Homewood Photography Project Documentation

Elizabeth England
National Digital Stewardship Resident, 2016-2017
University Archives, Sheridan Libraries
Johns Hopkins University

Table of Contents

1. Introduction

As the official photographers for Johns Hopkins University, Homewood Photography creates images that provide an unparalleled visual record of the university's history. The Archives has previously accessioned and made available Homewood Photography's analog work, which has proven to be a rich source of historical images for the university community as well as the general public.

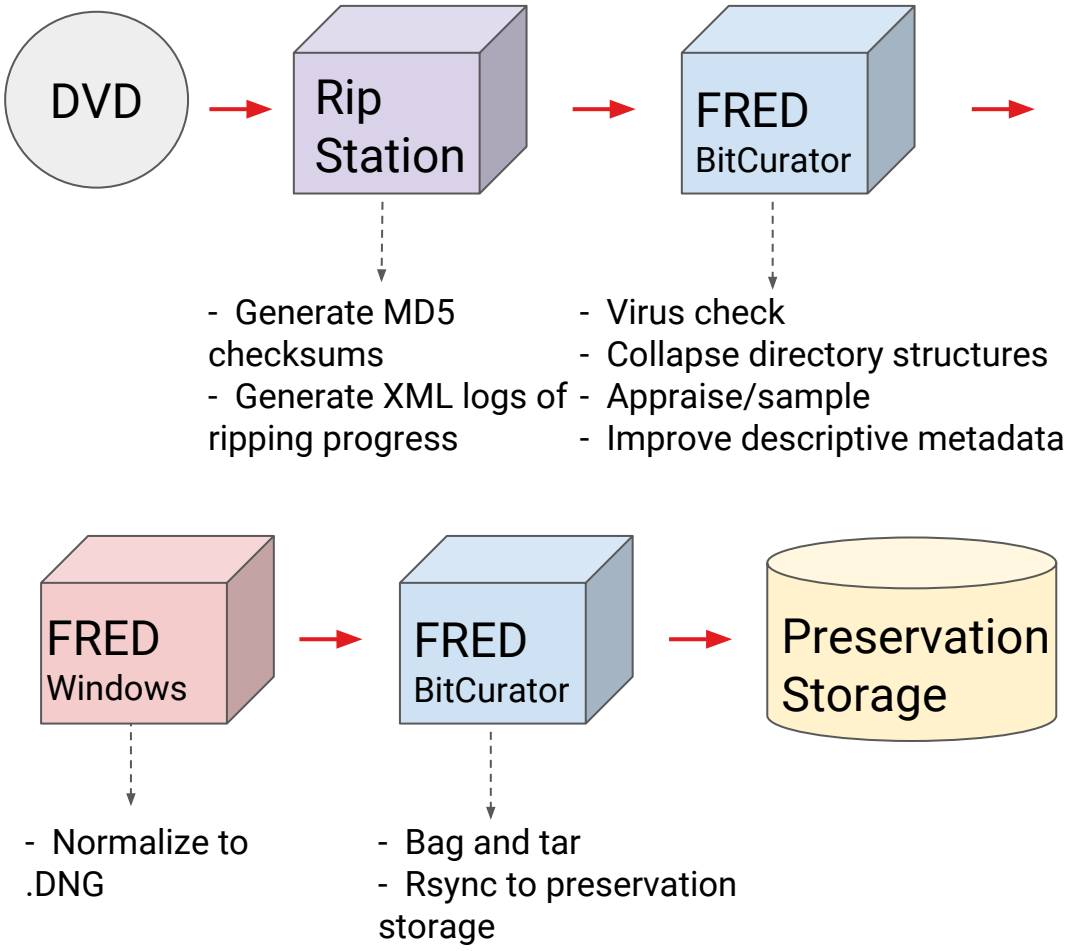
Since the mid-2000s, Homewood Photography has been using an exclusively digital workflow. Many of these digital images are stored on optical media which is beginning to fail. While Homewood Photography recognizes the historical importance of these images, they are unable to manage a strategy for long-term preservation and access. Metadata about these images is currently maintained in an Access database. However, these descriptions are inconsistently applied and sometimes incorrect.

Homewood Photography continues to produce digital images at a projected rate of 3TB per year, of which approximately one-half is appropriate for accessioning into the Archives based on how it aligns with the Archives' collection development policy. In order to ensure preservation of these images while preventing the buildup of another backlog, a process needed to be developed for the routine transfer of these images to the Archives.

Johns Hopkins University was selected as a host institution for the National Digital Stewardship Residency (NDSR), funded by the Institute for Museum and Library Services (IMLS) and administered by the Library of Congress. From October 2016 through September 2017, the Archives hosted a resident, Elizabeth England, to work on the capture, management, and preservation of Homewood Photography's born-digital content.

This descriptive workflow collects the documents that were produced for transferring, appraising, processing, preserving, and describing born-digital photographs from Homewood Photography during the course of the residency.

2. Workflow


DVD
Rip Station
FRED BitCurator
- Generate MD5 checksums
- Generate XML logs of ripping progress
- Virus check
- Collapse directory structures
- Appraise/sample
- Improve descriptive metadata
FRED Windows
FRED BitCurator
Preservation Storage
- Normalize to .DNG
- Bag and tar
- Rsync to preservation storage

2.1 Equipment

Hardware:

- Networked workstation - used for tracking progress, description of content
- RipStation - used for ripping DVDs
- Forensic Recovery of Evidence Device (FRED) - used for processing content
    - BitCurator partition
    - Windows partition
    - 8 TB RAID

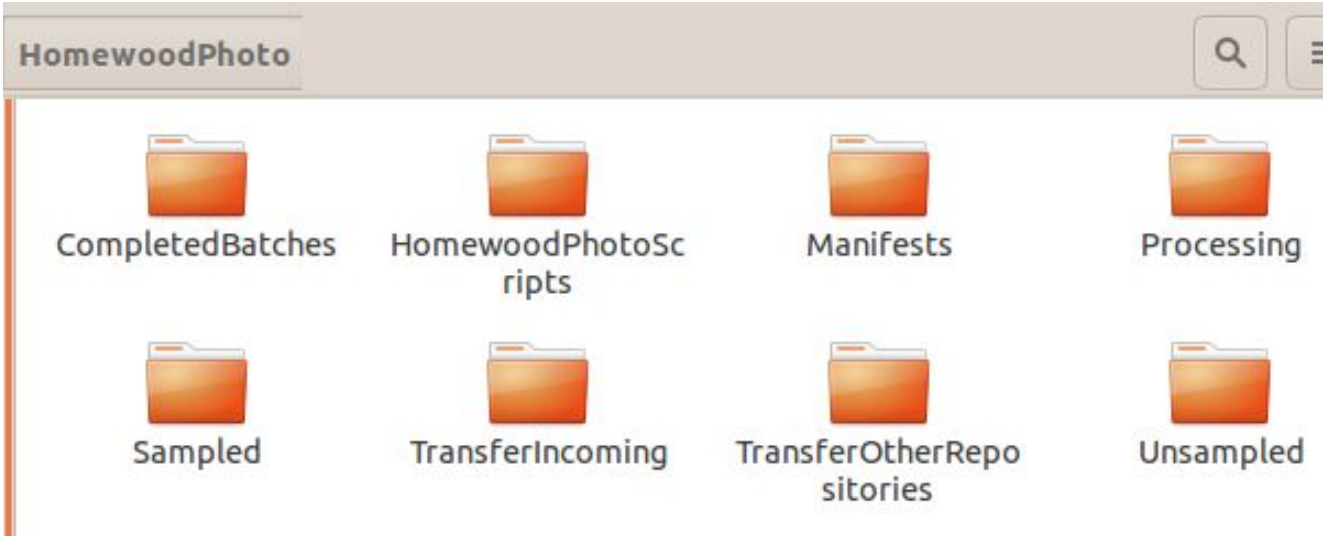
    See the Archives & Manuscripts wiki for hardware troubleshooting assistance.

Software/Tools:

- RipStation
    - RipStation DataGrabber - capture
    - Robocopy or Grsync - optional, transfer files to FRED
- FRED - BitCurator
    - ClamTk - virus checker
    - BagIt - package content for transfer to SAM
    - Rsync - transfer files to SAM for long-term preservation storage
    - Shotwell - viewer for camera raw thumbnails
    - PyExifToolGUI - optional, view embedded metadata
    - FITS - optional, detailed file format information
- FRED - Windows
    - Adobe DNG Converter - normalize files
- Workstation
    - OpenRefine - description clean-up
    - Excel - track progress; cleaned-up version of Homewood Photography database for reference. Can edit `Database_Copy.xls` as needed for creating description
    - Access - original Homewood Photography database
- GitHub - clone/download from here
    - Transfer progress template
    - Python scripts
    - ArchivesSpace templates
    - ArchivesSpace scripts

2.2 Processing Space and Documents

Processing takes place while the content is stored on the 8 TB RAID on the FRED.
A number of directories have been created on the RAID to help organize the content. Each directory will be discussed more in depth as it arises in the workflow.



A few versions of the Homewood Photography database live in the accession folder, view the `read me` file there for more information on each. `Database_Copy.xls` is a scratchpad for editing/enhancing description, and can assist in categorizing content.

https://github.com/jhu-archives-and-manuscripts/homewood-photo

Create a `TransferProgress` spreadsheet based on the template in GitHub for logging transfer progress and preservation events. The preservation events will ultimately be added to the accession record in ArchivesSpace.

2.3 Ripping Discs

The RipStation can hold a maximum of 300 discs, 150 on each side. Generally, a batch of 300 will take 12-15 hours. Load them in numerical disc order (by Homewood Photo's numbering system) with the lower disc numbers at the top. Consult the RipStation manual for additional assistance, located in the Electronic Records folder on the GDrive or desktop of the RipStation.

The workflow relies on ripping from the RipStation directly to the RAID on the FRED.[1] In order to designate this root path, boot up the FRED (in BitCurator/Ubuntu) first, followed by the RipStation. The connection between the RAID and RipStation may need to be re-established each time they are rebooted. On the RipStation, reconnect the RAID by navigating to the Computer, and the RAID should be listed as a Network Connection. Enter the password for BitCurator to establish the network connection.

