Guidelines for Digital Object Deposit

Version 1.0 - January 14, 2011
Version 1.1 - Updated number of partners September 21, 2011
Version 1.2 - Updated number of partners December 10, 2011

Summary

The purpose of these guidelines is to facilitate deposit of digital content from a variety of sources into HathiTrust. The guidelines contain a brief introduction to HathiTrust, a description of its guiding principles and design, an overview of the ingest process, and definitions, policies and procedures related to the ingest of digitized book and journal content and associated metadata. The guidelines are accompanied by a HathiTrust Deposit Form [1], a worksheet for depositors to complete to aid in efficient deposit of materials. The Deposit Form includes detailed specifications and validation criteria for submitted content. A checklist of all steps involved in the ingest process is available at http://www.hathitrust.org/ingest_checklist [2].

Outline

Summary

I. Introduction

II. Principles and Design

III. Ingest Overview

IV. Definitions

V. Policies

- Preservation [9]
- Content Formats [10]
- Preservation Description Information [11]
- Descriptive Information and Copyright Determination [12]
- Validation [13]
- Quality [14]

I. Introduction

HathiTrust is a large-scale digital preservation repository that was launched by a partnership of major research libraries in 2008. There are currently more than 60 partners (a full list can be found at http://www.hathitrust.org/community [15]). The initial focus of the partnership is on preserving and providing access to book and journal content digitized from their collections through a number of means, including digitization by Google, the Internet Archive, and through local initiatives. The partners aim to build a comprehensive archive of published literature from around the world and develop shared strategies for managing and developing their digital and print holdings in a collaborative way. The primary community that HathiTrust serves are the members (faculty, students, and users) of its partners libraries, but the materials in HathiTrust are available to all to the extent permitted by law and contracts, providing the published record as a public good to users around the world.

II. Principles and Design

HathiTrust is guided by principles of trustworthiness, openness and responsible stewardship. It strives to provide reliable long-term preservation of, and access to content in ways that maximize the contributions of partner institutions and make the most efficient use of available resources. The repository was designed according to the framework for Open Archival Information Systems (OAIS) [11] and is realized within the context of community-wide standards and criteria for Trustworthy Digital Repositories [2]. The logistics of operating a preservation repository at the scale of HathiTrust [3] have led to implementation solutions that favor consistency and standardization over variation, simplicity over complexity (in design, not function), and practicality over ideology. HathiTrust functions above all to meet the preservation and access needs of the HathiTrust partners. Although by extension HathiTrust serves a much broader constituency, it is these needs specifically that drive the development of HathiTrust services and capabilities.

III. Ingest Overview

There are two components to ingest in HathiTrust: ingest of bibliographic metadata and ingest of content.

Bibliographic metadata

Institutions must provide accurate bibliographic records for submitted content before content ingest can occur. The records act as a manifest of the digital content and are used as part of HathiTrust's digital object management strategy. Delivery of bibliographic metadata may occur via removable media, download from a source, or other mechanism determined by HathiTrust and the depositor. Specifications for bibliographic metadata are provided in Section II.F [14] of the HathiTrust Deposit Form [1].
Content

HathiTrust distinguishes two phases within content ingest: a pre-ingest phase where content is transformed and possibly remedia ted to repository standards, and ingest proper where validation of content is performed and objects are written to repository storage (including replication to HathiTrust’s redundant storage sites). Common modes of content transfer include delivery via removable media or download from a source. Specifications for deposited content are given in the HathiTrust Deposit Form, Sections III.B and III.C. Policies influencing these specifications are provided below.

IV. Definitions

The definitions below are adapted from the OAIS reference model and Brian Lavoie’s Introductory Guide to OAIS.

Archival Information Package (AIP) – The Information Package, consisting of the Content Information and the associated Preservation Description Information (PDI), that is preserved in HathiTrust.

- Core Information Package (CIP) – The set of files that must be present in an SIP and included in the AIP for a package to be eligible for ingest.
- Preferred Information Package (PIP) – The Core Package plus the set of files may be present in an SIP and included in the AIP for the package to be eligible for ingest.

Submission Information Package (SIP) – The Information Package that is delivered to HathiTrust for use in the construction of one or more AIPs.

Workstream – A Workstream is a group of SIPs that are similar in their technical characteristics (content formats, included metadata, digitization workflow, etc.).

Collection – A Collection is a group of materials within a Workstream that are related to one another by subject, date, or other categorization as understood by the depositor.

Content Information - The set of information that is the original target of preservation. It is an Information Object comprised of its Content Data Object and Representation Information.

- Content Data Object – The data object that, together with Representation Information, is the original target of preservation (in HathiTrust currently, page image files and associated OCR files and metadata)
- Representation Information – The information that maps a Data Object into more meaningful concepts (this includes at a very granular level standards such as Unicode and TIFF).

