ADDING NEW CONTENT TYPES TO A LARGE-SCALE SHARED DIGITAL REPOSITORY

Shane Beers
University of Michigan
Preservation and Conservation
3215A Buhr Building
837 Greene St
Ann Arbor, MI 48109-3209

Jeremy York
University of Michigan
300 Hatcher Graduate Library North
920 N. University Avenue
Ann Arbor, MI 48109-1190

Andrew Mardesich
California Digital Library
University of California,
Office of the President
Oakland, CA 94612

ABSTRACT

As a digital repository for the nation's great research libraries, HathiTrust brings together the immense collections of partner institutions. Initially, the Submission Information Packages (SIPs) deposited into HathiTrust were extremely uniform, being constituted primarily of books digitized by Google. HathiTrust's ingest validation processes were correspondingly highly regular, designed to ensure that these SIPs met agreed-upon qualities and specifications. As HathiTrust has expanded to include materials digitized from other sources, SIPs have become more varied in their content and specifications, introducing the need to make adjustments to ingest and validation routines. One of the primary sources of new SIPs is the Internet Archive, which has digitized a large number of public domain materials owned by HathiTrust partners.

Many of the technical, structural, and descriptive characteristics of materials digitized by the Internet Archive did not match previously developed standards for materials in HathiTrust. A variety of solutions were developed to transform these materials into HathiTrust-compatible AIPs and ingest them into the repository. The process of developing these solutions provides an example to other organizations that would like to add new types of materials to their repository, but are uncertain of the issues that may arise, or how these issues can be addressed.

1. INTRODUCTION

As a digital repository for the nation's great research libraries, HathiTrust brings together the immense collections of partner institutions. Partnership is open to institutions worldwide who share this vision.

HathiTrust strives to conform to the characteristics of a Trustworthy Digital Repository [1], and a significant amount of work has gone into developing ingest functionalities that analyze SIPs to determine whether they meet a number of standards. The standards include the technical aspects of the digital image files in a SIP (such as resolution, well-formedness, compression type, color and bit depth), descriptive elements of the SIP (including PREMIS preservation metadata and image header metadata), and structural metadata that explain what the digital image files represent and allow software tools to display the images correctly.

The majority of SIPs being deposited into HathiTrust initially were books that had been digitized by Google, Inc. The specifications Google uses in its digitization package were worked out collaboratively with Google library partners, resulting in a tightly controlled technical and descriptive SIP. The validation environment employed in HathiTrust was developed around the ingest of these materials. For some time, this ingest process has worked well in verifying SIPs against set standards, allowing content into the repository when compliant, and reporting when something failed.

However, the scope of digitization at HathiTrust institutions is much broader than Google digitization alone, and one of the partners' initial goals was to accommodate the outputs of the variety of digitization initiatives they had undertaken in a single repository. Because a number of partner institutions have had materials digitized by the Internet Archive (IA), expanding the capabilities of HathiTrust to preserve and provide access to these materials was a logical and highly desirable direction to pursue.

In the summer of 2009, the University of California (UC) was poised to deposit an initial set of nearly 100,000 IA-digitized volumes into HathiTrust. Talks were initiated between staff members at California Digital Library (CDL) and the University of Michigan on how to accommodate ingest of this content, and in the fall of 2009 a core team from the two institutions was formed to work out the details of ingest. The team worked over a period of nine months to develop specifications and routines for ingest of IA-digitized volumes generally, and HathiTrust began downloading UC content from IA in April 2010. This paper describes the issues the team encountered during this process and the solutions implemented to create a sustainable large-scale process for ingesting this new content.

2. ISSUES FACED

While partners wanted content digitized by IA to be preserved in HathiTrust, many of the technical, structural, and descriptive characteristics of this content did not match the previously developed standards for materials in the repository. The following are some of the issues the ingest team faced:

Issues related to IA Identifiers:

- The characteristics of primary identifiers would be problematic in HathiTrust systems.

- Filenames differed from HathiTrust conventions.

Resulting Questions:

- What can be used as a primary identifier?
- How will this decision be made?
- What accommodation, if any, will be needed for the different file-naming scheme.

Issues related to IA File Types and Metadata:

- Both raw original and edited page images were present.
- Metadata was located in a number of separate files, and metadata files were not present in a consistent manner between packages. Additionally, none used any obvious schema.
- Some files that the Internet Archive maintained were of undetermined value for preservation; a "preferred" package needed to be identified.

