## Uploading Transcriptions to CONTENTdm from DIYHistory

Scripts at:

* s-lib017.lib.uiowa.edu/local/vol00/diyharvest (harvesting scripts)
* s-lib012.lib.uiowa.ed/local/vol00/Content6/Website/public\_html/catcher (CDM upload scripts)

Transcriptions are uploaded to CONTENTdm monthly through this process:

1. **HARVEST--**Transcriptions (from any page modified in the past month) are automatically harvested from MySQL on the 25th via the “getTranscriptions.py” script and are checked against current CONTENTdm data with the “preCatcherDataChecks.py” script. Two csv files are created--“CatcherUpload\_OmekaTranscriptions[timestamp].csv” and “updates\_[timestamp].csv”--and notifications are emailed to Ellen and Shawn.
   * **“CatcherUpload\_OmekaTranscriptions[timestamp].csv”** is the file that will eventually get uploaded to CONTENTdm via Catcher
   * **“updates\_[timestamp].csv”** is a file of conflicting data that we need to spotcheck and manually clean up before we start the CONTENTdm upload.
2. **SPOTCHECK and MANUAL CLEANUP**—In between the time when pages were originally uploaded to DIYHistory and right now (as you are getting ready to upload transcriptions to CONTENTdm), the CONTENTdm items may have changed (possibly because of pages deleted and reuploaded, or items deleted and re-uploaded altogether—this always issues the item and pages new CONTENTdm ids). We want to make sure that we are uploading the transcriptions to their corresponding pages in CONTENTdm, but if we can’t rely on CONTENTdm identifiers remaining stable, the only way to connect them is by page label.   
     
   “*updates\_[timestamp].cs*v” is a list of pages that did not match exactly. Most of these non-matches are okay to upload, as page labels are sometimes edited after upload to DIYHistory. For some of these non-matches, the script will have already found the current page identifier and has included it in the csv file for you to update in DIYHistory. To perform manual cleanup, follow these steps:
   * **Spot check non-matches—**in the updates\_[timestamp].csv file, compare the “cdm\_label” column with the “omeka\_label” column. Most of these should be similar enough to be sure the pages’ CDM identifiers are still current. Delete these ‘close enough’ matches from this spreadsheet. If the labels do not appear to be describing the same page, flag these pages for the next step.
   * **Find current page identifiers for non-matches—**Omeka and CONTENTdm urls (taken from the Omeka database) are provided in two columns so you can easily check to see if Omeka page matches with its corresponding CDM page. If it matches, you can delete this row. If it doesn’t match, you’ll have to do some detective work to find the new CDM identifier and url. Once you’ve found it, enter the new CDM page identifier in the blank “Identifier” column (format this identifier as [alias\_itemid\_pageid], ex. “cwd\_6687\_1689”) and enter the new reference url under “Source”. If there is already id and url in the Identifier and Source fields, this means the script already found the lost page! (You should check the urls anyway to make sure the script did its job right.) If the parent id has also changed, you’ll need to update this in DIYHistory (in the admin interface find the item and update both the Identifier field and the url in the Source field). Once you’ve gotten through the list, delete any pages that need updating from the “CatcherUpload\_OmekaTranscriptions\_[timestamp].csv” file. We’ll update the metadata for these pages in DIY and their transcriptions will get harvested and uploaded to CDM next month. (Note that anything the script already corrected for you will have already been deleted from “CatcherUpload\_OmekaTranscriptions.csv” for you.)
   * **Update page identifiers in DIYHistory—**Next we need to update the identifers and urls in Omeka for the pages that have changed in CDM. In Omeka admin, upload the “updates.csv” file with CSVImport as you would for file metadata. You only need to map three of the columns:
     1. Original\_filename -> Filename? Box
     2. Identifier -> Identifier
     3. Source -> Source
   * **Make sure the CDM\_field is populated in every cell in the upload file—**Check that every cell is populated in “CatcherUpload\_OmekaTranscriptions\_[timestamp].csv”. This field tells Catcher which metadata field to overwrite. If there is nothing in this field, it is probably because a new collection was added to DIYHistory and the CDMnickKey.csv file wasn’t updated. Find the nickname for the full text field (or the field you want to put the transcription in) in the AllCollectionFields csv file (currently in the Catcher folder on CDM server), and add this nickname plus the collection alias to CDMnickKey.csv.
   * **[optional] Unlock locked records in CONTENTDM**—Check the collections listed in the “CatcherUpload\_OmekaTranscriptions\_[timestamp].csv” file to see if any contain locked records. Metadata in CONTENTdm cannot be updated if the record is locked. Unlock what you can before the next step to save yourself from reuploading failed uploads later.
3. **UPLOAD TO CONTENTDM VIA CATCHER—**To avoid disruption to the CDM server, the Catcher script runs every evening (if there is a file available to upload). Put the “CatcherUpload\_OmekaTranscriptions\_[timestamp].csv” file on the CDM server in the Catcher directory. Later in the evening, the upload script will run, upload the transcriptions, then email Ellen and Shawn a transaction report. (Talk to Brian if you’d like it emailed to others.) This report and the original upload file are moved to Catcher/Completed folder after the upload is finished. Check the report to see if all transcriptions were uploaded successfully:
   * A “Transaction ID:” indicates that the transcription was uploaded successfully
   * “Error detail: This item is currently locked” means that the item is locked in CONTENTdm and has not been uploaded. You will need to unlock this item and reupload the transcriptions from this item.
   * “No content—not uploaded” indicates that there was no transcription in the transcription field to upload. Check the upload file to make sure there was a transcription to be uploaded.
   * “Malformed…” indicates an error with the upload. This could be from a network interruption during the upload. The data itself is usually fine—just try reuploading the next day.
4. **UNLOCK and RE-UPLOAD LOCKED ITEMS—**If there were errors because of locked items or malformed data, unlock these items in CDM administration (if locked), re-index the collection, and re-upload **just these pages**. (You will need to create a new upload file with just the pages that weren’t uploaded.) Follow the instructions for step 3 until all pages are uploaded.
5. **RE-INDEX the COLLECTIONS—**Index each of the collections that are listed in upload file. Once collections are indexed, the transcriptions will be searchable in CDM.