On the FRED, create a new directory within `TransferIncoming` where the ripped discs will be stored, using the date you are beginning the batch:

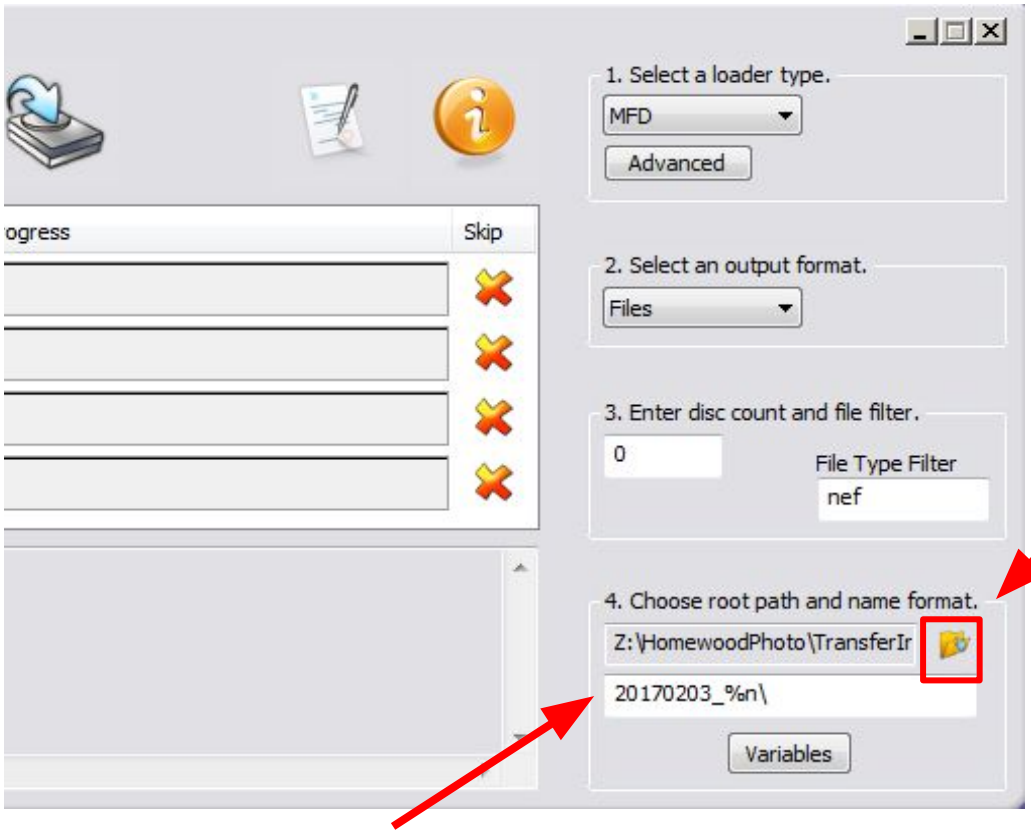
```
[…]/HomewoodPhoto/TransferIncoming/[YYYYMMDD]
```

[1] Alternatively, you can rip to local storage on the RipStation and then transfer using Robocopy or Grsync to transfer to the FRED. The RipStation has 1 TB internal capacity, and one full batch of 300 discs could fill this storage.

Next, on the RipStation, open the “Ripstation DataGrabber” program on the desktop

The settings will default to the last used settings, but should be checked each time and updated for the new batch. On the main screen, the settings should be:

1. Loader type: `MFD`
2. Output format: `files`
3. Disc count: `0` (for continuous ripping)
   File type filter [2]: `nef`
4. Root path: `[set to the directory you just created]`(folder button)
   In the text box, set the name format: `[YYYYMMDD_%n]` (date_incrementalCounter)

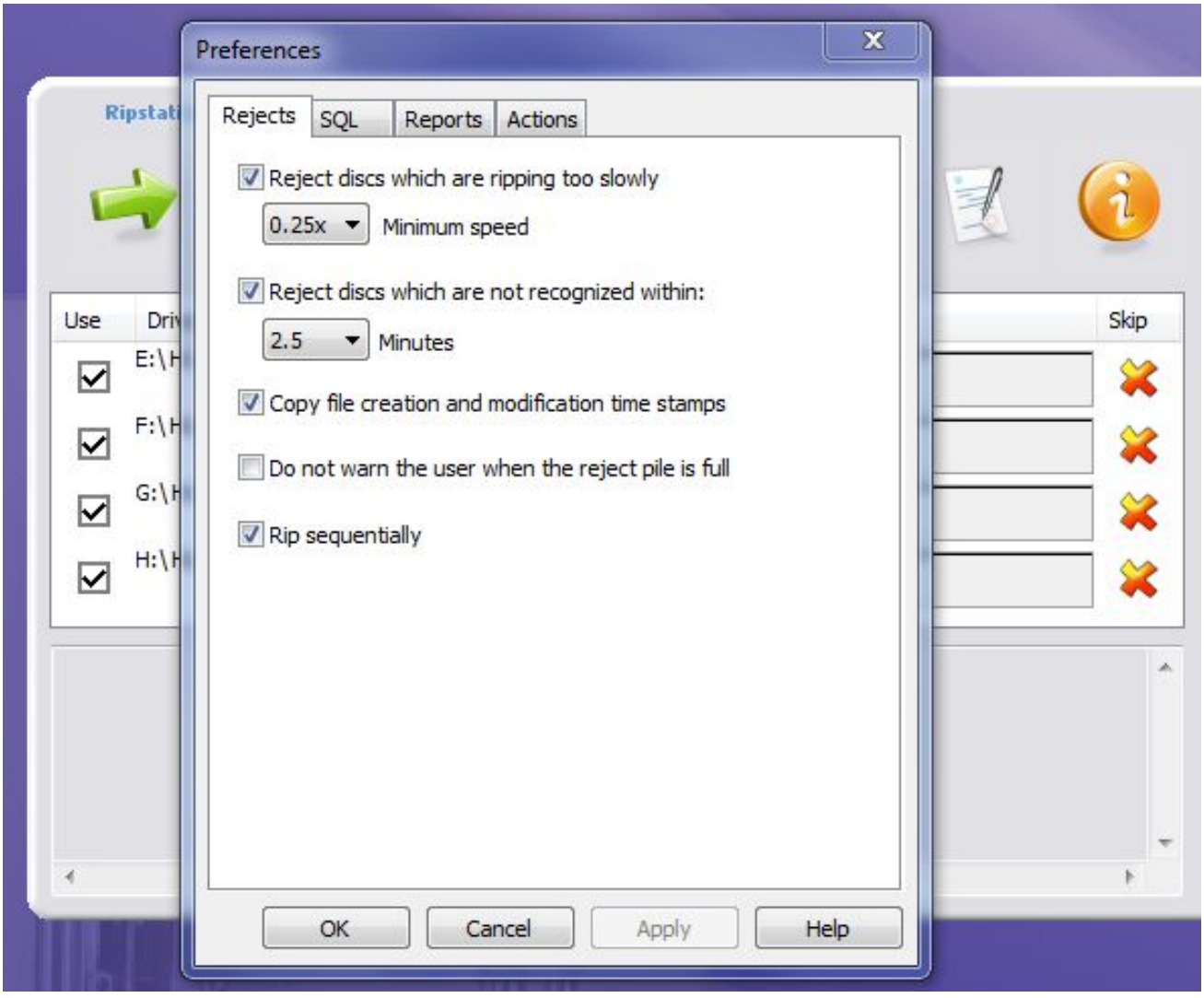


[2] The RipStation will still preserve empty directories that contained other file formats, such as JPEGs, which will be deleted later. Alternatively, to rip all files regardless of format, leave this field blank.

Next, check the settings under `Preferences -> Rejects`

Select all options except “Do not warn”:

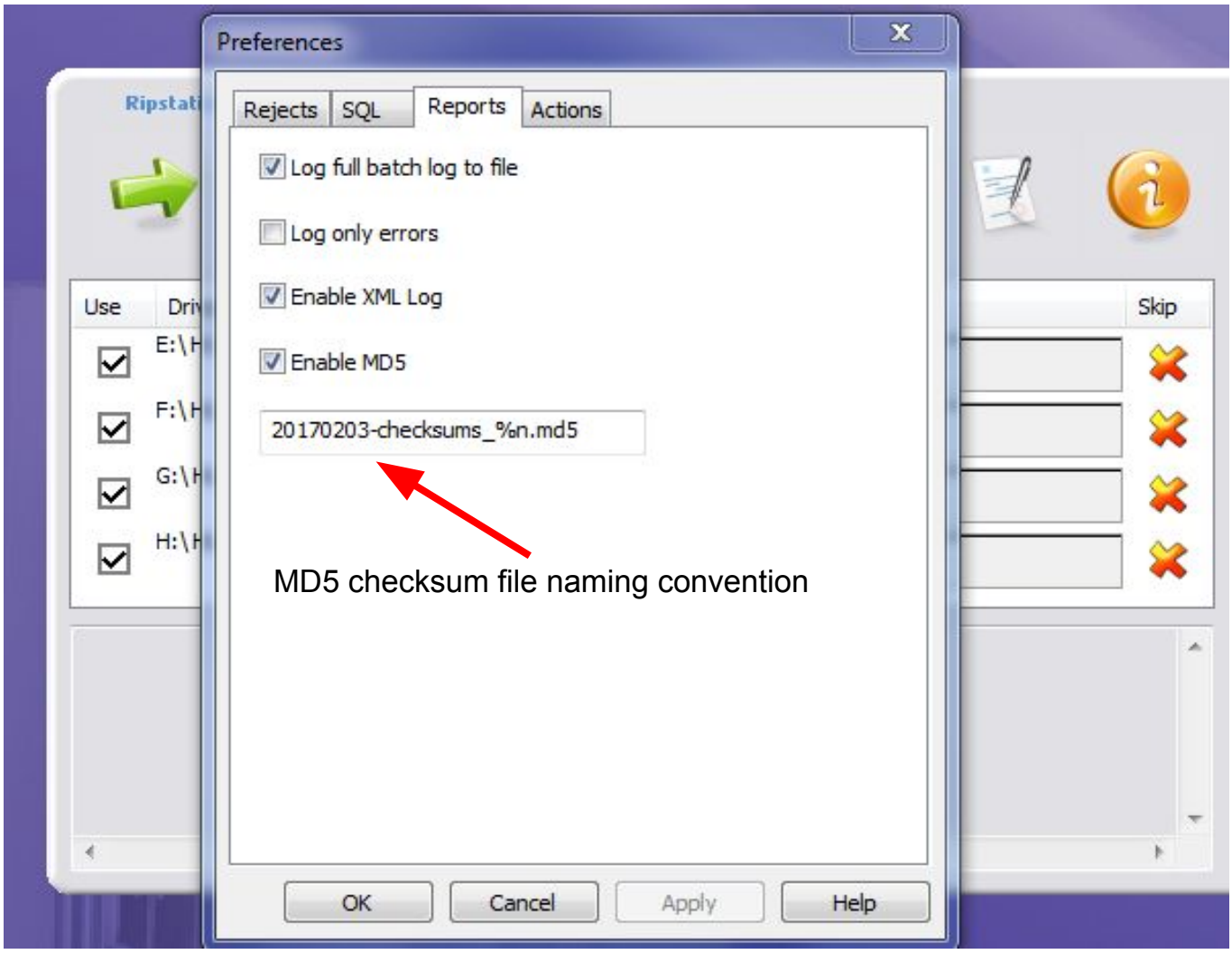
- `Reject discs which are ripping too slowly: 0.25x minimum speed`
- `Reject discs which are not recognized within 2.5 minutes`
- `Copy file creation & modification time stamps`
- `Rip sequentially` [3]



[3] Ripping sequentially is not necessary but is highly recommended. While it slows down the ripping process, it helps in overall organization and makes putting the discs away much easier.