Resulting Questions:

- Is it prudent to preserve raw originals and make them accessible?
- What information captured by IA meets the requirements in the existing HathiTrust AIP specification?
- How do we deal with missing metadata and metadata that does not meet the requirements (e.g. differently formatted dates, invalid MARCXML, etc)?
- What should be done with the metadata in the IA SIP that is not part of the current HathiTrust METS profile?
- What PREMIS syntax do we use to properly record the transformations made to the SIP?

Issues related to IA Page Captures:

- Captured images did not always represent actual page data (e.g. captures of the cradle, tissue papers, and scanning targets).
- Some page types indicated a lack of label authority control (e.g., "Title Page" and "Title" being used to represent the same type of page) or contained errors (e.g. "Norma" instead of "Normal").
- Some required technical and descriptive metadata elements were missing from the image file headers.

Resulting Questions:

- How do we manage structural issues, such as erroneous page types and scanned pages that should not be displayed?
- How do we map the IA page tags to the standard HathiTrust values?
- If image header information is missing, can it be safely and reliably derived from the image data or assumed to be a standard value?

These separate questions led to two overarching issues for the team to address: what transformations would be needed to create HathiTrust-compatible AIPs from IA SIPs, and in what ways could the ingest verification process be modified to accommodate IA-digitized content, but still maintain a high degree of consistency and corresponding reliability for preservation across the repository?

3. SOLUTIONS DEVELOPED

To address these issues, the team of staff members from CDL and Michigan met over a period of months, consulting both with HathiTrust partners and non-partners who had digitized content with IA, to overcome the technical hurdles to ingest. Ensuring the long-term preservation of the digital materials was the highest priority in the development of strategies, with the simultaneous desire to ensure access to the ingested objects. Successful alignment of the Internet Archive SIP to HathiTrust standards required team members to balance the following specific objectives:

- retain components of the SIP that were most useful for preservation and access purposes
- create the most efficient ingest package in terms of size and number of component parts
- maintain functional consistency across the repository
- develop procedures and policies that could be generalized to future types of new content

IA Identifier

One of the first questions the ingest team encountered was whether to continue using IA's primary identifier for volumes as their identifier in HathiTrust. Tagged as "<identifer>" in the object's meta.xml file, the IA ID is used in the names of all files associated with a given object, and is also embedded in the object's URL hosted by IA. While the IA ID works well for Internet Archive's own purposes, the ingest team found it could not easily be integrated into the HathiTrust environment:

- While the majority of IA IDs contained only lowercase characters, several were found with uppercase characters. IDs need to function in case-insensitive contexts in HathiTrust, and team members found that IA IDs were not necessarily unique when lowercased.

- IA IDs had no distinct length. A set of identifiers representing 190,000 objects averaged 24 characters long; a small proportion of this set was found with over 30 characters, and some over 40 characters. Lengthy identifiers would strain the HathiTrust catalog, as well as the pairtree implemented directory structure.
- IA IDs contained embedded semantics: author, title, volume, and scanning facility. Semantics put unnecessary weight on an identifier when the goal is long term preservation. For instance, a string of letters could carry a different unintended meaning in some other time or place.

Fortunately, through collaboration with CDL in developing its processes, IA also generated a NOID (nice opaque ID) for each object [2], prefixed with “ark:/” and written to the meta.xml file. Ultimately, the NOID was chosen to be the primary identifier within the HathiTrust AIP. The original IA ID was retained in the METS for purposes of posterity, but is not used to access the object.

The NOID identification scheme was chosen as a primary identifier for this new ingest type because: 1) The NOID was already embedded in the object’s metadata record. 2) The NOID was created directly by the digitizing agent instead of by a receiving institution. 3) Being an opaque and short identifier, a NOID is unique across all providers 4) NOID supports the ARK (Archival Resource Key) scheme [3], which – although not fully implemented in the current HathiTrust instance – dictates a tight binding between an ARK URL (a combination of a NOID with some name mapping authority) and its metadata.

Ideally, an identifier should correspond to the identifier in use for the physical object, embodied, for instance, in a scannable barcode. Although the identifier scheme decided upon for these books in question did not involve a tight binding between identifier and object, the team believed it arrived at a durable compromise.

