Ethan Gates
CINE-GT 1807
9/23/14

The HathiTrust Digital Library (HDL) represents the efforts of a large coalition of American-based research universities to create and share a massive, online digital archive of their print resources. Founded in 2008 as a collaboration between the University of California and the Committee on Institutional Cooperation (itself a consortium between the Big Ten universities and the University of Chicago), and now expanded to over 90 partner institutions, the HathiTrust effort has digitized over 4 billion pages of print content, including over 6 million books and 300,000 serial titles, composing an estimated total of 523 terabytes of data. Similar to the Google Books project (with whom HathiTrust in fact also collaborates), HathiTrust aims to build a full-text searchable database out of the print collections of its partner institutions, making complete scans of pages available online whenever legally possible – thus any member of the public can see all of HathiTrust's content in the public domain; whereas with material under copyright users may be limited to only accessing catalog metadata (unlike Google Books, which provides actual samples or snippets of text even from copyrighted resources, the HDL only displays the frequency and page number of search terms to help users decide if they should pursue a physical copy of the book/journal). HathiTrust seems a more targeted version of Google Books, aimed primarily to serve the academic community: both allowing authenticated users (students and scholars of HathiTrust's partnered institutions) greater access to a large amount of historical print material, and reducing capital and operating costs for academic libraries by creating a shared storage space.

HathiTrust has committed to conforming to what standards exist for digital archives; they use the OAIS model for describing their archival operations and the Trustworthy Repositories Audit and Certification (TRAC) criteria for the specific requirements of preservation of digital files. Since the number of institutions submitting material to the HDL is so varied, stringent SIP requirements have been made to ensure that all ingested material contains a certain amount of metadata, and can then be easily accessed within HDL's catalog with a minimal amount of processing. There are two acceptable file formats for the SIP, depending on the nature of the content for each scanned page: all-text pages must be submitted as a bitonal TIFF file, while JPEG2000 is used for images. There is also a strict naming convention, through which individual books/journal volumes are assigned a unique identifier (HathiTrust members can access a database used to find, identify and assign unique identifiers, to reduce duplication within the archive), and then each individual scanned page within those books are assigned a further 8-digit identifying number. Bibliographic metadata for each volume scanned is also required as part of the SIP, in the form of a MARC21-conforming MARCXML file.

The actual digital objects (scanned pages) are ingested at the University of Michigan, where they are replicated to an active mirror site in Indiana, as well as backed up to LTO tape (and stored in a third facility some distance from Ann Arbor). The bibliographic material for each volume, meanwhile, is submitted separately from the digital object and managed by the University of California. Administrative metadata for

each SIP is also submitted to the HDL via a Google Doc (again, managed by University of California). HDL also maintains a third database regarding rights information for its materials, hosted by the University of Michigan's Mirlyn catalog; when materials are ingested, the rights database uses the volume's unique identifier to check the bibliographic database for certain metadata (publishing date, publication location, etc.) and assigns a copyright attribution to the volume – which then in turn is used by the PageTurner access system to determine whether a volume can be fully displayed to users (further depending on their authentication status; see below).

For every AIP deposited in the HDL, a "HathiTrust" METS file is created for the digital object, containing provenance, reference and fixity information. This preservation metadata is conformed to the PREMIS data dictionary, and is intended as a record of the object from the time it enters the repository onward; in some cases a "Source" METS file may also be provided in the SIP, which gives preservation metadata relating to the object before it was ingested into the HDL (often in cases where a non-partner organization, such as Google Books, digitized the original volume).

Curation for what volumes are included in the HDL is generally left to the discretion of HathiTrust's member institutions; at the moment, the HDL supports ingest of any digital book or journal content, as well as digitized manuscripts, as long as the digitized files conform to the University of Michigan's digitization specifications (described in general terms above, provided to the public in detail on the HathiTrust site in a 50-page PDF). HathiTrust also has agreements with Google Books and the Internet Archive, so that the HDL can automatically ingest content digitized and uploaded by those projects. Pilot projects are also in the works to support the ingest of digital audio and image content as well as born-digital publications.