Lastly, check the settings under `Preferences -> Reports` Select three reports, which will later be stored in the accession folder:

- `Log full batch log to file`shows the path to each file on each disc. Will create a file called “Log.txt”
- `Enable XML Log` shows the times discs started, failures, when completed. Will create a file called “Batch_Log.xml”
- `Enable MD5` creates MD5 checksums for each file. In the text box, set the file naming convention: `[YYYYMMDD-checksums_%n.md5]`in which the date is the batch date.[4]



[4] If the content is going to be processed right away, generating the checksums at this stage is unnecessary because the files will be normalized. However, if the .NEF content is going to be captured but not processed right away, checksums should be generated.

The batch is started with the green arrow button.

As the discs are being ripped, any discs that have errors will be sequestered on the short spindle; consult the `Batch_Log.xml` for more information on any errors that occur. It is recommended to start a batch late afternoon, turn off the monitor when leaving, and then check it in the morning.

Document the Batch

When the batch is finished, update the `TransferProgress` spreadsheet.

Record the batch date, disc range (beginning and ending disc numbers), transfer status, and note any errors.

Batch ID (Date Ripped)	Disc Numbers	Transferred	Errors
20170103	100100-100250	X	
20170104	100251-100500	X	
20170106	100501-100800	X	100604

Repeat the above steps as necessary, creating a new directory for each batch in `[…]/HomewoodPhoto/TransferIncoming/[YYYYMMDD]`. Keep an eye on how much free space is on the RAID before beginning new batches, and be cautious with how much you use the FRED for other processing tasks while batches are being ripped.

For each new batch of discs, update:
1) the root path, 2) the name format, 3) MD5 log name

2.4 Processing

Virus Check

On the FRED, run a virus check using ClamTk in BitCurator and add the date of the virus check to the Transfer Progress table. You can virus check multiple batches simultaneously.[5] Record the date of the virus check on the `TransferProgress` sheet.

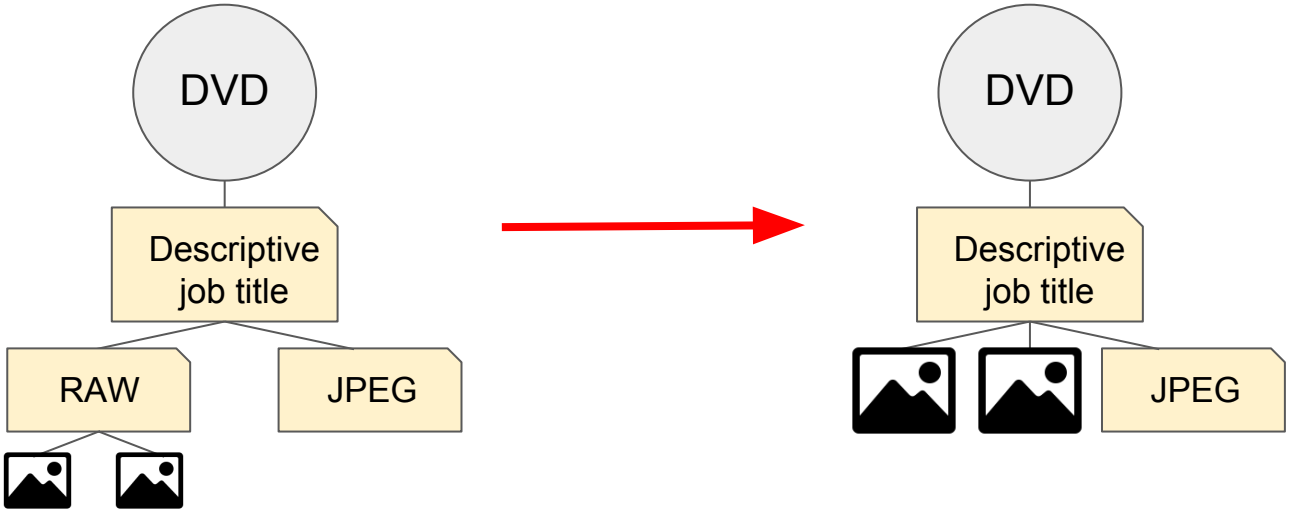
Batch ID (Date Ripped)	Disc Numbers	Transferred	Errors	Virus Check
20170103	100100-100250	X		20170113
20170104	100251-100500	X		20170117
20170106	100501-100800	X	100604	20170123

[5] It isn't ideal to virus check content once it is already on the network, but the FRED can't be switched to offline when accessioning this content, since the ripping process relies on a network connection.

The content needs to be organized so that it can be worked with at the job level. While virus checks can be done simultaneously for multiple batches, the following steps must be done for each batch individually.

Collapse Directories

First, run the python script `collapseDirectories.py` to collapse the directory structures, which will move the individual photos within each job so they are nested directly under the descriptive job titles:



This and other python scripts that will be used later are stored in the File Management submodule in the Homewood Photo GitHub.

```
$ cd [directory where Python script is stored]
$ python collapseDirectories.py
```

type the path (do not drag & drop):

```
[…]/HomewoodPhoto/TransferIncoming/[batchID]
```

This will generate a CSV file called `moveLog[date-time].csv` that lists each file's old location and new location. The log is generated wherever you ran the script from.

Create a new folder with the batch ID in the `CompletedBatches` folder and move the log here. Later, the CompletedBatches folder will be moved to the accession folder for retention.

Delete Empty Directories

```
$ cd [filepath to the batch]
$ find . -type d -empty -delete
```

This will delete all the empty JPEG directories resulting from the ripping process, as well as deleting empty directories resulting from collapsing the directory structures.

Log that directories were collapsed and empty directories were deleted on the `TransferProgress` sheet.

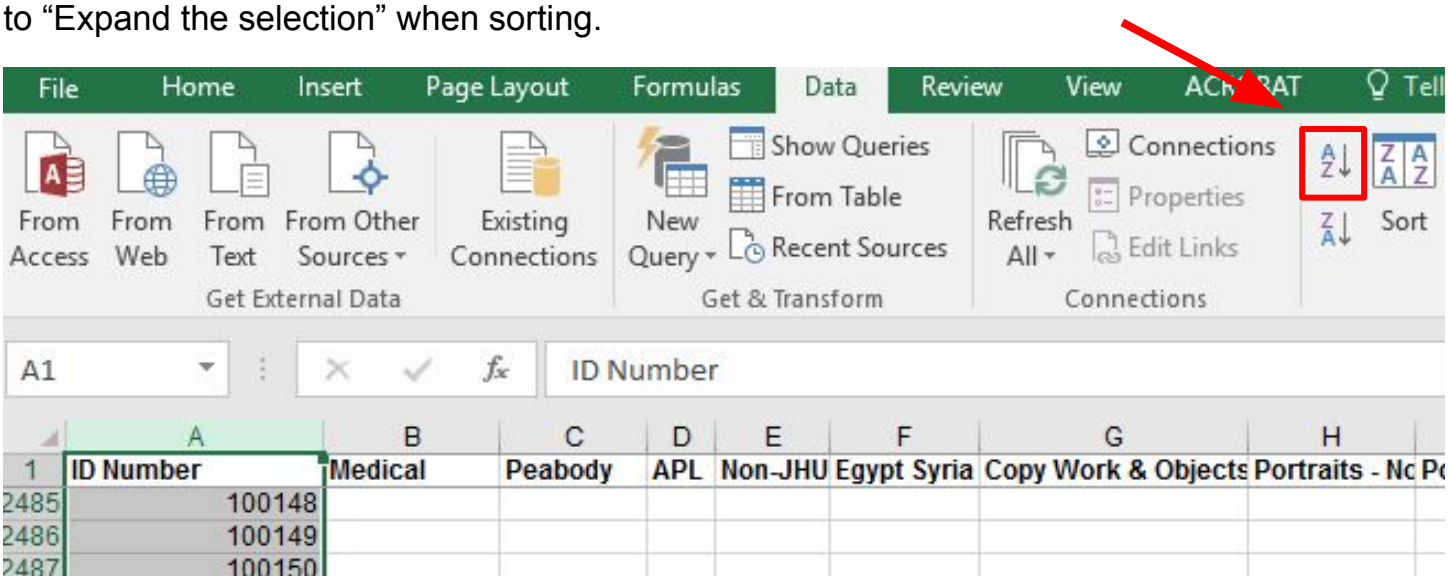
Batch ID (Date Ripped)	Disc Numbers	Transferred	Errors	Virus Check	Collapsed Dirs	Deleted Empty Dirs
20170103	100100-100250	X		20170113	X	X
20170104	100251-100500	X		20170117	X	X
20170106	100501-100800	X	100604	20170123	X	X

Arrangement and Appraisal

Now you will start breaking down the batch to the job level by moving content out of `TransferIncoming`. The majority of content will be moved to the `Processing` directory, but a significant portion will be for other repositories, or will be out of scope and can be deleted. You will want to have four tabs or windows open: 1) the batch within TransferIncoming, 2) Processing, 3) Peabody, and 4) Chesney.[6]

As you are moving large amounts of content, monitor the system performance in the top right corner and allow time for the files to be moved before trying to move more.

Within the batch, work your way methodically through, starting with the content from disc 001. You will have to open the folder for the disc, look at the names of each job, and drag & drop the job to the appropriate place (Processing, Peabody, Chesney, etc.). It helps to have the `Database_Copy` open on the workstation computer, as the descriptions in the database may be different from the folder names. With the database sorted by “ID Number” (disc number), find your starting point based on the first disc number of the batch. When prompted by Excel, choose to “Expand the selection” when sorting.



For more information on what is considered out of scope, see the [Content Categories for Weeding and Transfer].

You can save yourself some descriptive work later by flagging jobs in `Database_Copy.xls` that are going to be retained, and enhancing the database description based on the folder titles, if you choose. Note that the database is not a comprehensive inventory, and some jobs may not be listed.

[6] Peabody refers to the Johns Hopkins Peabody Institute Archives, and Chesney refers to the Alan Mason Chesney Medical Archives of the Johns Hopkins Medical Institutions.

