HathiTrust Preferred Internet Archive Book Package

Overview & Background

As a by-product of the Internet Archive scanning process, a variety of different files and formats are available to everyone, everywhere. This differs from the Google output, which offers no file-level variations or options. However, this also means that files chosen for ingest into the HathiTrust repository must be carefully selected, with an eye towards both near-term and long-term utility. The process of selecting files that is described below attempted to balance the following important criteria: a baseline, cross-partner standard; functional consistency with the Google work products; a desire to keep the highest quality master images; a disinclination to discard useful information; and an attempt to minimize overall package size to reduce storage costs.

Ingest into the HathiTrust repository will require pre-processing of the original file set described below in order to normalize files to an expected format. This normalization will allow HathiTrust processes to accommodate content from all partners. This process is currently in development and a link to the documentation of the process will be included here, once it is finalized.

File Selection Criteria

In the following section, the files selected for ingest into the HathiTrust repository are identified, along with a justification for why they were selected. Also listed are files that are available from the Internet Archive, but have not been selected. A description of each file can be found in the All Available Files & Characteristics section below. All files below are named using the Internet Archive identifier, preceding the underscore (ex. balladsotherpoem00tennrich_jp2.zip).

Preserved Files:

- _jp2.zip/tar - These are required for most access services built on the book content. These image files represent the highest quality, unaltered images that are available to us as a result of scanning and do not require significant pre-processing before they can be used for access services.
- _djvu.xml - Future access services could require coordinate data for added functionality. Also valuable for accessibility reasons.
- _djvu.txt - Current indexing processes require text to provide robust search and retrieval capabilities. Likewise valuable for accessibility reasons.
- _meta.xml - Provides a basic level of self-description required for preservation. Also provides a set of important identifiers and some metadata about the scanning cycle, which will feed into the PREMIS data recorded in the HathiTrust METS.
- _marc.xml - Provides a full description of the object.
- _files.xml - A manifest of all files available for the scanned object.
- _scandata.xml - Tracks the modifications made to the raw image captures to get them into the _jp2.zip state.
- _scanfactors.xml - Records problems detected with the raw image captures and the steps taken to correct them. Important for facilitation of future automated error correction and possible selective re-scanning of pages.

Excluded Files:

- _.pdf - This file can easily be re-created using the page image files, the OCR text and open source software.
- _bw.pdf - This can also easily be re-created using the page image files, the OCR text and open source software.
- _.djvu - Can be re-created, with appropriate proprietary software.
- _.gif - Can easily be re-created with open source software.
- _raw_jp2.zip - (and _orig_jp2.zip) The original, unprocessed (other than format conversion) image captures. These are much larger files per image (>240%) than other available images and also must be modified/transformed before they can be used for display purposes.
- _wm_jp2.zip - (and lib_jp2.zip) Modified image files cannot have any watermarking removed without extensive processing; the watermark applied may not be desirable.
- _abby.gz - Very large file that contains more information than access services will need; the file size would probably slow down access; may possibly be re-created in the future (by re-OCR'ing page images).
- _dc.xml - May be recreated from existing metadata (marcxml and metaxml files).
- _meta.mrc - An XML version of this file is being captured; the binary version does not offer any additional benefits.
- _metasource.xml - The bibliographic metadata contains the information to determine the source
- _scandata.zip - Contains a great deal of extraneous data that describes portions of the workflow that are unnecessary for understanding the object and its formatting

Rationale for Excluding Raw JP2 Files

While it may seem advantageous to include the raw JP2 files in the set of preferred files, there are some practical reasons against doing so. First and foremost is the size of the raw JP2 ZIP, which averages 264.33 mb per book. This increases the overall package size by 240% (to ~453.61 mb) and will more than double the cost per partner for each book stored. The HathiTrust system balances both preservation and access services and this focus has influenced this decision. The raw JP2s would need extensive (and computationally heavy) processing to prepare them for use in access services, which would either significantly slow down the ingest queue (if processed during ingest) or would affect response time for end users (if processed on-the-fly at display time).

