Specifications for submitting print holdings updates to HathiTrust
10/14/22

Introduction

The purpose of this specification is to describe the requirements governing the naming, format and content of print holdings submission files. These requirements enable HathiTrust to automate the processing of these files.

A print holdings submission consists of 1-3 files.
Each file must conform to:

- the File format convention
- the File naming convention
- the Header line convention

... and must be uploaded to the current-year sub-folder under "print holdings" in your HathiTrust DropBox account.

Submitted files that do not conform to this specification may be rejected by HathiTrust. HathiTrust staff will work with you to get the files into acceptable format, where possible.

General guidelines

HathiTrust requires print holdings information from partner institutions in order to:

1. Support analysis of the overlap of institutions' print holdings with digital holdings in HathiTrust. These calculations are used to generate annual member fees.
2. Facilitate collaborative collection development and management operations.
3. Support the HathiTrust Shared Print Program.
4. Enable access via the Emergency Temporary Access Service.

Files should be uploaded to the Dropbox folder for your institution, under:

[member]-hathitrust-member-data/print holdings/2022 holdings

For questions about Dropbox access, please contact support@hathitrust.org.

Please also see our Print Holdings FAQ page.

Content guidelines

Record export

Please provide holdings information for books and book-like materials (e.g., pamphlets, bound newspapers or manuscripts) that are:

- in print format
- have OCLC numbers and
- are cataloged as a single unit

Do not send holdings records for microform, eBooks, or other non-print materials.

Record type

Single-part monographic holdings, multi-part monographic holdings, and serial holdings should ideally be submitted in separate files (see Item Type).

For monographic holdings: Please include separate records for all print holdings, including multiple copies of the same title. Each record should appear as a separate row in the file (since they can have separate conditions and statuses).

For serial holdings: Please include a single title-level record for each holding. Each record should appear as a separate row in the file.

For bound-withs, when possible, please include all constituent OCNs and place each on a separate row.

Format

The data should be provided in tab-delimited text format.

File format convention

Print holdings files submitted to HathiTrust must be tab-separated text files and have the `.tsv` file extension.[1] Each field value must be delimited with a tab, and each record must appear on a new line. Field values themselves must not contain any tabs.

Please note that simply changing the file extension of a file to .tsv does not make it a tab-separated text file. If you do an export to tab-separated text from e.g. Excel, please check the exported file before submitting to ensure that large OCLC numbers and Local IDs were preserved. If you see OCLC numbers or Local IDs that look like "1.79699E+11" then that might indicate that numbers were changed into something unusable during export.

Files may optionally be compressed with gzip, in which case the `.gz` file extension is added to make `.tsv.gz`.

[1] Tab-separated text is a non-proprietary format that is both human- and machine-readable. For more information about tab-separated text files, see https://www.loc.gov/preservation/digital/formats/fdd/fdd000533.shtml

File naming convention

The purpose of the filename is to identify which member institution it belongs to, when it is from, and hint at what it contains.

The file name consists of 5 required parts, illustrated below with <>'s, and 2 optional parts, illustrated with ()'s. Parts are separated from their preceding neighbor by an underscore ("_"), except file extensions which are separated by a period (".").

The parts should be in the order:
`<member_id>_<item_type>_<update_type>_<date>(_rest).<file_ext>(.compression)`

Part name	Required?	Description
`member_id`	Y	Your organization's HathiTrust member_id.
`item_type`	Y	One of: "mix", "mon", "spm", "mpm" or "ser". See Item type.
`update_type`	Y	One of: "full" or "partial". See Update type.
`date`	Y	A date string, in 8-digit YYYYMMDD format following the Gregorian calendar. Ideally a date with relevance to the file, such as when it was requested, generated, last edited, or uploaded.
`rest`	N	An optional string, which can contain anything and will be ignored. For use by member institutions for their own purposes as needed.
`file_ext`	Y	File extension, `.tsv`, to indicate tab-separated values.
`compression`	N	An optional second file extension, to indicate that the file is compressed (`.gz` for gzip).