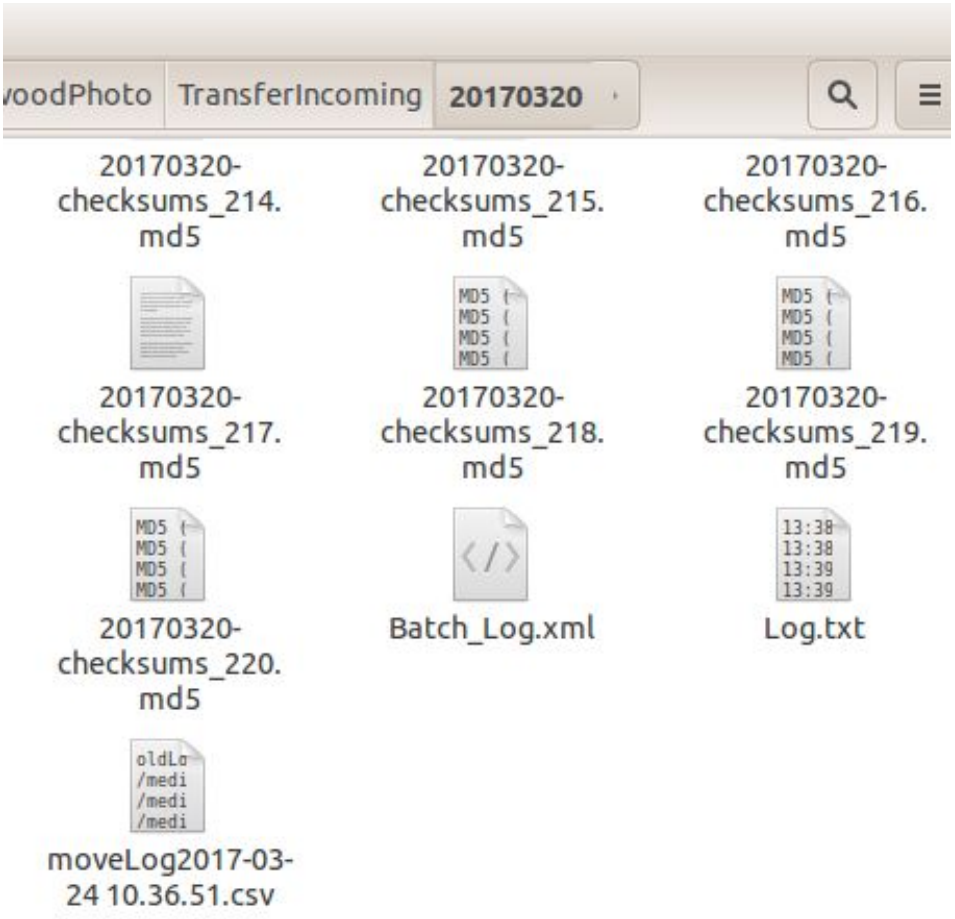
Precaution

When moving jobs from `TransferIncoming` into `Processing`, it is likely you will encounter a message that a folder with that name already exists, and be prompted to replace the folder. In this scenario, ***do not replace*** *the folder as they may in fact contain different files from the same job. Instead, add a "2" to the end of the folder you were attempting to move and try moving it again. In the next processing stage, this content will be checked for possible duplicates.*

When you have finished moving content out of each disc folder within the batch, delete the empty directories **from the command line** to ensure you don't accidentally delete any directory that still has content in it.

```
$ cd [filepath to the batch]
$ find . -type d -empty -delete
```

The batch directory should now only contain the logs from the RipStation, including checksums, and the moveLog.csv generated from collapsing the directories. Move all these files out of `TransferIncoming` and into the appropriate batch directory within `CompletedBatches`.



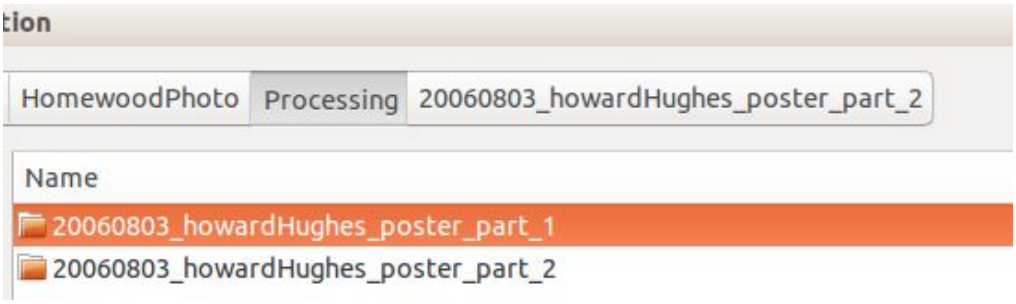
Repeat the steps to collapse directories, delete empty directories, arrange, and appraise batches as needed.

Because the following processing steps can be done at scale, it is most efficient to get as many batches as possible to this point before continuing. Additionally, one job can be split between different DVDs in different batches, and it is best to reunite this content prior to sampling it, so that the sampling is conducted evenly across the files.

Reunite Split Content

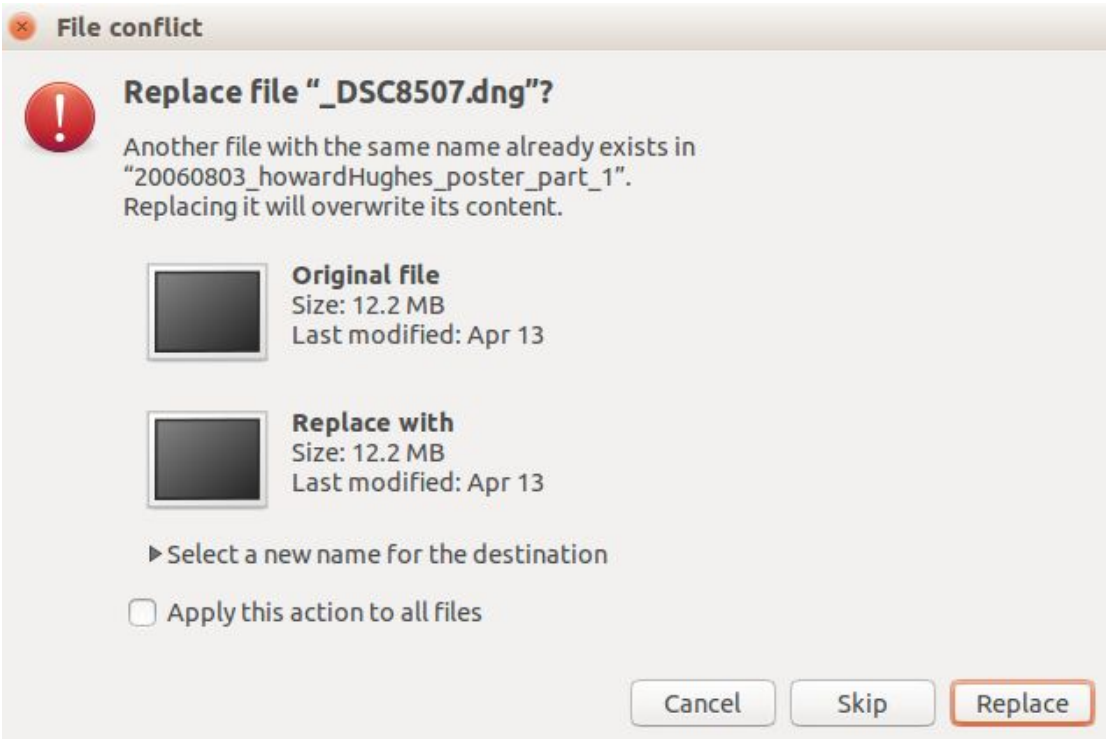
Now that the batches have been broken down and content for other repositories has been moved, you can begin working with the content at the job level to reunite split content.

Within the `Processing` directory, the jobs will now be organized chronologically by shoot date, because each folder title begins with [YYYYMMDD]. To reunite content with the same date and description (such as part 1 and 2 of the same job), open the folder for part 2, select all files, right click, select “Move To” and choose the destination. Again, monitor the system performance to allow time for the files to be moved before attempting to move more.



Although some jobs span multiple dates (such as 20090823_pre_orientation_hiking and 20090824_pre_orientation_hiking), **only reunite the content if it begins with the same date** (such as 20090823_pre_orientation_hiking_part_1 and 20090823_pre_orientation_hiking_part_2).

You may encounter a message that the file names already exist in the destination folder; this is most likely duplicate content that Homewood Photography had saved on multiple DVDs.



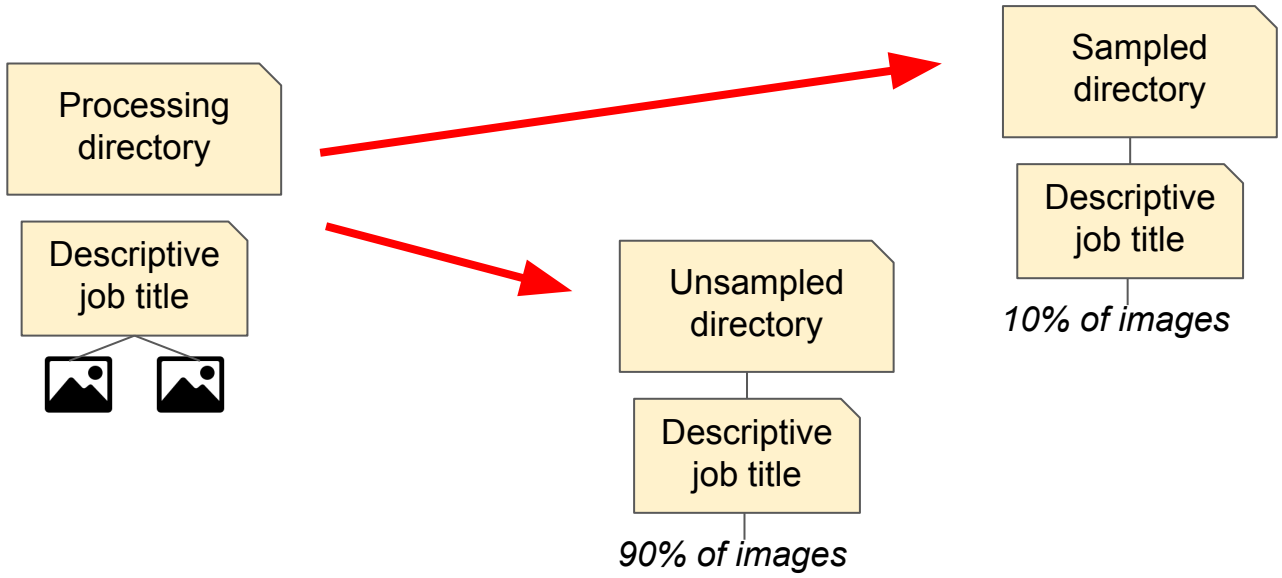
If in doubt, you can visually check a few files against one another (.../part1/_DNG1157 vs .../part2/_DNG1157) before proceeding. Once you have confirmed the content is identical, try moving the files again and choose to replace the files in the destination folder. By overwriting, you will empty out the source directory, whereas if you didn't overwrite, you would be left with files for the same job still in multiple folders.

