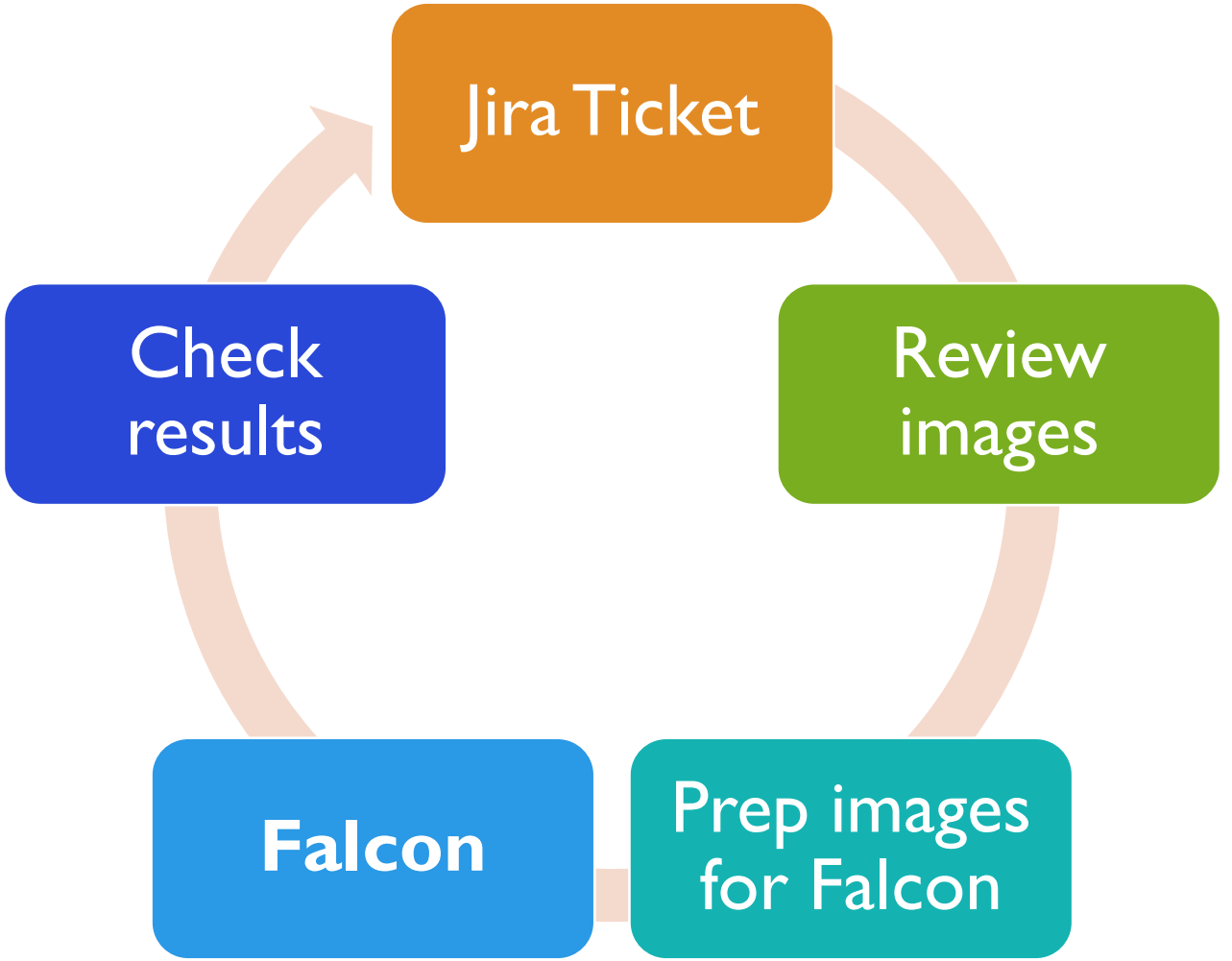
IMAGE PREP WORKFLOW FOR HATHITRUST SUBMISSION

Ying Hu

Digital Initiatives, MU Libraries

September 2024

ROAD MAP


Jira Ticket
Review images
Prep images for Falcon
Falcon
Check results

START WITH JIRA TICKET

- Find image location
- Review any notes and comments on the ticket
- Create a process tracking table if the ticket doesn't have one

HathiTrust batchds610

Edit | Add comment | Assign | More | Awaiting 3rd party res...