Access to HathiTrust is controlled by the PageTurner system. Any user, whether they are an authenticated member of an HDL partner institution or not, can access PageTurner and complete full-text searches of the HDL database; however, once a user is authenticated (via Shibboleth), they are given the option to download full PDFs of public domain documents from the HDL. For in-copyright or out-of-print volumes, the HDL has been integrated into the OCLC's WorldCat catalog, so users can jump directly from an HDL catalog entry to its WorldCat equivalent, through which they can locate and/or request a print copy of the original volume. Special allowances are also made if the user is authenticated as print-disabled (e.g. allowing access to the full scans of in-copyright material), although some of these allowances are currently being legally contested. The PageTurner system also allows authenticated users to create their own "collections," wherein individual volumes or pages can be bookmarked and saved to a curated page (which the user can also make public if desired).

In terms of sustainability, HathiTrust appears to be preparing for both physical dangers related to storage, and maintaining the HDL system in future digital environments. Replacement of storage equipment at the data centers that hold the HDL's content, bibliographic and rights databases is conducted annually and assumes that all equipment has a useful life of 3-4 years; the storage system's N+3 Reed-Solomon parity redundancy ensures that data can be automatically redistributed and storage nodes replaced without manual movement of data. This storage system also performs periodic

data integrity checks for both “at-rest” and “in-flight” files (data lying dormant in storage or being actively transferred/accessed), using the parity redundancy to repair errors. HathiTrust also uses checksums to make periodic data validations outside of the storage system's built-in security. The HDL itself, meanwhile, was built using open-source technologies (including PERL for its primary programming language, Linux for its operating system, Apache for its web requirements, MySQL for its database server). From a financial standpoint, HathiTrust's infrastructure is paid for by its partners, with a model that distributes fees among the various institutions depending on the benefits each partner derives from the total collection (e.g. how many volumes in the HDL overlap with a library's physical holdings). The HathiTrust board considers this a flexible model for operations that will ensure the HDL can adjust to both evolving digital technologies and shifts in operating cost with minimal or no interruption to the service itself.

Whatever the unique specifications of HathiTrust's operation may be, the technical details of the project have largely been overshadowed in the public eye, as the HDL has turned into something of a testing ground for the legal boundaries of digitizing copyrighted material and fair use. Following their ongoing lawsuit against Google Books, the Authors Guild sued HathiTrust in September of 2011, claiming that the libraries involved in HathiTrust had violated copyright by creating full-scan digital copies of print materials, even if the HDL's copyrighted material was not made accessible in its entirety to the general public. A district court ruling in October of 2012 determined that building a full-text searchable database was a transformative use of copyright material and that therefore the HDL was acting within fair use. The Second Circuit Court of Appeals upheld that decision in June of 2014, although HathiTrust's attempts to use the database to replace books in its holdings was not ruled on, and will likely become the subject of a separate suit. In any case, HathiTrust has become a major example of the possibilities for fair use in digital preservation.

Six years into the project, HathiTrust appears likely to only grow further; it combines the mass digitization scale of Google Books with the combined resources and organization of dozens of world-class research universities. Now that it appears to have weathered the legal challenge from the Authors Guild, HathiTrust is free to continue growing its print collection; the next obstacle will likely be whether or not the HDL can be expanded to accommodate other forms of media, including audio or moving images. Also recently, Indiana University and the University of Illinois jointly launched the HathiTrust Research Center (HTRC), an online portal for teachers and nonprofit users that provides computational and data mining software to perform research in the public domain section of the HDL. Expanding and improving the tools available in the HTRC seems another top priority for HathiTrust in the future.

Webography

“About.” HathiTrust. Accessed Sep. 20, 2014. <http://www.hathitrust.org/about>.

"Authors Guild v. HathiTrust" Electronic Frontier Foundation. Accessed Sep. 20, 2014. <https://www.eff.org/cases/authors-guild-v-hathitrust>.

Furlough, Mike. "Sharing Collections through Shared Stewardsip: A HathiTrust Progress Report." July 23, 2014. Accessed Sep. 21, 2014. <http://www.hathitrust.org/documents/HathiTrust-TRLN-20140723.pdf>.

"Large-Scale Text Analysis through the HathiTrust Research Center." Poster, presented July 10, 2014. Accessed Sep. 21, 2014. <http://dharchive.org/paper/DH2014/Poster-356.xml>.

"Narrow Fair Use Ruling Permits Limited Library Uses, Shoots Down Replacement Copying." The Authors Guild. Accessed Sep. 20, 2014. <http://www.authorsguild.org/advocacy/narrow-fair-use-ruling-permits-limited-library-uses-shoots-down-replacement/>.