Once all content has been reunited, delete empty directories again:

```
$ cd [filepath to the batch]
$ find . -type d -empty -delete
```

Sampling

Each in-scope job is sampled down to 10%, due to the volume of content (see Sampling Strategy for more information.) This is accomplished through the python script `samplingScript.py` which will grab every 10th image within each job folder, push that 10% to be retained into the `Sampled` directory, and move the remaining 90% into the `Unsampled` directory.




```
$ cd [directory where Python script is stored]
$ python samplingScript.py
```

type the path (do not drag & drop):

`[…]/HomewoodPhoto/Processing`

The script prints 2 logs to the where ran you the script from, showing what was sampled and what was unsampled. Create a folder within the `CompletedBatches` directory called `SamplingLogs` and move the logs. Do not delete the unsampled content yet, in case you later need to reunite content for transfer to another repository.

Manually further edit down individual and group portraits to 1-3 images per job.

Log that content was sampled on the `TransferProgress` sheet.

E	F	G	H
Virus Check	Collapsed Dirs	Deleted Empty Dirs	Sampled
20170113	X	X	X
20170117	X	X	X
20170123	X	X	X

2.5 Folder Names

Use the python script `extractDirectoryNames.py`to create a CSV called JobsListing.csv of all the folder names. The CSV is placed where the script was run from.

```
$ cd [directory where Python script is stored]
$ python extractDirectoryNames.py
```

type the path (do not drag & drop):

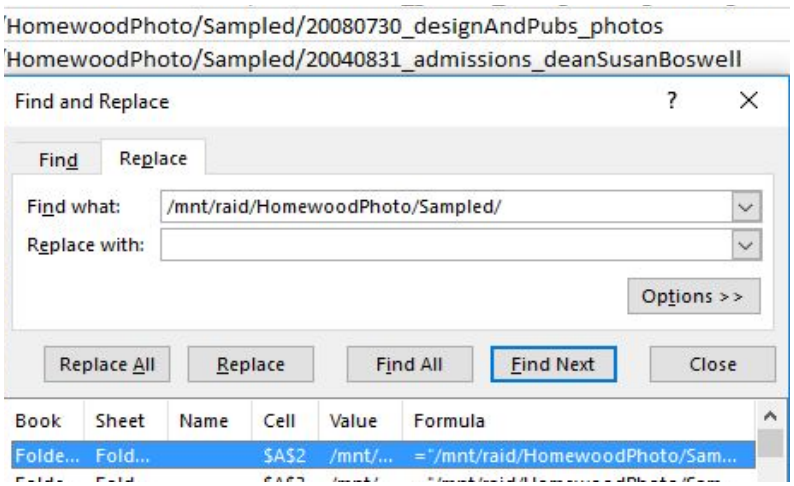
`[…]/HomewoodPhoto/Sampled`

How much, or by what method, to standardize and clean up the folder names is at the processor's discretion, but option 1 or 3 required approval from the University Archivist. For more on the recommended folder naming, see Folder Naming Conventions. Generally, there are three options:

1. **Acceptable**: Leave the folder names as assigned by the creators
2. **Standard**: Use Excel to do minimal clean up: removing "raw," spaces, special characters such as "&"
3. **Advanced**: Use OpenRefine to do the above clean-up and also standardization: removing additional underscores, shortening the length of the folder names, standardizing abbreviations for clients and events. This should only be done if you are already familiar with how to use OpenRefine.

For option 1, save the CSV on a workstation, as you will use it later when you create archival description. Skip ahead to the section on normalizing the files.

For option 2 or 3, open the CSV in Excel on the workstation, select everything in column A, and do a Find and Replace to remove `[…]/HomewoodPhoto/Sampled/`from the beginning of the folder names. Leave the `Replace with` field blank.



Copy and paste everything from column A into column B and rename column B from “oldFolder” to “newFolder”

A	B
oldFolder	**newFolder**
20080327_gazette_YoungInvestigatorBrigdet	20080327_gazette_YoungInvestigatorBrigdet
20080730_designAndPubs_photos	20080730_designAndPubs_photos
20040831_admissions_deanSusanBoswell	20040831_admissions_deanSusanBoswell
20041020_admissions_marcusArtis	20041020_admissions_marcusArtis
20070128_ksas_students	20070128_ksas_students
20070201_ksas_students	20070201_ksas_students

The folder names in column A can not be edited, they will be needed later in order to overwrite the names on the actual folders. To continue with option 2, follow the below instructions. For option 3, save the file as a CSV and skip ahead to that section.

Option 2 - Excel

Select column B and continue using Find and Replace to remove special characters, spaces, and unnecessary description. Be cautious with removing the word “part” using Find and Replace, as there are many folder descriptions that include the word “party” - instead try “part_1” or “part1” to work around this. Find and Replace in Excel is not case sensitive.

It’s recommended to choose “Find All” before “Replace All" to ensure you are only acting on column B. Skip past option 3 to the Importing Folder Names section.

Option 3 - OpenRefine

The following expressions are a few options available for experienced OpenRefine users, thus are accompanied by minimal instructions.

- Remove commonly occurring words:

```
Edit cells -> Transform
```

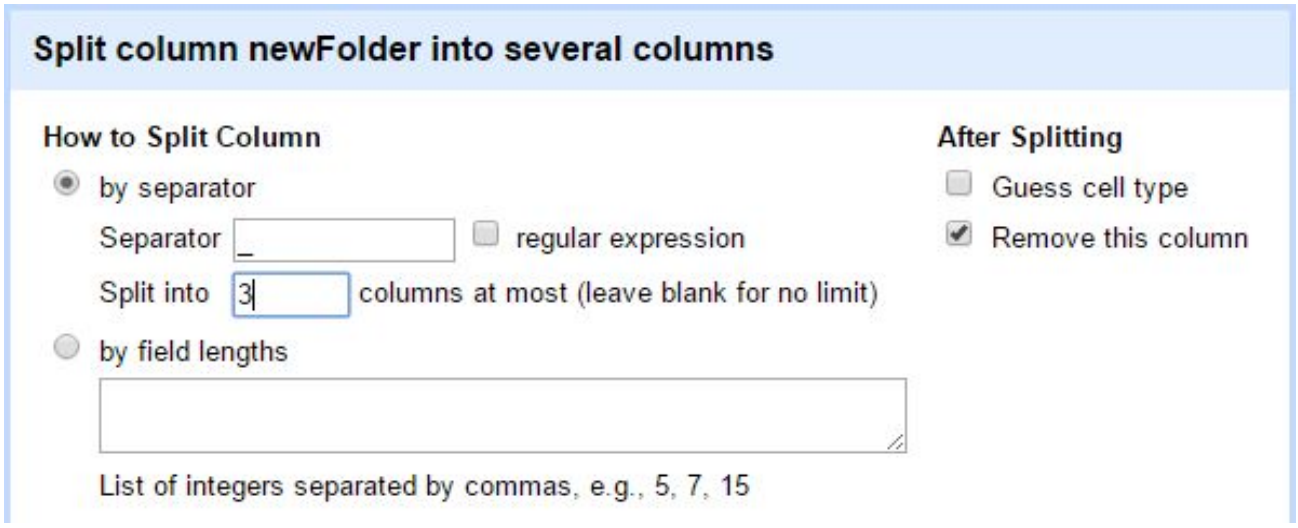
In the “Expression” box, enter expressions such as:

```
replace(value, "Raw","")
replace(value, "raw","")
replace(value, "part_1","")
```

And so on, keeping in mind that OpenRefine is case sensitive.

- To split the newFolder column into three parts (date_client_event) so you can work on each distinct part of the name:

```
Edit column -> Split into Several Columns
```



- Some folder names will still have underscores remaining in the third part of the name. In order to remove them, you will have to run a text transformation until it says “Text transform on 0 cells.” This expression will also capitalize each letter that follows an underscore, so that it is easy to read. From the dropdown next to “newFolder 3” select:

```
Edit cells -> Transform
```

In the “Expression” box, enter:

```
if(indexOf(value,'_') >= 1, substring(value, 0,
indexOf(value,'_'))+toTitlecase( substring(value,
        indexOf(value,'_')+1)), value)
```

Select “Re-transform up to `10` times until no change”

oldFolder	newFolder 1	newFolder 2	newFolder 3
20081002_alumni_crab_feast_raw_part_1	20081002	alumni	crabFeast
20081003_alumni_leadership_dinner_raw_part_1	20081003	alumni	leadershipDinner
20081004_alumni_presidential_address_raw	20081004	alumni	presidentialAddress
20081011_alumni_pumpkin_farm	20081011	alumni	pumpkinFarm
20081002_alumni_council_chair_raw	20081002	alumni	councilChair

- Additional standardization can be done through alphabetical sorting, faceting, and filtering, such as to assign ‘jhu’ as the default client for jobs that only have an event, assign ‘photos’ as the default event for jobs that only have a client, or to standardize abbreviations for clients, such as ‘carey’ instead of ‘cbs’ and ‘carey_business_school’

- The last OpenRefine step is to concatenate the date + client + event. From the dropdown next to “newFolder 3” select:

```
Edit column -> Add column based on this column
```

Name the new column “newFolder” and in the “Expression” box, enter:

```
cells['newFolder 1'].value+'_'+cells['newFolder
   2'].value+'_'+cells['newFolder 3'].value
```

Export the project, and follow the next steps to rename the folders.

Importing Folder Names

When you are done with Excel or OpenRefine, save the file as a CSV called “FolderNames.csv” on the BitCurator desktop and run `renameDirectories.py`to import the new names to the job folders.

```
$ cd [directory where Python script is stored]
$ python renameDirectories.py
```

This generates a CSV log called “`renameLog[Date][Time]`” and saves it wherever the script was run from; move the log to the `CompletedBatches` directory.

2.6 Normalizing

The .NEF, Nikon proprietary camera RAW files, are converted to .DNG for preservation, Adobe's open RAW format (see File Formats for Preservation for more information). Adobe Digital Negative Converter is the best tool for normalizing to DNG because it will recurse through directories, but it isn't Linux compatible. Restart the FRED and boot into the Windows partition to use the Adobe DNG Converter.

Check the settings:

1. Select Folder: `…\HomewoodPhoto\Sampled\`
   Include images contained within subfolders: `Yes`
   Skip source image if destination image already exists: `Yes`
2. Select location to save converted images: `Save in Same Location`
3. Select name for converted images: `Document Name`
4. Preferences – (default) Compatibility: `Camera Raw 7.1 and later`

You can open multiple converter windows at the same time, and convert `Sampled` and `Peabody` simultaneously (*do not normalize Chesney*).[7]

When the converter is done, go back to the BitCurator partition to delete the .NEF files:

```
$ cd […]/HomewoodPhoto/Sampled
$ find . -name "*.nef" -type f -delete
```

Due to case sensitivity and changes to how the files were saved over time, repeat this command to delete both "*.NEF" and "*.nef"

Repeat the above commands for the `Peabody` directory.

Log that content was normalized on the `TransferProgress` sheet.

Collapsed Dirs	Deleted Empty Dirs	Sampled	Normalization
X	X	X	20170309
X	X	X	20170309
X	X	X	20170309

[7] As of writing this documentation, Peabody wants .DNG files and Chesney is keeping the .NEF files.