Details

Type:	Task	Resolution:	Unresolved
Priority:	Normal		
Component/s:	HathiTrust		
Labels:	None		

Description

Items for this batch:"S:\HT-1-Gathering\TimeAndLabor-GiftItems-DS-610".

!!!NOTES: We digitized copies that Marie purchased on ebay. These items were not added to MU collection. These are great materials to add to HT & for Marie's Prices and Wages project. Talked to Marie 07/25/2024. She will donate these books to the library and give them to Seth for cataloging.

Item	Review, Code, & yaml file	Falcon	Metadata xml	Submit metadata	Submit images
Automobile1924					
HousePrices1978-FederalGovDoc					
Statia1881					
WardwayHomesMagazine1924					

REVIEW IMAGES

1. Check images in both Maters and Access folders
   - Note if there is any differences between the two folders (total number of files, file naming convention, page sequence, dpi, image size…etc.)
     - if so, figure out if anything needs to be fixed
2. We will use images in Access folder for HT submission, and they should pass these quality check points:
   - Every image should be a .tif file
   - Every image should have 600 dpi
   - Pages in correct sequence
   - No curvature and distortion
   - Foldouts in full view (not folded)

PREP IMAGES FOR FALCON

1. Create a set of files for Falcon process

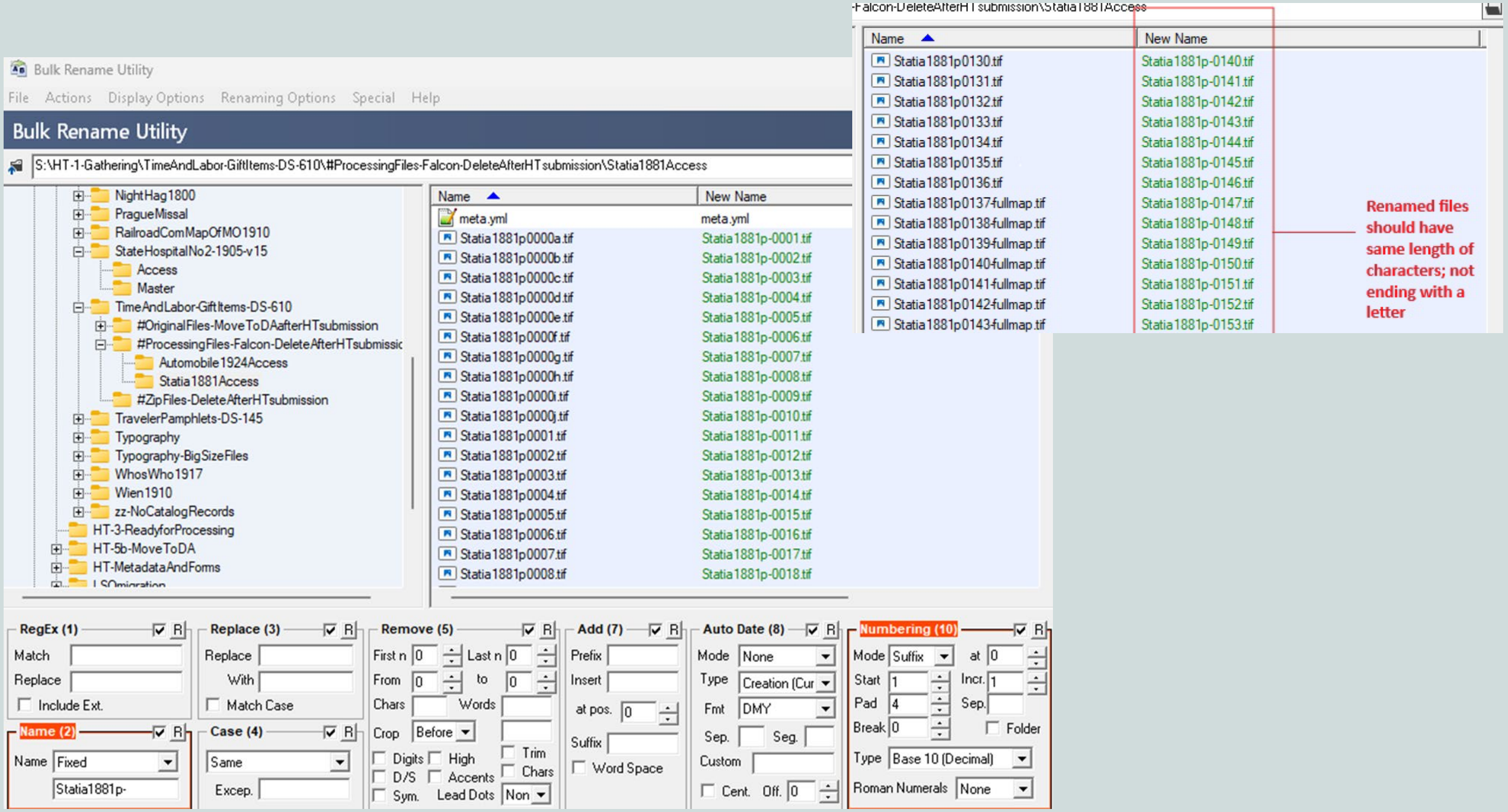
   e.g., Copy access folder and paste it into a new folder "#ProcessingFiles-Falcon-DeleteAfterHTsubmission"