Preservation Description Information - The information that is necessary for adequate preservation of the Content Information and which can be categorized as Provenance, Reference, Fixity, and Context information.

- Provenance Information – Documents the history of the Content Information, including its creation, any alterations to its content or format over time, its chain of custody, any actions (such as media refreshment or migration) taken to preserve the Content Information, and the outcome of these actions.
- Reference Information – Uniquely identifies the Content Information within HathiTrust (e.g., repository identifier), as well as in relation to entities and systems external to HathiTrust (e.g., OCLC number, ISBN, etc.). Note that some reference information, such as ISBN, OCLC number, and other standard identifiers, are contained in MARC records with the Descriptive Information.
- Fixity Information - Validates the authenticity or integrity of the Content Information: for example, a check sum, a digital signature, or a digital watermark.

Packaging Information – Information that is used to bind all of these information components into a single logical package. Packaging Information serves to associate all of the various components of an AIP, permitting them to be identified and located as a single logical unit within the archival system.

Descriptive Information - Information that supports the discovery, retrieval, and ordering of Content Information in HathiTrust (bibliographic and rights metadata).

V. Policies

Preservation

Preservation in HathiTrust encompasses the content characteristics, metadata, and processes that enable us to maintain the bit-level integrity of content over time and migrate content to new formats as technology, standards and developments in the library community necessitate.

From the HathiTrust website (http://www.hathitrust.org/preservation):  

HathiTrust is committed to preserving the intellectual content and in many cases the exact appearance of materials digitized for deposit...HathiTrust is committed to bit-level preservation and format migration of materials...as technology, standards, and best practices in the digital library community change.

HathiTrust strives to ensure that the digital content it preserves is accurate, complete, suitable for long-term preservation, and useful for a variety of access purposes. It does this through consideration of content file formats, preservation and descriptive metadata, validation...
Content Formats

1. HathiTrust supports formats for Content Data Object Files
   - That are widely held standards in the library community
   - That have proven to be robust as far as carrying digital data and minimizing loss of that data over time
   - That enable multiple alternative and downstream uses (such as print-on-demand) with high confidence
   - For which, should a more robust format be developed or technological or other developments dictate the need to alter or abandon the format standards, there is high confidence that the library community as a whole will devote energy and resources to forward migration paths

2. We support four Content Data Object formats currently: TIFF ITU G4, JP2 (JPEG 2000 part 1), JPEG, and Unicode OCR with and without coordinates. The first three are base image formats from which the other can be derived.

3. We do not currently (as of September 2010) support schemas or DTDs that describe publication structures (e.g., DocBook, TEI, ePub).

4. We do not include derivative formats in the AIP unless there are compelling preservation or access reasons.

5. We do not perform additional processing to de-skew or crop files for display to end-users. Post-processing may be done (typically by recipients of content) “downstream” to support specific purposes such as print on demand.

Preservation Description Information (PDI)

1. PDI in HathiTrust is typically stored in two files conforming to the Metadata Encoding and Transmission Standard (http://www.loc.gov/standards/mets/), which contain Provenance, Reference, Fixity, and Context Information for Content Information as follows:
   - A “Source” METS file is assembled from metadata provided to HathiTrust in the SIP, and contains information about the Content Information from the time of its creation to the time it enters the repository;
   - A “HathiTrust” METS file is created on ingest and includes a subset of the Source METS file data, but is primarily a record of the object from the time it enters the repository forward. The Source METS is kept for preservation purposes only. The HathiTrust METS is used for both preservation and access purposes (i.e., in both the archival and dissemination information packages).
   - More information, including sample METS files is available at http://www.hathitrust.org/standards/mets/


3. Reference Information – Each HathiTrust AIP has a single primary identifier in the repository, which is composed of two parts: a namespace, and an item identifier, (e.g., physical volume barcode). Namespaces are selected by the contributing institutions and delineate content in the repository contributed by that institution from content contributed by other institutions. If items from an institution have more than one distinct identifier scheme, multiple namespaces are used. Namespaces do not delineate content based on technical characteristics, so multiple Workstreams may exist within a single namespace. Some examples of current HathiTrust identifiers include:
   - University of Wisconsin: wu.89094366424
   - University of California: uc1.3543486
   - University of California: uc2.ark:/13960/t26973133
   - University of Michigan: mdp.39015037375253
   - University of Michigan: mdp.39015037375253
   - University of Wisconsin: wu.89094366424
   - University of California: uc1.3543486
   - University of California: uc2.ark:/13960/t26973133
   - University of Michigan: mdp.39015037375253
   - University of Michigan: mdp.39015037375253