2.7 Transferring Processed Content

Transferring to Preservation Storage

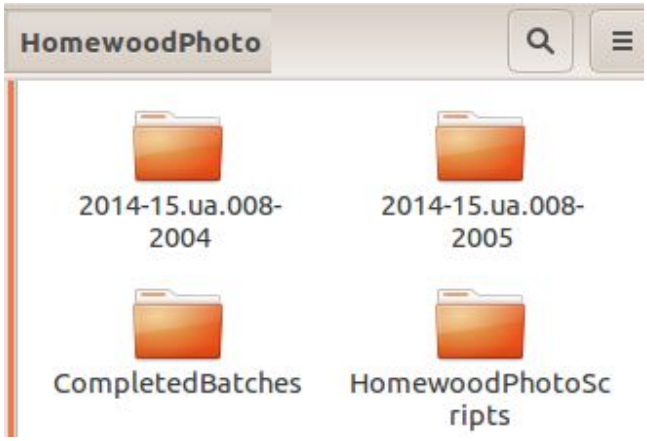
The steps to transfer to preservation storage generally follow the archives' Electronic Records Accessioning Workflow[8] to:

- Mount the GDrive in BitCurator
- Print directories
- Bag
- Tar
- Rsync
- Create event records in ArchivesSpace

Before printing the directories, create a new folder for each year of content, directly within the `HomewoodPhoto` directory, using the following naming convention:

```
[accession-number]-[year] Ex.: 2014-15.ua.008-2004
```

Move the content out of the `Sampled` directory and into the appropriate year.



You can now follow the electronic records accessioning workflow to print directories, bag, tar, and rsync the content by year. If storage space permits, it is best to wait until the whole accession has been processed before going through these steps; to add more content from the same accession number and year later, copy the tarred bag back from SAM, untar, add the content, and follow the process again.

[8] The internal version is located in the Wiki and should be consulted for the most up-to-date workflow.

Transferring Manifests to GDrive

Create copies of the manifests from each bag and place then in the `Manifests` folder. When you are done bagging everything, copy the manifests to the accession folder to the GDrive. First, mount the GDrive, then:

```
$ sudo cp -a [source folder] [destination folder]
```

The `source folder` is `[…]/HomewoodPhoto/Manifests`
and the `destination folder` is `/mnt/[path/to/accession-folder]`

Log the message digest calculations and ingest to SAM on the `TransferProgress` sheet.

H	I	J	K
Sampled	**Normalization**	**Message Digest Calcuation**	**Ingestion**
X	20170309	20170706-20170707	20170721-20170804

Note that these activities are no longer linked directly to the batches as the content is now organized by year, however, you can provide a range of dates for the preservation events associated with the content.

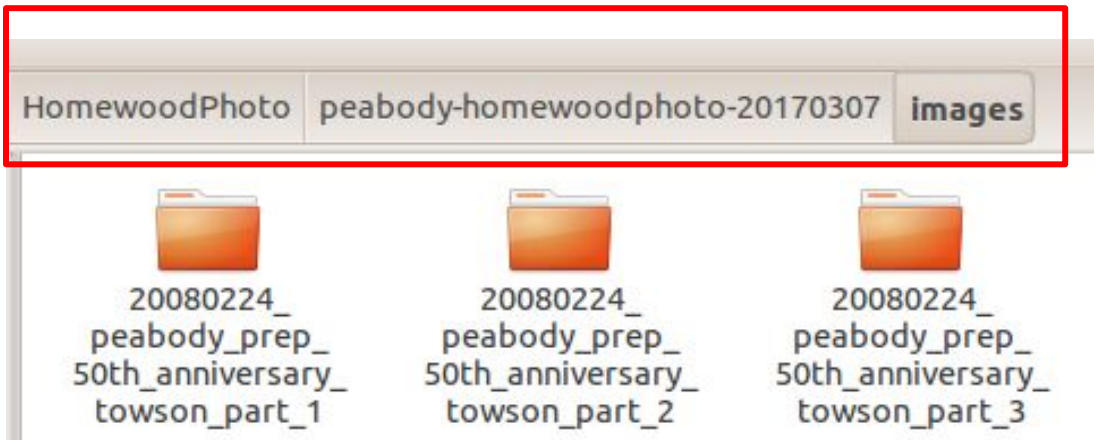
Also copy the `CompletedBatches` folder, containing the batch logs, sampling logs, and folder name conversions (if applicable) to the GDrive accession folder.

Transferring to Other Repositories

It is a slightly different process to prepare content for transfer to other repositories. Create a new folder directly within the `HomewoodPhoto` directory, using the following naming convention:

```
[reponame]-homewoodphoto-[todaysdate]
Ex.: peabody-homewoodphoto-20170825
```

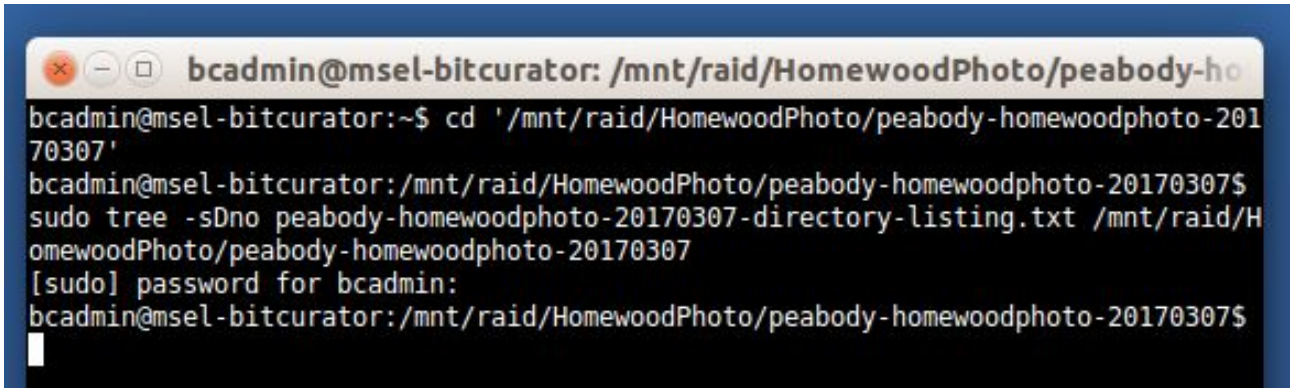
Move the entire **folder** of content into this new folder, rename the folder containing the content "images"



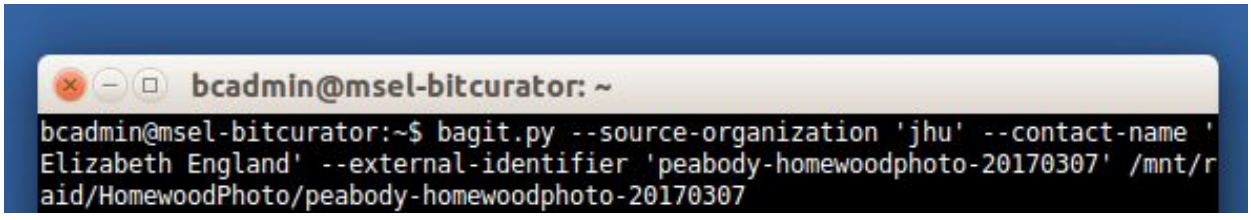
Print a directory listing:

```
$ cd […]/HomewoodPhoto/[reponame]-homewoodphoto-[date]
$ sudo tree -sDno
[reponame]-homewoodphoto-[date]-directory-listing.txt
[…]/HomewoodPhoto/[reponame]-homewoodphoto-[date]
```

This will place the directory listing inside the folder for the repository, but outside the images folder.



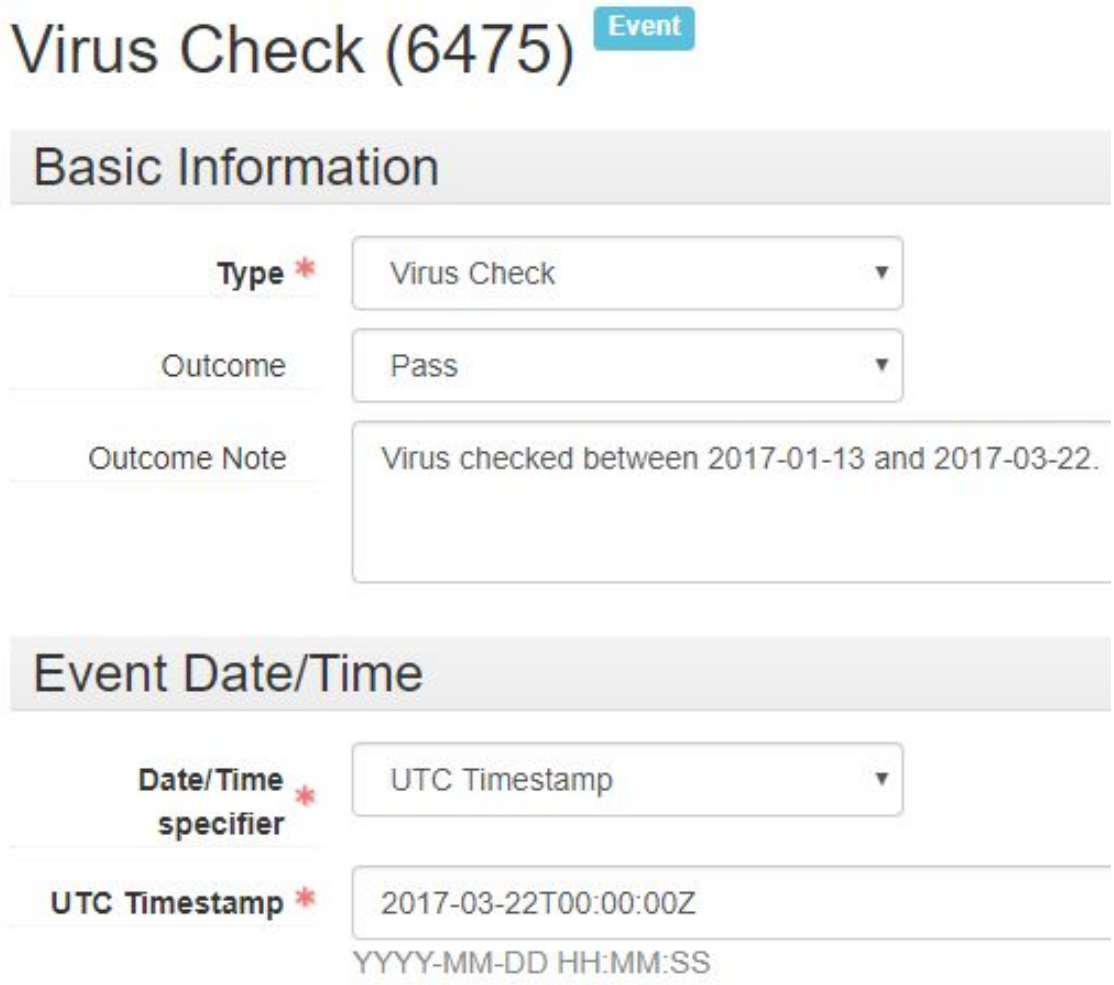
Proceed with bagging as usual, and the directory listing and images will both now be contained in the data directory.