2. Add and edit meta.yml file into the copied access folder

   Change the capture/compression dates, and scanner used for the images

3. Rename files. In this step, we will ignore the local file naming convention and best practices.

   Can keep prefix of the files, but **numbering should start from 0001 and contains numbers only (no additional letters appended in the end)**. This is to prevent page sequence from changing in the next steps.


Bulk Rename Utility
File Actions Display Options Renaming Options Special Help
Bulk Rename Utility
S:\HT-1-Gathering\TimeAndLabor-GiftItems-DS-610\#ProcessingFiles-Falcon-DeleteAfterHTsubmission\Statia1881Access
NightHag1800
PragueMissal
RailroadComMapOfMO1910
StateHospitalNo2-1905-v15
Access
Master
TimeAndLabor-GiftItems-DS-610
#OriginalFiles-MoveToDAafterHTsubmission
#ProcessingFiles-Falcon-DeleteAfterHTsubmissic
Automobile1924Access
Statia1881Access
#ZipFiles-DeleteAfterHTsubmission
TravelerPamphlets-DS-145
Typography
Typography-BigSizeFiles
WhosWho1917
Wien1910
zz-NoCatalogRecords
HT-3-ReadyforProcessing
HT-5b-MoveToDA
HT-MetadataAndForms
Name
New Name
meta.yml
Statia1881p0000a.tif
Statia1881p0000b.tif
Statia1881p0000c.tif
Statia1881p0000d.tif
Statia1881p0000e.tif
Statia1881p0000f.tif
Statia1881p0000g.tif
Statia1881p0000h.tif
Statia1881p0000i.tif
Statia1881p0000j.tif
Statia1881p0001.tif
Statia1881p0002.tif
Statia1881p0003.tif
Statia1881p0004.tif
Statia1881p0005.tif
Statia1881p0006.tif
Statia1881p0007.tif
Statia1881p0008.tif
Statia1881p-0001.tif
Statia1881p-0002.tif
Statia1881p-0003.tif
Statia1881p-0004.tif
Statia1881p-0005.tif
Statia1881p-0006.tif
Statia1881p-0007.tif
Statia1881p-0008.tif
Statia1881p-0009.tif
Statia1881p-0010.tif
Statia1881p-0011.tif
Statia1881p-0012.tif
Statia1881p-0013.tif
Statia1881p-0014.tif
Statia1881p-0015.tif
Statia1881p-0016.tif
Statia1881p-0017.tif
Statia1881p-0018.tif
RegEx (1)
Match
Replace
Include Ext.
Replace (3)
Replace
With
Match Case
Remove (5)
First n
Last n
From
to
Chars
Words
Crop
Before
Digits
High
D/S
Accents
Sym.
Lead Dots
Non
Trim
Chars
Add (7)
Prefix
Insert
at pos.
Suffix
Word Space
Auto Date (8)
Mode
None
Type
Creation (Cur
Fmt
DMY
Sep.
Seg.
Custom
Cent.
Off.
Numbering (10)
Mode
Suffix
at
Start
Incr.
Pad
Sep.
Break
Folder
Type
Base 10 (Decimal)
Roman Numerals
None
Name (2)
Name
Fixed
Statia1881p-
Case (4)
Same
Excep.
-Falcon-DeleteAfterHTsubmission\Statia1881Access
Name
New Name
Statia1881p0130.tif
Statia1881p0131.tif
Statia1881p0132.tif
Statia1881p0133.tif
Statia1881p0134.tif
Statia1881p0135.tif
Statia1881p0136.tif
Statia1881p0137-fullmap.tif
Statia1881p0138-fullmap.tif
Statia1881p0139-fullmap.tif
Statia1881p0140-fullmap.tif
Statia1881p0141-fullmap.tif
Statia1881p0142-fullmap.tif
Statia1881p0143-fullmap.tif
Statia1881p-0140.tif
Statia1881p-0141.tif
Statia1881p-0142.tif
Statia1881p-0143.tif
Statia1881p-0144.tif
Statia1881p-0145.tif
Statia1881p-0146.tif
Statia1881p-0147.tif
Statia1881p-0148.tif
Statia1881p-0149.tif
Statia1881p-0150.tif
Statia1881p-0151.tif
Statia1881p-0152.tif
Statia1881p-0153.tif
Renamed files should have same length of characters; not ending with a letter