File Types and Metadata

One of the most significant issues faced was the difference between the structure and content of the Internet Archive book packages and packages already preserved in HathiTrust. The Internet Archive scanning process creates a variety of files in different formats, and generates significantly different metadata than those produced through Google process, for example, or other locally-digitized content contained in HathiTrust. Files chosen for long-term preservation from IA had to be carefully selected, with attention to both near- and long-term utility and viability.

The ingest team decided to select certain files from the IA SIP for preservation and exclude others. Any file that contained information determined to be valuable was kept. These included primary images in the JPEG2000 format, information describing how raw images were captured and modified, MARC cataloging information, and OCR data. Any file that could be re-created from the preserved content was excluded, such as a PDF version of the book, .GIF images, .DJVU files, and Dublin Core metadata. After some debate, the raw, uncropped page captures were not preserved for several reasons: their value above that of the cropped images was unclear; they required an additional 1.5 to 1.75 times more storage space than the cropped page images, which were already of significant size; and they would need to be processed to be used in the same manner as the cropped images, which HathiTrust did not support.

A set of pertinent files were thus selected for inclusion in the HathiTrust AIP. However, further analysis of IA SIPs found that not all of these files were present consistently in the SIPs. The files were therefore further divided into “core package” files that would be required in each IA SIP and “non-core package” files that were highly desired, but determined in the end to be optional. The package designations were based on the ingest team’s determination of which files were most valuable for preservation and access purposes. The core package contains the image files, OCR data, and the core descriptive metadata and scanning process metadata. The non-core package contains file checksum data, and potentially useful but non-essential scanning process metadata. The team decided to use PREMIS metadata [4] to document any non-core package files that were missing from an SIP. If core package files were missing, the volume would not be ingested.

IA METS Document

Perhaps the most interesting decision made in the process of accommodating the IA SIPs was one to create a separate METS file in the AIP to store the information contained within the metadata files retained from IA, and then discard the original IA metadata files themselves. This was consistent with the existing practice for Google packages, where a Google-produced METS file is stored in the HathiTrust AIP in addition to a functional METS file (the HathiTrust METS) created by HathiTrust for its own use in the repository. A single METS container for information from the IA files would allow the team to save valuable information in a way that simplified management of files and maintained consistency in the repository, both in the overall package specification and in the HathiTrust METS. The HathiTrust METS would therefore not need to be modified to accommodate these new elements. Instead, including some base information from the IA METS file (such as creation date, as is the practice for Google-provided METS file), the HathiTrust METS could be a record primarily of actions and events occurring in relation to an object after its ingestion into the repository, while the IA (or generically, digitization source METS), could function as the record of the digital object prior to ingest. Though previously a peculiarity of Google-digitized content, the idea of combining all information about digital materials prior to ingest in a single file took hold in the IA ingest process, and has become integral to strategies for ingesting content from a variety of digitization sources.

The IA METS is built by parsing the separate metadata files inside the IA SIP and copying their contents into a METS file similar to the one that is part of each HathiTrust AIP. This takes place during a pre-ingest phase, which the team developed to effect all modifications relating to metadata and content in the IA SIP (e.g., image headers), prior to final validation and ingest. The IA METS is similar in format to the HathiTrust METS that is part of each AIP. Most of the IA METS is boilerplate structure, filled in with information downloaded from the IA book package or the objects as they are processed for HathiTrust compatability. The information in the IA METS includes MARC XML, descriptive metadata, OCR information, and metadata about the scanning process – all of which were part of the IA SIP but not necessarily appropriate for inclusion in the HathiTrust METS.

PREMIS Events

The transformations and processes that occur during the pre-ingest transformation are documented in the IA METS using PREMIS metadata in order to maintain the digital provenance record, with the goal of providing additional trustworthiness. The decision was made to employ PREMIS 2.0, as opposed to the PREMIS 1.0 used in Google- and other partner-digitized AIPs, because it allowed for new preservation elements, and repository-wide plans included transitioning all content to PREMIS 2.0. The transformation events include processes such as MD5 validation, IA SIP inspection, image header modification, file renaming, OCR splitting, IA METS creation, and final validation. PREMIS is utilized to document the processes and actions performed, the institution that performed it, and the software tools employed.