Contact the archivist(s) at the appropriate repository to arrange a transfer. Upon completing the transfer, verify they have access to the content before deleting it from the RAID.

2.8 ArchivesSpace Event Records

The information recorded on the `TransferProgress` table can now be translated to ArchivesSpace event records in the accession record. Although there are multiple instances of one event type, such as virus checks, **create one event record per event type**. Use the latest date associated with the event type for the `UTC Timestamp`, and use the `Outcome Note` field to provide the date range for the events:



See the accession record for 2014-15.ua.008 for more example event records, and the definitions of events in the Electronic Records Accessioning Workflow.

Event types to record are:

- **Capture** - Batch dates
- **Virus Check -** Outcome of virus scans
- **Normalization** - Conversion to .DNG
- **Message Digest Calculation** - Bagging
- **Ingestion** - Transfer to SAM

2.9 ArchivesSpace Description

The final, but time consuming, step is to create archival description for the Homewood Photography records collection, Series 2: Digital photographs. The series is a file-level listing of each job, in alphabetical order by client and job. See Series 2 for example records, and note the following:

- Agents and subjects are linked at the file-level, whenever possible:
    - Agents: Homewood Photography is assigned role: creator, relator: photographer, and the job client is assigned role: creator, relator: sponsor.
    - Each file-level archival object is assigned DVDs as a subject, and an additional 1-2 subjects may be provided from the ArchivesSpace Subjects list.
- Each archival object has a corresponding digital object for the preservation file version
- For each job that was recorded in the Homewood Photography access database and therefore the associated DVDs are known, the archival object for that job should be linked to the appropriate container instances for the boxes and DVDs
- The series and collection level descriptions should be modified as necessary
- Each file-level archival object is assigned a Conditions Governing Access note that the digital content is available offline (this is subject to change if an access system is implemented)
- See 3.8 Access Policy for University Photographs for information about restricted content.

Multiple CSVs will need to be created and pushed into ArchivesSpace in this order:

1. New agent records for Homewood Photo clients (if applicable)
2. Container profiles
3. Digital objects
4. Archival objects

More on the order of operations and templates to use can be found on the GitHub. The easiest way to get to this end result is in this order:

1. Create the container profiles CSV.
2. Create the archival objects CSV, starting from the CSV of folder names you generated earlier. Use the `Database_Copy` to assist in assigning titles and DVDs. While working, keep the client and event description in separate columns, which makes standardization easier. Add in the ArchivesSpace IDs for all clients that have agent records. This is most efficiently done in OpenRefine; when complete, concatenate the client and event fields to create the titles.
3. The titles and folder names can be copied into a separate CSV to create the digital objects CSV.
4. Create the new agent records CSV for the agents that do not already have ArchivesSpace IDs, based on your description spreadsheet.

3.1 Content Categories for Weeding and Transfer

Category	Description	Common Abbreviations	Action
Medical	Includes all School of Medicine, School of Nursing, School of Public Health, JHMI, East Baltimore, Bayview, “Corporate Communications,” photos taken for publication (Gazette, Magazine) of content in these areas, and related units. The Biomedical Engineering Department is partially run by School of Medicine and is in scope for both Homewood and Chesney archives.	son, som, jhmi, sph	Out of scope, send to Chesney. Current arrangement is to borrow hard drive from the repository to transfer content.
Peabody	Includes all Peabody-related clients and events. *Does not include* Peabody Library.		Out of scope, send to Peabody. Current arrangement is to borrow hard drive from the repository to transfer content.
Applied Physics Laboratory	Includes all APL clients and events.	apl	Out of scope, barely any content thus far. Will contact APL once there is more.
Athletics[9]	Includes all photos taken for JHU of athletes/athletic events, team photos, Blue Jays Unlimited. *Does not include* Inside Lacrosse, non-JHU sports events.		Likely retain; verify with University Archivist.

[9] If the archives works out a transfer arrangement directly with Athletics Communications to accession their described and culled selection, may not retain these photos from Homewood Photography.

Content Categories for Weeding and Transfer, continued

Category	Description	Common Abbreviations	Action
Non-JHU	Includes content or clients unaffiliated with JHU.		Out of scope, do not retain.
Egypt and Syria	Photos taken to document original NES/archaeological research in Egypt and Syria. *Does not include* 2008 Admission trip to Egypt.	egypt, syria	Out of scope (research data), do not retain.
Objects and copy work	Photos of museum objects, books, etc. Includes photos taken for publication purposes, art history ("hoa") department, Special Collections, etc.	hoa	Most jobs don't have any descriptive information. Low archival value, do not retain.
Portraits #1	Passport and identification photos, student portraits where they are the client, family portraits. *Does not include* Medical, Peabody, APL.	id	Low archival value, do not retain.
Portraits #2	All other group and individual portraits, including Faculty, Staff, Student portraits where the client is a JHU unit. *Does not include* Medical, Peabody, APL.		Retain 1-3 images per shoot.
Publications & Communica-tions	Includes photos for publication-related clients, such as Design & Publications, Gazette, News & Information, JHU Magazine, JHU Press. *Does not include* photos taken for publication of Medical, Peabody, or APL.	d&p, mag, nais	Select images have already been published, but hard to ID which. Only retain jobs where the description warrants inclusion (irregardless of the publication status).
Other Campuses	East Baltimore, Evergreen, Montgomery County, Peabody Library		Retain, standard sampling strategy.

3.2 Sampling Strategy

Guiding characteristics from relevant archival literature

Two ways to reflect the characteristics of a series:

- **Sampling** – reliable representation of the whole; quantitative
- **Selection** – qualitative reflection of some predetermined significant characteristic of the whole
- Systematic sampling: space evenly, that is, every “nth” case selected (rather than using a random number chart)
    - If selecting quantitatively and qualitatively, the quantitative sample is determined first – then qualitative selection made from the remaining case files. Result is retention of both the routine and the most important.
- Series with greater homogeneity are good for quantitative sampling, where the individual file is not important
- Samples should be large enough to contain all the key elements of evidence or information within the series
- Often a 5% sample is sufficient standard of retention – can go smaller or larger based on size of the series and its content

Process

- Examine the record group to determine the homogeneity of content by looking at the beginning, end, and middle files
- For highly homogenous content (portraits, lower archival value): qualitatively select. Select 1 or 2 images per job.
- For all other types of in-scope content (campus, reunions, symposia, etc.): quantitative, systematic sampling. Select every 10th image, beginning with the 2nd image, because the 1st is often a color chart (resulting in a 10% sample).
    - When the resulting sample is more heterogeneous (indicating a job with diverse content): do a second pass, looking at thumbnails of the remaining case files and qualitatively select, if necessary

Resources Consulted

Boles, Frank. “Sampling in Archives.” *American Archivist* 44, no. 2 (Spring 1981): 125-130.

Bradsher, James Gregory and Bruce I. Ambacher. “Archival Sampling: A Method of Appraisal and A Means of Retention.” *Mid-Atlantic Regional Archives Conference* Technical Leaflet no. 8 (1992): 1-24.

Cook, Terry. “‘Many are called, but few are chosen’: Appraisal Guidelines for Sampling and Selecting Case Files.” *Archivaria* 32 (Summer 1991): 25-50.

3.3 Folder Naming Conventions

Since ~ 2005, Homewood Photo has largely followed the following convention:

YYYYMMDD_client_event (ex.: 20160301_office_of_investment_management_staff_portraits)

Folder naming should follow the following characteristics:

1. Use underscores (_) to separate discrete parts of the naming, not entire words. It is currently too hard to understand where the client name ends and event name begins (is the above “Office of Investment Management” or “Office of Investment”?). Want to control the overall number of parts in the file names, thus no other special characters, dashes, or spaces.
2. Camel case the beginning of words (“officeOfInvestment” instead of “officeofinvenstment”)
3. If character limit is a concern, can reduce the length of the folder names through shortening the event, as necessary (“portraits” instead of “staff portraits”).
4. Preserve original folder names in the GDrive accessions folder, detailed in the workflow, Part 2.5 Folder Names.
5. Occasional typos and misspellings should be corrected. Use OpenRefine to standardize client and event names, such as correcting duplicate clients that are the same (“Office of Investment” and “Investment Office”).

Parts of the Folder Names

1. **Date String**: begin with an eight-digit date string. Follow the date with an underscore.
   Example: 20070515_
2. **Client**: The second part of the folder name contains the client. In the absence of a client name, use “jhu.” Standardize the client naming so “jhu magazine” and “jhu mag” both become “jhuMag.” Follow the client with an underscore.
   Examples: 20070515_jhu_ (unknown client)
   20070515_jhuMag_
3. **Event**: The final part of the folder name contains the event. In the absence of an event name, use “photos.” (“20050810_wse_photos”)
   Examples: 20070515_jhuMag_photos (unknown event)
   20070515_ksas_staffPortraits

3.4 File Formats for Preservation

This document serves to recommend a preservation file format for the Homewood Photography collection. In researching formats, I looked at best practices within other institutions (Library of Congress, NARA, and Harvard), resources such as PRONOM and the Library of Congress' "Sustainability of Digital Formats," documentation on the file formats used in applications such as Archivematica, and consulted the SAA Electronic Records listserv and Digital Curation Google group for input on what formats others have chosen, and why. From reviewing these resources, I decided to do more targeted research into three file formats: TIFF, JPEG/JFIF, and DNG. Main considerations in my research and the resulting recommendations are the resources required/anticipated for managing images stored in the selected format, namely: time, expertise, and storage space.

I selected TIFF because it is the preferred format for digital photographs by the Library of Congress, NARA, Harvard, and numerous other institutions. TIFF is the longest running format used for still image preservation, and its use has carried over from digitized to born-digital images. Its wide adoption makes it a risk-adverse choice, although the files require significant storage space.

JPEG/JFIF is considered primarily because of the small file sizes, which would reduce the storage space required, thus allow for saving a greater number of images and/or reduce the cost of storage for the collection. While JPEG/JFIF isn't the most popular choice for a preservation format because of its lossy compression, it is a well-supported format and nevertheless still the second choice preferred format for NARA and in the top five choices for both Harvard and LOC. On the choice to use JPEG for a collection of campus photographs, SUNY Albany University Archivist Greg Wiedeman wrote to me, "For our university photos we converted all of our NEF to just lossy jpegs. I'd imagine that this is fairly controversial as we deleted a lot of data and jpeg compression can be pretty terrible. The reason is that I thought that uncompressed or raw masters would never see any use as the purpose of these photos was to document university events. I thought jpegs retain the information that matters and maintaining terabytes of unused masters has a cost. In our tests, jpeg compression preserved the look of the images better than png or other formats."