PREP IMAGES FOR FALCON CONTINUED

4. Code the files. Add a suffix letter according to the end results we desire for each page. There are 4 letters we use; their meanings are as following:
   - b = bitonal & OCR
   - B = bitonal; No OCR
   - z = color/grayscale & OCR
   - Z = color/grayscale; No OCR
5. Check the image folder again before starting Falcon

S:\HT-1-Gathering\TimeAndLa

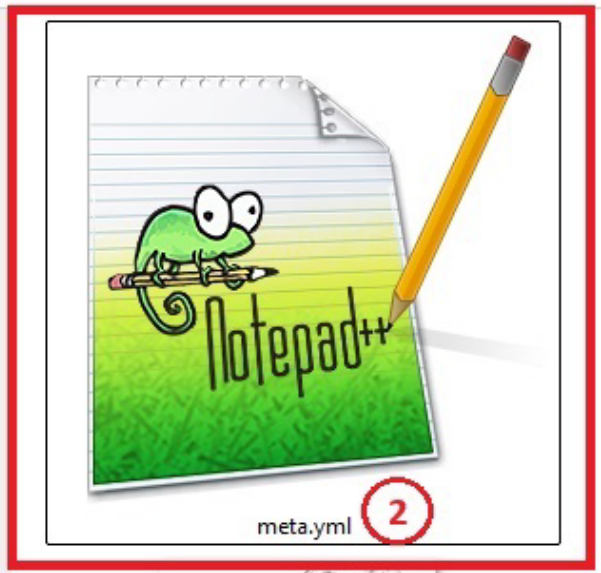
#ProcessingFiles-Falcon-DeleteAfterHTsubmission > Statia1881Access

