the Mass-digitized Library Environment.” OCLC, 2011. http://www.oclc.org/resources/research/publications/library/2011/2011-01.pdf.

Research Libraries Group, and OCLC. *Trusted Digital Repositories: Attributes and Responsibilities*. Mountain View, CA: Research Libraries Group, 2002. http://www.oclc.org/content/dam/research/activities/trustedrep/repositories.pdf.

Task Force on Archiving of Digital Information. “Preserving Digital Information : Report of the Task Force on Archiving of Digital Information.” Commission on Preservation and Access and Research Libraries Group, 1996. http://library.oclc.org/cdm/singleitem/collection/p267701coll33/id/272/rec/11.

York, Jeremy. “Building a Future by Preserving Our Past: The Preservation Infrastructure of HathiTrust Digital Library.” In session 157 - ICADS with Information Technology. Gothenburg, Sweden, 2010. http://www.hathitrust.org/documents/hathitrust-ifla-201008.pdf.