# Ingesting Our Internet Archive Books Into HathiTrust

### Introduction

We deposit \*almost\* all of our Internet Archive volumes into the HathiTrust.

### Hathi Workflows

Hathi has copious documentation about how to deposit content at www.hathitrust.org/ingest.

In a nutshell, we simply submit one ALMA record per volume. Prior to submission, the ALMA record is enhanced with a 955 field containing (1) the book id, (2) the ark id, and (3) the volume number.

### Local Workflow

Our local workflow is tracked in an ACCESS database and is designed to accomplish the following:

* Keep track of what has already been submitted so we make duplicate submissions
* Keep track of which books were digitized in house and which were digitized by the IA, as they have to be submitted separately.
* Monitor whether or not HT ingests have failed or been successful
* Follow up on failed ingest
* Pull together three key data elements to enhance ALMA records for submission to Zephir
  + ALMA ID and IA Book ID (on returned picklists)
  + Ark id (in the volume’s meta.xml file)

### Special Requirements for Volumes Digitized In-House

Volumes digitized in-house need to be deposited separately from the books digitized by the Internet Archive. Special file naming conventions are used so that Hathi knows to indicate that the volume was digitized by Boston College (as opposed to the IA). In addition, the ppi for the volume must be updated in the Internet Archive scandata.xml file for the volume.

### Identifying New Volumes for Deposit

How do we know what is and is not in Hathi? There are a few ways and all of them are useful.

1. A record of the previously deposited IA book identifiers is maintained in an ACCESS database. This ACCESS database includes other information and is currently maintained on Betsy’s desktop.
2. Items in a newly completed IA picklist will not be in Hathi
3. Using ARK ID to join the files together, compare all the volumes in the Internet Archive to all the volumes in Hathi. This is a good thing to do periodically, to make sure nothing has been missed.

### HathiFiles Description

## The [HathiFiles Description](https://www.hathitrust.org/hathifiles_description) page gives a field-by-field description of the elements in the tab-delimited HathiFiles and is good to have on hand for reference.

### Local Workflow: Step by Step

⮊ Make sure all volumes from the previous month’s submission were ingested successfully.

Hathi publishes [tab delimited data files](https://www.hathitrust.org/hathifiles). The files contain a row for each volume in the repository. The file for first of each month is cumulative (eg. hathi\_full\_20160801.txt.gz).

* Use [bc.pl](https://github.com/BCDigLib/BC-HT-IA-Ingests/blob/master/bc.pl), a script on Github to pull out all the BC rows.
* Import the resulting text file as a new table in the IA-Hathi-Deposit-Tracker database.
* Delete the “bc.” prefix from the Hathi volume identifies
* With the ark id as the key, create a query that finds all entries in the Hathi table and all entries in the Hathi-IA-Master-Log where HTrecNo in the master log is null
* Examine the results and, as appropriate, update the Hathi-IA-Master-Log.status to “ingested” and the HTrecNo to the Hathi Trust Record Number.

⮊Find out what’s new in the IA and add it to the IA-Hathi-Deposit-Tracker database

Use the Internet Archive’s [*Advanced Search returning JSON, XML, and more*](http://archive.org/advancedsearch.php) to get information for all of our volumes. The most reliable search string is:

* Collection:Boston\_College\_Library

Select the option to download into CSV format (allow more than the total number of books we have in the IA). Select the following data elements: identifier; identifier ark; contributor; image count. Sort the results in Excel to remove the collection identifiers (they will have no page count).

After identifying all the IA volumes, use ACCESS to find the identifiers not yet in Hathi and add them to the Hathi-IA-Master-Log table in the database.

⮊Get the ark ids and other metadata from the Internet Archive

Put the new book ids (identified in the previous step) in a text file and run get-ia-metadata.pl to get the xml metadata for each volume the Internet archive.

Run the xslt hathi-meta.xml to extract a pipe delimited file of the following metadata elements:

identifier|ark|link|volume|title

Import the resulting file into a temporary table in the database and using the identifier as a key, run an update query that adds the ark, link, volume, and title to the appropriate rows of the Hathi-IA-Master-Log table.

⮊Obtain ALMA numbers and add them to the master log

Review the new rows in the Hathi-IA-Master-Log table (the status field will be blank). Import the relevant new completed picklists for the new titles to the Full Picklists table for that year. Next use a query or manual update to add the picklist file name.

Run an update query that uses identifier as the key to update the picklist name and the ALMA number to the Hathi-IA-Master-Log table.

⮊Mark volumes digitized in house

Sort by picklist. For volumes digitized in house, add “in house” to the notes field.

⮊Sort on the status field of the master log

For rows where the status is blank add “pending” or “pending (in house)” if you are ready to ingest. Add other appropriate status values where problems exist.

⮊Export data needed to enhance our local MARC records

Depending when whether you are working on an ingest for “pending” or “pending (in house)”, export the data that is needed to prepare our local MARC records for submission: (alma number;book id;ark;volume) to a tab delimited text file.

⮊Get MARC records from ALMA

Retrieve all the MARC records with a 901 digitized from ALMA. Retrieve in the .mrc (binary) format. Use MarcEdit to convert them to .mrk format

⮊Enhance ALMA records

Run modify-marc.pl to add 955 fields to ALMA records using the text file export from the ACCESS database.

⮊Check to make sure all ALMA records in the batch had their 901 fields

Use the MarcEdit report “Field Count” to make sure there newly created .mrk file has the same number of 001 and 901 fields. If it does not, figure out which 901(s) are orphans and re-do the ALMA export.

⮊Validate MARC records

Use the Report called MARC validator to make sure all of the records are valid. Ignore errors in the 0xx fields. If there are errors in other fields, make a pdf of the report and send it to Kelly so the ALMA records can be upgraded. After records are upgraded, export ALMA records and run through the process of enhancing them again.

⮊Convert and rename file

Convert the resulting file to .mrc and then to MARCxml. Rename the file according to Hathi’s guidelines which vary depending on whether the volume was digitized by us or by IA (or even born digital!)

Digitized by us: mchb\_ia-bc-loc\_yyyymmdd\_bc.xml

Digitized by Internet Archive: mchb\_ia-bc\_ yyyymmdd \_ia.xml

⮊FTP file to California Digital Library using Coreftp

⮊Send message to CDL staff as outlined in HT documentation

⮊Send message to HathiTrust staff, indicating file was submitted to Zephir

Indicate whether or not the volumes were digitized in house or at the Internet Archive. If digitized in house, include book ids and ark ids in the message.

⮊In the database, change the status of these volumes to submitted with the data

Also add the file name of the submission file in the appropriate field