# Ingesting Our Internet Archive Books Into HathiTrust

## Introduction

We deposit all of our Internet Archive volumes into the HathiTrust. Hathi has about how to deposit content on their web site.

### Books Digitized In-House

Originally, the HathiTrust/Internet Archive package ingest only worked for volumes that were digitized at an OCA Scanning Center. However, in July 2014, Hathi modified their code so we are now able to deposit books from the Internet Archive packages containing books that we digitized in-house.

These books digitized in-house need to be deposited separately from the books digitized by the Internet Archive. There are special requirements for file names and the DASI sheet has to indicate the ppi for the images. Without this extra information, the ingest should fail. It should fail because there is information missing from the IA package for locally digitized books that is needed from Hathi.

### Identifying New Volumes for Deposit

A record is of the IA book identifiers that have previously been deposited into Hathi is maintained in an ACCESS database. This ACCESS database includes other information and is currently maintained on Betsy’s desktop.

Two strategies can be used to identify volumes that need to be ingested. The simplest is to add the take book identifiers from a newly completed picklist, add them to the ACCESS database and work with them.

A second method is to use Internet Archive’s [*Advanced Search returning JSON, XML, and more*](http://archive.org/advancedsearch.php) to get information for all of our volumes. I use searches such as:

* Contributor:Boston College Libraries
* Collection:Boston\_College\_Library

I select the option to download into CSV format (allow more than 50) and select the following data elements: identifier; identifier ark; contributor; scancenter, details.

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

### Getting all the metadata I need from IA

I put the identifiers for the volumes I want to add to Hathi in a text file and run get-ia-metadata.pl to get the xml metadata for each volume the Internet archive.

I put a <records> wrapper around the resulting file and remove all those extra xml declarations.

Then I run the xslt hathi-meta.xml to extract a pipe delimited file of the metadata I need for the next step of the process.

<xsl:value-of select="call\_number"/><xsl:text> |</xsl:text>  
 <xsl:value-of select="identifier"/><xsl:text>|</xsl:text>  
 <xsl:value-of select="identifier-ark"/><xsl:text>|</xsl:text>  
 <xsl:value-of select="identifier-access"/><xsl:text>|</xsl:text>  
 <xsl:value-of select="volume"/><xsl:text>|</xsl:text>  
 <xsl:value-of select="title"/><xsl:text>&#10;</xsl:text>

We need:

Alma

Book id

Ark

Identifier access

Volume

Title

### Next Steps:

I enrich the ACCESS database with this information. When importing text data, it is important to select the advanced feature and indicate that you are importing UTF-8 otherwise diacritics get deleted.

ID, DASIFileName, status, hathiReportStatus, hathiReportTime, note, identifier, aleph, ark, link, volume, title