Image Headers

Addressing missing image header metadata was somewhat complex. HathiTrust requires JPEG2000 files to have technical and descriptive metadata in the XMP box, but this information was not always present in the IA images. The ingest team decided to use ExifTool to modify and/or populate metadata in the image headers if it could be reliably derived or taken from metadata provided outside the headers. Some of this metadata, such as TIFF:SamplesPerPixel and TIFF:PhotometricInterpretation, could be derived from the bitstream using JHOVE. TIFF:Orientation was assumed to be 1 (which indicates a horizontal, or normal, orientation), as images were captured in the orientation in which it should be displayed.. Some elements were able to be copied from the JPEG2000 metadata elements such as the image width and height. One of the more difficult issues faced was missing JPEG2000 resolution information. Here the team decided to determine the resolution value from data found in the file header in the JPEG2000:CaptureResolution and CaptureResolutionUnit fields. If this was not present the resolution was determined by using information captured in the IA metadata files, which appeared to match the resolution metadata in the header when present.

Page Types

There were a number of issues with individual page captures in the IA SIPs that needed to be resolved. Among the page captures were images of the scanning stand, scan targets, tissue pages, and miscellaneous pages that were tagged as “delete”. A lack of documentation of this portion of the digitization process required the ingest team to deduce what was meant by some of the labels (e.g., identifying tissue pages, blank pages, title pages, tables of contents, etc.). Even after these variations were clarified and misspellings were normalized, these labels did not always neatly fit into the standard array used in the HathiTrust AIP. In the end, original IA page type values were stored in the IA METS and normalized to HathiTrust values during the creation of the HathiTrust METS. Where applicable, some page type values were incorporated as additions into the standard HathiTrust schema for labeling pages. While it would not have been a burden to accommodate IA labels in the HathiTrust access system (where they are used to browse content) instead of normalizing them, the team determined that asking downstream users of HathiTrust content to analyze the different DIPs (Dissemination Information Packages) to understand multiple labeling schemes would unduly inhibit use of the content.

4. FINDINGS AND CONCLUSION

Much was learned in the process of developing successful and appropriate methods for ingesting IA-digitized materials into HathiTrust:

- Documentation of process is essential to downstream uses of content. Significant time was spent by the ingest team in analysis and interpretation of IA processes and digitized content because documentation was not available. In some cases such as foldout images (which are not gone into above) no special action was needed for preservation or display purposes, but extensive investigation was required to determine that this was the case.
- File names are just names, and should not be invested with too much meaning. Much deliberation occurred around filenames but in the end the team decided to use the IA names instead of normalizing them. The issue of primary concern is that metadata

exists to indicate the proper order of files, not the filenames themselves.

- In a collaboration of this size, with expertise required in so many areas, open and clear communication is the key to success. Agendas for meetings, facilitators of conversations, and individuals at each institution to coordinate efforts, talk through issues, and bring in additional team members for perspectives, insights, and expertise as needed, were essential to the success of this project.
- The trustworthiness and effectiveness of a shared repository does not rely on specifications and sound technology alone. They are based as well on the relationships of the people involved in building and sustaining the repository over time. Through the conversations and experiences working together, the teams from CDL and Michigan gained greater trust in one another, and in the methods and processes we use for getting things done. Building relationships through in-person, phone, and video conferences throughout the project helped the team accomplish its goals, and will strengthen HathiTrust in its collaborative efforts going forward.

The collaboration between the HathiTrust partners set precedent for future approaches to ingesting content from new sources. The contributions from each institution and other HathiTrust partners led to a strong shared philosophy on digital preservation and content management.

This type of experience is likely to be far more common as digital repositories seek to expand their stores of digital content to content produced by a variety of providers and partners, while simultaneously attempting to create strict validation routines and a manageable, consistently-structured store of AIPs. It is hoped that this case study will provide a model for other organizations and collaborations to follow as they expand their collections.

A large number of staff from the University of Michigan and California Digital Library contributed to the success of this project. The authors would like to acknowledge their efforts and offer thanks for their contributions to this paper.

5. REFERENCES

[1] TRAC. (2007). *Trustworthy Repositories Audit &Certification: Criteria and Checklist.* Center for Research Libraries and OCLC. http://www.crl.edu/sites/default/files/attachments/pages/trac_0.pdf

[2] NOID: Nice Opaque Identifier (Minter and Name Resolver). (2006). https://wiki.ucop.edu/display/Curation/NOID

[3] ARK: Archival Resource Key. (2008). https://wiki.ucop.edu/display/Curation/ARK

[4] PREMIS: Preservation Metadata Implementation Strategies. http://www.loc.gov/standards/premis/).