Preferred File Set and Size Estimates

Preferred Files	Description	Average Size
_jp2.zip/tar	zip/tar of all processed (rotated, cropped, etc.) JP2 page images	176.74 mb*
_djvu.xml	OCR with coordinates as a single file	4.98 mb
_meta.xml	descriptive metadata; see below for more details	2.04 kb
_marc.xml	MARCXML as retrieved using z39.50	2.68 kb
_djvu.txt	Raw OCR as a single file	560.53 kb
_files.xml	list of all available files	16.36 kb
_scandata.xml	information about each page image's scan event	143.7 kb
_scanfactors.xml	page errors and automated corrections	.36 kb
	Total for All Books	189.28 mb

* This figure also includes JP2 archives that are .tar

Core vs. HathiTrust Preferred Book Package

Some of the files included in the preferred file set above are more important for both long term preservation and access than other files and it is critical that those files are present to represent a sense of a digital surrogate for a book. For this reason, we can differentiate the files above into 'required' (core) files and 'desirable' files. Given that past experience has shown file retrieval from Internet Archive servers to be erratic and at times inconsistent, ingest should not fail if a file deemed desirable is missing or unavailable, so long as failure occurs if a required file (ex. the image archive) is missing.

We propose to define the book package in two ways, as follows:

Core Book Package Files:	**HathiTrust Preferred Book Package Files:**
_jp2.zip/tar	core book package files, plus:
_djvu.xml	_files.xml
_djvu.txt	_scandata.xml
_marc.xml	_scanfactors.xml
_meta.xml	

Ingest will attempt to retrieve and store the complete book package in the repository even if all files are not available. Ingest will be considered successful only if all of the core files are retrieved and saved. This allows for two varations of ingest success: complete ingest and core ingest.

All Available Files & Characteristics

- PDF:
  - .pdf - The presentation version in PDF.
  - _bw.pdf - A black and white PDF compiled using binarized versions of the images. The binarized images are not made available.
- .djvu - Similar to PDF, a proprietary compressed document format.
- _flippy.zip - An open source, image-based page turner format to be used with the Open Library's Flip Book (beta) software; more information can be found at http://openlibrary.org/dev/docs/bookreader
- Images:

- .gif - A looping, animated thumbnail of the first 20 pages of a book. Usually a 100x152 pixel GIF.
- _jp2.zip - A ZIP archive of all of the cleaned, cropped, etc. JP2 page images. These are the highest quality, least modified images that are available after the raw/orig file set.
  - .tar - Some books are so large that the volume of images exceed the maximum size for a ZIP archive. For these books, the images are compressed and delivered using TAR. These TAR archives average 2.07 gb and occur .39% of the time (738 out of 191,568 books total).
- _raw_jp2.zip - A ZIP archive of the uncropped, unprocessed page images. The original raw camera capture format has been saved as a JP2 with minimal processing, so this is not the file straight from the camera. Alternately, this file may be named _orig_jp2.zip.
- _wm_jp2.zip - A ZIP archive of the clean images that includes a watermark. Alternately, this file may be named _lib_jp2.zip. The watermark varies by institution; the UC version reads:
  Univ Calif - Digitized by Microsoft ¨

- OCR:
  - _abby.gz - GZipped version of the full ABBYY FineReader XML output, which includes all character-level information (confidence, location, etc.)
  - _djvu.xml - OCR text formatted as XML that includes word coordinates and rough page formatting (column, paragraph, line, word, etc.)
  - _djvu.txt - the raw OCR text
- Metadata files:
  - _dc.xml - OAI record in Dublin Core XML
  - _meta.xml - Internet Archive's internal "management" metadata; a proprietary XML format, this file includes information about the scan event (date, # of pages, operator, station, etc.), the contributor, basic bib data (title, author, subject, language), and a set of identifiers
  - _meta.mrc - the binary MARC record as retrieved using z39.50
  - _marc.xml - the MARC data in XML
  - _files.xml - the manifest that records all of the files available for this book; also gives 2 checksums and a format definition for each file; provides the only mechanism for validating that the component data has been downloaded successfully
  - _metasource.xml - a proprietary XML file recording where the MARC record came from (catalog, operator, zquery, etc.)
  - _scandata.xml - a proprietary XML file recording information about each page image (handSide, cropBox, original width & height, etc.)
    - often this in contained in a scandata.zip archive that also collects logs of the scanning session (for the scribe) and the transfer of the files to the data center, along with some lighting target captures, camera profile information and a ZIP archive of the original data captured during scanning (as the individual XML files, used to build _scandata.xml)
  - _scanfactors.xml - a proprietary XML file recording page errors and automated corrections; this file may not be present, although 0 error files do exist