Search Statia1881Access

New | Sort | View | Details

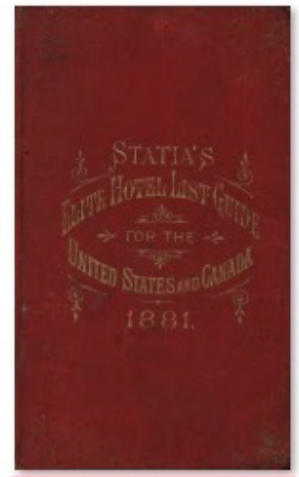
- TimeAndLabor-GiftItems-DS-610
  - #OriginalFiles-MoveToDAafterHTsubmission
  - #ProcessingFiles-Falcon-DeleteAfterHTsubmission
    - Automobile1924Access
    - Statia1881Access
    - #ZipFiles-DeleteAfterHTsubmission
  - TravelerPamphlets-DS-145
  - Typography
  - Typography-BigSizeFiles
  - WhosWho1917
  - Wien1910
  - zz-NoCatalogRecords
- HT-3-ReadyforProcessing
- HT-5b-MoveToDA
- HT-MetadataAndForms
- LSOmigration
- LSOmigration-overflow
- MOspaceOnlineSubmissions
- Projects

1 separate copy of files for falcon, not directly on original access files



meta.yml 2

.yml file is needed for each folder



Statia1881p-0001z.tif

3 Renamed and coded

Statia1881p-0002B.tif

Statia1881p-0003b.tif

Statia1881p-0004B.tif

Statia1881p-0005B.tif

159 items

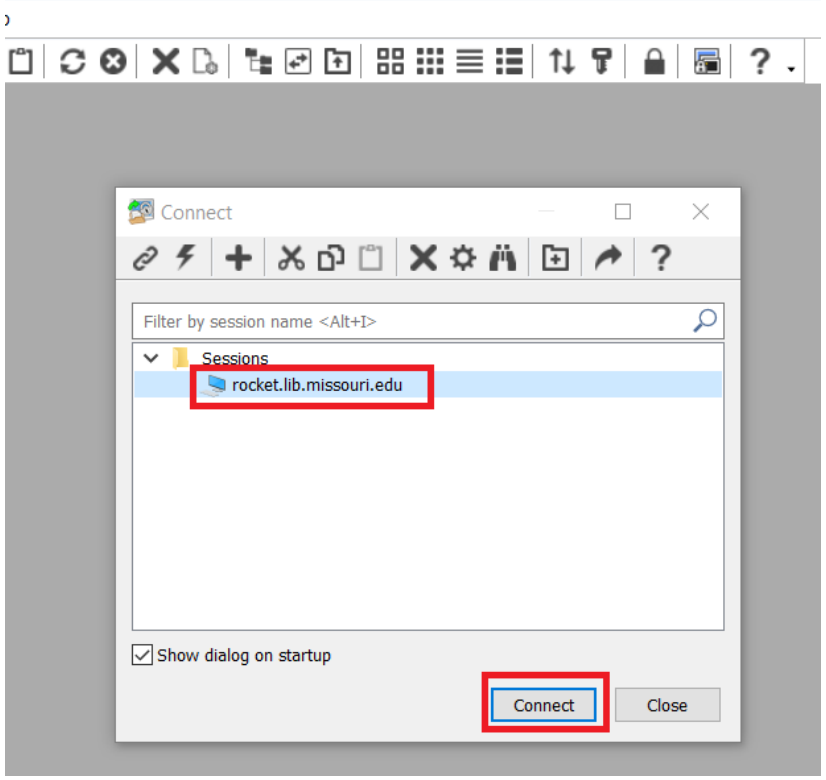
FALCON PROCESS

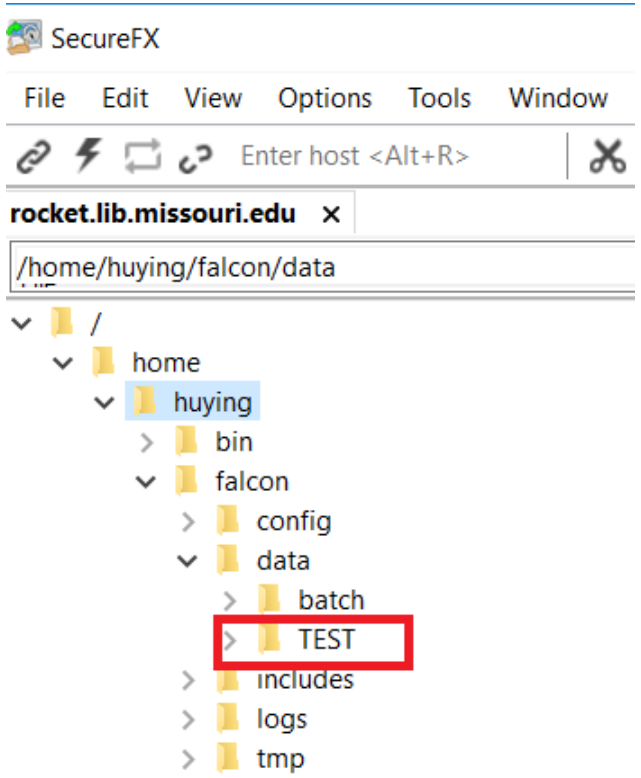
Note: You will need a software for FTP transfer and a terminal to execute a command line. We currently use SecureFx (for FTP transfer) and Secure CRT for running command lines.

1. Open SecureFx software and connect to **rocket.lib.missouri.edu**
2. Upload the image folder to falcon/data or falcon/data/batch depends on how you want to process the images
3. For individual folder processing, when uploading is complete, open Secure CRT, connect to rocket.missouri.edu. Run a commend line “falcon foldername”. Batch processing will start at 5pm each day and run automatically.

SecureFx:

Upload images





Secure CRT:

Run script

```
[huying@rocket ~]$ falcon TEST
+++ [Fri Sep 27 12:51:24 CDT 2019] [INFO] Start falcon on TEST
+++ [Fri Sep 27 12:51:24 CDT 2019] [INFO]   Create /home/huying/falcon/data/2019-09-27/TEST
+++ [Fri Sep 27 12:51:24 CDT 2019] [INFO]   Decompress images...