Examples:
`test_mono_full_20210530.tsv`
`test_mono_full_20210603_ocnfix_version2.tsv.gz`

Header line convention

Each submitted file must have a header line.

The purpose of the header line is to identify how many columns there are in the file, and to indicate which column is which. This helps make the file both human-readable and machine-readable, allows automatic validation of the data in a given column, and removes the need for guesswork.

The number of columns should be consistent within a file. If the header of a file identifies 5 columns, then all subsequent lines in the file should also have 5 columns. If you need to put multiple values (e.g. OCLC numbers or ISSNs) in one field, check the detailed column specification for which delimiter to use.

The oclc and local_id column headers are required in all files. The rest of the fields are optional and may not be applicable depending on the item type of the file. Fields that aren't applicable for a given item type should not be included, e.g. do not include an issn column in a spm-file or a status column in a ser-file.

Do not include optional columns that are completely empty. That is, if you are not sending any status information in your spm-file, then please omit the entire column from that file.

Which columns are optional/required/not-applicable in which file:

Name	spm	mpm	ser	mon	mix
oclc	req	req	req	req	req
local_id	req	req	req	req	req
status	opt	opt	n/a	opt	n/a
condition	opt	opt	n/a	opt	n/a
enum_chron	n/a	req	n/a	opt	n/a
issn	n/a	n/a	opt	n/a	n/a
govdoc	opt	opt	opt	opt	opt

Each column is described in further detail in the section Detailed column specifications.

Item type

The item type must be included in the filename. It is used to indicate what kind of holdings records the file contains.

There are 5 item types:

- **spm** for files containing single-part monograph records only
    - allowed fields: ocn, local_id, status, condition, gov_doc
- **mpm** for files containing multi-part monographs records only
    - allowed fields: ocn, local_id, status, condition, enumchron, gov_doc
- **ser** for files containing serials records only
    - allowed fields: ocn, local_id, issn, govdoc
- **mon** for files containing both single-part and multi-part monograph records
    - allowed fields: ocn, local_id, status, condition, enumchron, gov_doc
- **mix** for files containing a mix of single-part monographs, multi-part monographs and serials
    - allowed fields: ocn, local_id, gov_doc

Ideally, single-part monographs are separated from multi-part monographs and serials.
In this case the holdings are split into 3 different files, using the item types:

- **spm** for single-part monographs (this file does not have an enumchron column)
- **mpm** for multi-part monographs (this file has an enumchron column)
- **ser** for serials

If it's not possible to separate single-part from multi-part monographs, the single-part and multi-part monographs can be put together in one file, with serials in another. In this case using the item types:

- **mon** for single-part monographs and multi-part monographs, enumchron field optional
- **ser** for serials

If it is not possible to separate monographs from serials, all holdings records can be put in the same file. In this case using the item type:

- **mix** for all holdings

The serials-file differs from the other two in that the serials-file should only list records at the title level, but monos-files and multis-files should if possible list each copy held. Serials-files may contain issn. Serials-files must not contain status, condition or enum_chron.

Update type

Update type must be part of the filename and should be either "full" or "partial".

A "full" update file contains your full print holdings (for the given item type), and HathiTrust should delete your previously submitted print holdings and reload with the ones in the file.

A "partial" update file contains only new (as in not previously submitted) holdings, and HathiTrust should add them to your previously submitted holdings. Records in a partial update that match previously submitted holdings will be ignored.

Note that HathiTrust currently accepts full update files only. Please use “full” in the filename.

Detailed column specifications

The oclc column

Required for: all item types.
Used for:

- Partner fee calculations - overlap analysis of member member print holdings with HathiTrust digital holdings.
- Collection analysis, collection management, collection development.

Each OCLC number should be a continuous string of digits with no intervening spaces. Multiple OCLC numbers should be delimited with a comma or semicolon. The list of accepted OCLC prefixes is as follows:

Prefix	**Description**	**Example**
`ocl7`	7-digit numbers.	`ocl71234567`
`ocm`	8-digit numbers.	`ocm12345678`
`ocn`	9-digit numbers.	`ocn123456789`
`on`	10- or more digit numbers.	`on1234567890`

`(OCoLC)` or `OCoLC`	Frequently used to prefix oclc numbers in 035 fields, with or without stripping the other prefix.	`(OCoLC)12345`
		`OCoLC67890`
		`(OCoLC)ocm12345678`
		`(OCoLC)ocn123456789`

	No prefix.	`12345678`

Please include only OCLC primary numbers, not institution (IR) OCLC numbers.