We prefer to use item identifiers that are currently in use at the contributing institution because

- introducing another identifier scheme requires mapping and/or bulk updates to existing records and ongoing additional maintenance, and
- by leveraging an existing identifier scheme, institutions may more easily make references to HathiTrust representations of items by automated means

If such identifiers are unavailable, or undesirable because they do not have good identifier properties, a vendor or secondary identifier may be used instead. If valid identifiers do not exist, HathiTrust may be able to assign them. Good identifier properties include:

- guaranteed uniqueness
- a deterministic process for creating new identifiers
- an internal check scheme
- an explicit, existing association to the physical item to be digitized (such as an item barcode)
- either an accurate correlation or no correlation to existing names or other identifying characteristics of the item (i.e. no implied or misleading relationships)

Further details are provided in the volume identifier specifications in Section III.B of the HathiTrust Deposit Form.

Other Reference Information such as ISBN, OCLC number, and other standard identifiers may be present in the Descriptive Information (bibliographic metadata) that is stored in the Source and HathiTrust METS files. However, these identifiers are stored for preservation
purposes only and are not used actively in access systems. The specifications for bibliographic metadata in Content Data files are less directive than bibliographic metadata submitted to the bibliographic metadata management system and need only be valid MARCXML. Specifications for Descriptive Information submitted to the bibliographic management system are provided in Section II.F [19] of the HathiTrust Deposit Form [19].

Descriptive Information and Copyright Determination

1. Bibliographic data must be submitted to HathiTrust before ingest of corresponding AIPs can begin, and should follow the specifications listed in Section II.F [19] of the HathiTrust Deposit Form [19].

2. Bibliographic Data is used to make an initial automated rights determinations for every deposited item, which are saved in a Rights [26] Database [26] (not in the AIP). Bibliographic Data should therefore be as complete and accurate as possible. If insufficient data exists to make a rights determination, volumes default to a status of in copyright, with the same limitations on visibility and use that apply to in copyright volumes. Details on the bibliographic fields used and processes involved in automatic rights determination can be found in Section II.F [19] of the HathiTrust Deposit Form [19].

3. Rights information and permissions pertaining to specific volumes may be attached to the Digital [27] Assets [27] Submission [27] Inventory [27], which must accompany each deposit. A Permission [28] Agreement [28] is available for rights holders to open access to their works in HathiTrust, including the assignment of Creative Commons licenses, and to enable the production of reprints. An outcome of review, changes are sometimes made to bibliographic data submitted by partner institutions. Determinations resulting from the review process are also saved in the Rights [28] Database [28].

4. Staff from multiple HathiTrust institutions are involved in ongoing manual copyright review of volumes published in the United States between 1923 and 1963. More information about this work can be found at http://www.crl.edu/sites/default/files/attachments/pages/trac_0.pdf. As an outcome of review, changes are sometimes made to bibliographic data submitted by partner institutions. As an outcome of review, changes are sometimes made to bibliographic data submitted by partner institutions. Determinations resulting from the review process are also saved in the Rights [28] Database [28].


Validation

1. It may not be possible to validate some components of the AIP (e.g., coordinate OCR in hOCR format). In these cases, we make limited exceptions based on need and the co-occurrence of open specifications (i.e., the presence of the same content in an open format specification). We are not able to commit to migration of materials that it is not possible for us to validate.

2. There are no processes currently in place to test the well-formedness or intellectual value of OCR text files beyond assuring it is valid UTF-8. The partners understand that OCR detection is an evolving area and include OCR text because of its value to users in searching and accessing materials.

Quality

1. Although strategies to expand quality review processes in HathiTrust are in development, quality review remains the responsibility of partner institutions.

2. Quality is a high priority for HathiTrust and we take action where possible to prevent content of poor quality from entering the repository. In addition to quality assurance that may be performed by partners on the content they submit, we use quality metrics provided by Google for volumes it has digitized to gate volumes that do not meet a certain quality threshold. We have engaged with Google and other Google library partners on an ongoing basis to improve the metrics and processes Google employs. We also work with Google to make corrections and improvements to individual volumes based on user feedback.

3. We are exploring opportunities to extend the manual review that partners undertake on their own content more broadly across the repository and develop criteria to certify content in HathiTrust as useful for distinct purposes (see http://www.crl.edu/sites/default/files/attachments/pages/trac_0.pdf). We are also investigating ways to communicate aspects and characteristics of volumes that effect quality (such as missing pages, crop errors, warp errors, etc.) in the user interface.

Source URL: http://www.hathitrust.org/deposit_guidelines

Links:
[3] As of May 2010, HathiTrust contained more than 6 million volumes, more than 1 million of which are in the public domain.