+++ [Fri Sep 27 12:51:26 CDT 2019] [INFO]   Convert to grayscale: 4/5
+++ [Fri Sep 27 12:51:27 CDT 2019] [INFO]   OCR images: 4/5
+++ [Fri Sep 27 12:52:07 CDT 2019] [INFO]   No OCR: 1/5
+++ [Fri Sep 27 12:52:07 CDT 2019] [INFO]   JP2 conversion: 5/5
+++ [Fri Sep 27 12:52:08 CDT 2019] [INFO]   Copy yaml file
+++ [Fri Sep 27 12:52:08 CDT 2019] [INFO]   Start checksum...
+++ [Fri Sep 27 12:52:08 CDT 2019] [INFO]   Start zip...
+++ [Fri Sep 27 12:52:08 CDT 2019] [INFO] End falcon on TEST
[huying@rocket ~]$
```

CHECK RESULTS AND DELETE FILES FROM FALCON

- When Falcon process is finished, a .zip file will be created and shown in the Falcon/data folder.
- Download the zip file and spot check a few files to ensure the process was done correctly and no file was corrupted.
- After downloading the zip files, delete the original folder and the folders created by Falcon. **This is very important** for folders in the batch folder, as the script will re-process them at 5:00 pm **each day**.

MU Libraries Digital Services / DS-610

HathiTrust batchds610

Edit | Add comment | Assign | More | In Progress

Details

Type: Task
Priority: Normal
Component/s: HathiTrust
Labels: None

Resolution: Unresolved

DON'T FORGET TO UPDATE JIRA TICKET

Description

Items for this batch:"S:\HT-1-Gathering\TimeAndLabor-GiftItems-DS-610".

!!!NOTES: We digitized copies that Marie purchased on ebay. These items were not added to MU collection. These are great materials to add to HT & for Marie's Prices and Wages project. Talked to Marie 07/25/2024. She will donate these books to the library and give them to Seth for cataloging.

Item	Review, Code, & yaml file	Falcon	Metadata xml	Submit metadata	Submit images
Automobile1924	x				
HousePrices1978-FederalGovDoc					
Statia1881	x				
WardwayHomesMagazine1924					