Please do not include any other types of numbers that occur in the 035 field of records, as they may be erroneously interpreted as OCLC numbers.

If possible, deduplicate similar OCLC numbers (similar in the sense that their numeric part is the same), so that instead of submitting:

```
ocn000000001,ocn000000001,(OCoLC)1,000000001
```

... just submit one of them (which one does not matter), e.g.:

```
000000001
```

Records without an OCLC number will be ignored.

The local_id column

Required for: all item types
Used for:

- Tracking updated holdings submissions over time.
- Matching Shared Print Commitments to specific Print Holdings records.
- Communicating/reporting back to you about specific Print Holdings records.

Requirements:

Either the bibliographic system ID or holdings ID is acceptable, as long as it is used consistently in submissions over time.

The status column

Optional for: files with item type spm, mpm & mon.
Used for:

- Determining eligibility for inclusion in the HathiTrust Shared Print Program.

Values accepted:

Value	**Meaning**
`CH`	**C**urrent **H**olding
`LM`	**L**ost or **M**issing
`WD`	**W**ith**D**rawn
	No value given

Records where no status value is given, as well as records in a file that lacks a status column, will be assumed by HathiTrust to be currently held and internally assigned `CH` status.

If you do not have information that can be mapped to these values, please do not include a status column. HathiTrust will assign those holdings `CH` status.

The condition column

Optional for: files with item type spm, mpm & mon.
Used for:

- Determining eligibility for inclusion in the HathiTrust Shared Print Program.

Values accepted:

Value	**Meaning**
`BRT`	**BR**i**T**tle, damaged and/or deteriorating
	No value given

Records where no condition value is given, as well as records in a file that lacks a condition column, will be assumed by HathiTrust to NOT be brittle, damaged and/or deteriorating.

If you do not have information that can be mapped to these values, please do not include a condition column.

The enum_chron column

Required for: files with item type mpm.
Optional for: files with item type mon.
Used for:

- Volume-level overlap analysis for multipart monographs. Providing enumeration and chronology information will enable more precise matching for multi-part items, likely reducing calculated fees. It will also increase precision for holdings-based services such as ETAS.

Requirements:

There is no standard format requirement for enumeration and chronology. When providing this data, please draw from item-level enumeration and chronology fields rather than unrelated fields, e.g. SuDoc, or call number.

If you hold several volumes of a title, please list each on a separate line. That is, if you hold 2 volumes of *Marketing dairy products*, please list them as:

```
oclc        local_id    enum_chron  ...
1011851340  b567        no.1 1922   ...
1011851340  b678        no.2 1922   ...
...         ...         ...         ...
```

Please do not include:

- Copy count details, e.g. "volume 1, copy 2"
- Shelving location
- Purchase price
- Call number
- etc.

The issn column

Optional for: files with item type ser.
Used for:

- May be used in the future as a secondary check on overlap analysis for serials.

Requirements:

Multiple ISSNs should be separated by a comma or semicolon.

ISSNs will only be accepted if matching `DDDD-DDDC`, where:

- The hyphen is optional.
- `D` is a single digit, 0-9.
- `C` is a single digit 0-9, or the letter "X".

Example:

```
local_id    oclc        issn
7113730     6415579     0022-362X
7113751     9974250     8755-0393
7113730     642030352   2041-7365,2041-7373
...         ...         ...
```

The govdoc column

Optional for: all files.
Used for:

- Collection analysis, in conjunction with HathiTrust's U.S. Federal Government Documents Program and the HathiTrust Shared Print Program

Values accepted:

Value	**Meaning**
`0`	not a U.S. federal government document
`1`	is a U.S. federal government document
	No value given.