File Formats for Preservation, continued

DNG is the only open image format specifically designed for the preservation of camera RAW. In the Homewood Photography collection, the files from which preservation copies will be derived are NEF, which is proprietary Nikon camera RAW. While DNG would not make sense for all still image collections, institutions are moving away from TIFF or JPEG2000 in favor of DNG for collections that begin with camera RAW files. Those who have not yet made a switch to DNG generally consider it a good option (although caution that it isn't supported by all image viewers). Archivematica is also considering designating DNG as the preservation format for RAW files in the future (the current policy is to keep the proprietary RAW formats as the preservation copies).

I tested workflows for normalizing to these file formats, as well as other formats that I ultimately decided against, namely JPEG2000 and TIFF with LZW compression. As I found from testing the workflow, TIFF-LZW files were too large to justify the added complexity of compression. While JPEG2000 has smaller file sizes, it has never been fully adopted in the digital preservation community due to its complex format, and now many institutions are looking to DNG to replace JPEG2000. The accompanying table provides a comparison of the three selected formats; the actual details for each workflow would be cemented after selecting the format.

File Formats for Preservation, continued

	TIFF (revision 6.0)	**JPEG/JFIF (version 1.01)**	**DNG (version 1.4)**
Pronom Identifier	fmt/353	fmt/43	fmt/730
Why Considered	Community best practice; widely accepted and supported; bit lossless.	Saves on storage space; opportunity to use same format for preservation & access; auto preserves EXIF metadata	Suggested by Homewood Photographers; alternative to proprietary RAW camera formats to save unprocessed camera data; compatible with TIFF-EP standard
Adoption	Widely deployed as preservation format; widely supported	Widely adopted, but typically not as a preservation format	Not yet a standard format in ISO, but is based on several open formats/standards; increasingly adopted as a preferred format
Compression	Uncompressed	Lossy	Lossless; starting with version 1.4, option for lossy compression
Disadvantages	Storage space required; might be best for low-volume/high-value collections	Lossy compression; can't later get back lost information	Interoperability issues outside Adobe products; not yet a standard format
Retains embedded metadata when normalizing	No	Yes	Yes
Approx. storage implications (per individual image, unresized)	50 MB	2 MB	10 MB (lossless compression) 20 MB (uncompressed)

Recommendations

The ultimate goal of preserving Homewood Photography's born-digital photographs is to make them accessible. It is worth considering the collection's anticipated use and how it will best serve researchers. I do not foresee it being used as a fine art collection and am not sure there is a use case for needing to preserve the raw image data; rather I believe the photographs will be used for their informational value about people, events, and places related to the Homewood campus.

While DNG is arguably the best-suited format for this particular collection because it is specifically for camera RAW and is lossless, JPEG/JFIF may be better suited for the archives at a broader level, considering resources. If the JPEG/JFIF files are left alone and are not being resaved, there is little known risk of the images further compressing and becoming lossier overtime. However, while a DNG file could always later be converted into a JPEG, it is not possible to recreate the RAW camera data from a processed JPEG, thus the decision could not be made later to convert JPEG to DNG. JPEG is a "safer" choice in that it is already an acceptable format that the archives is well-equipped to manage in the long-term, and DNG is a "safer" choice in that it permits greater long-term flexibility and is more secure from data loss.

Resources Consulted

"Appendix A: Tables of File Formats, 4.1 Digital Photographs." *National Archives*. Accessed December 6, 2016. https://www.archives.gov/records-mgmt/policy/transfer-guidance-tables.html#digitalphotographs.

"Format policies." *Archivematica*. Last modified December 24, 2015. https://wiki.archivematica.org/Format_policies.

Goethals, Andrea and Patricia A Patterson. "Formats Supported by the DRS." *Harvard Library Digital Preservation*. Last modified December 1, 2016. https://wiki.harvard.edu/confluence/display/digitalpreservation/Formats+Supported+by+the+DRS.

Goethals, Andrea and Jack Holm. "Camera Raw Images / Adobe DNG." *Harvard Library Digital Preservation*. Last modified April 5, 2016. https://wiki.harvard.edu/confluence/pages/viewpage.action?pageId=207554565.

LeFurgy, Bill. "Is JPEG-2000 a Preservation Risk?" *Library of Congress: The Signal*. January 28, 2013. http://blogs.loc.gov/thesignal/2013/01/is-jpeg-2000-a-preservation-risk/.

Resources Consulted, continued

"PRONOM." *The National Archives*. Accessed January 5, 2017. http://www.nationalarchives.gov.uk/PRONOM/Default.aspx.

"Recommended Formats Statement 2016-2017." *Library of Congress*. Accessed October 31, 2016. http://www.loc.gov/preservation/resources/rfs/RFS%202016-2017.pdf.

Robbins, Kevin Patrick. "Why I Stopped Using the DNG File Format." *PetaPixel*. July 16, 2015. http://petapixel.com/2015/07/16/why-i-stopped-using-the-dng-file-format/.

"Sustainability Factors." *Library of Congress*. Last modified March 20, 2013. http://www.digitalpreservation.gov/formats/sustain/sustain.shtml.

"Sustainability of Digital Formats: Format Descriptions for Still Images." *Library of Congress*. Last modified December 16, 2015. http://www.digitalpreservation.gov/formats/fdd/still_fdd.shtml.

"Williams College Archives and Special Collections File Format Recommendations." *Williams College* Accessed December 6, 2016. http://archives.williams.edu/files/FileFormatRecommendations_FINAL.pdf.

3.5 ArchivesSpace Subjects

Proposed Subject (for HIPS database)	ASpace subject	ASpace ID
Academics	FAST: “Universities and colleges—Curricula”	subjects/938
Administration	FAST: “Universities and colleges—Administration”	subjects/836
Alumni	FAST: “Universities and colleges—Alumni and alumnae”	subjects/798
Athletics	FAST: “Universities and colleges—Alumni and alumnae”	subjects/939
Buildings	FAST: “College buildings”	subjects/837
Campus Life	FAST: “Universities and colleges—Social aspects”	subjects/940
Commencement	FAST: “Commencement ceremonies”	subjects/446
Construction	FAST: “College buildings—Design and construction”	subjects/941
Faculty	FAST: “Universities and colleges—Faculty”	subjects/835
Lectures	FAST: “Lectures and lecturing”	subjects/600
Portrait - Group	AAT: “group portraits”	subjects/943
Portrait - Individual	AAT: “portraits”	subjects/944
Staff	FAST: “Universities and colleges—Employees”	subjects/942
Stock	FAST: “Stock photography”	subjects/945
Students	FAST: “College students”	subjects/441

3.6 Recommendations for Homewood Photography

Folder Naming

- **Avoid Special Characters** – such as spaces, apostrophes, “&” symbol, periods as they can cause interoperability issues between operating systems and can result in errors

- **Consistent formatting** – YYYYMMDD_client_event (ex.: 20170423_AlumniRelations_Awards)
    - Continue using underscores to separate these three parts because underscores are considered an acceptable special character
    - Instead of separating words with additional underscores, use camel casing (ex.: CamelCasing) to ensure legibility *or* hyphens (ex.: 20170423_Alumni-Relations_Awards)
    - Doing this will make it easier to visually see where the client name ends and event name begins. It also helps to later parse the job names and increases the ways to sort job names (by client, by event, etc.)

- **Consistent client names and abbreviations** – use a controlled list for client names that are short but clear (ex.: “carey” instead of “CBS,” “carey_business_school”)
    - I created a table in Access called “Lookup_Clients” and begun populating it with full, proper client names (for content I have gone through thus far)
    - New clients can be added as needed, and it is recommended to assign each client a corresponding, unique abbreviation to use in folder naming

- **Consistent event names** – for events that occur regularly, use standardized names (ex.: “graduation” versus “commencement”)

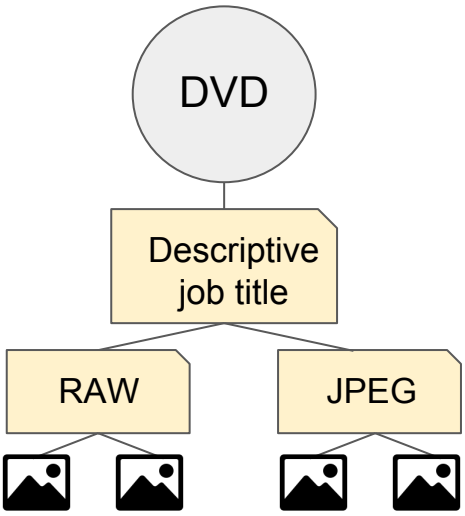
- **Length –** be descriptive, but brief
    - Different operating systems, software, and storage devices can impose different length limitations for file paths, generally a maximum length of 255 characters
    - I recommend limiting the folder name to around 30 characters, and using the Access database to expand the description

- **Spelling** – particularly with peoples’ names, I found many instances in which the Access database entry and folder name were spelled differently. It can be hard to find someone when the only information is “portrait Sloan” and “portrait Sloane” (common name), or conversely for a more unique name such as “Konchin” in the database when the actual name is “Karchin”

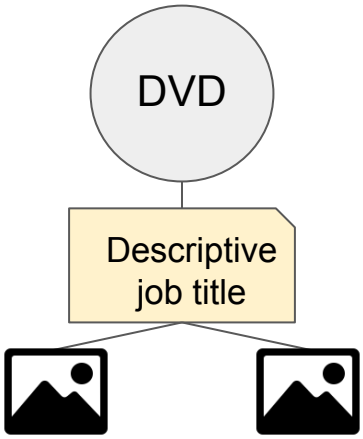
Organization

- **Consistent structuring of the folders that organize the jobs** – the below structure recommendations apply regardless of the content carrier (DVD, external hard drive, etc.)

    - If the images are being saved as more than one file format, continue to separate into different folders by file format. These folders (“raw,” “jpeg” etc.) should be nested directly below the descriptive job folder name:



    - If the images are being saved as only one file format, such as has been discussed with only keeping .DNG files, the individual images should be nested directly below the descriptive job folder name:



    - If saving a smaller selection of edited shots, create a folder nested directly below the descriptive job folder name, and choose a consistent name for the folder such as “Edited”

Recommended Database Fields

Database Field	Example	Notes
Date *Required, 8 char. limit*	20071203	-8 digit string, YYYYMMDD
Client *Required, controlled list*	Whiting School of Engineering	-Will pull from “Lookup_Clients” table and auto-complete -If the client doesn’t already exist, enter it on the Lookup_Clients table to store it as the master name for future use -Should always be the name of an office, department, person, family (Department of Biology, Bloomberg Family, etc.) rather than an event (Black History Month, Senior Week, etc.) -Avoid abbreviations -If there is no identifiable client, such as for stock photos of campus, the client should be “Johns Hopkins University”
Event *Required, free text*	Professor Ed Bouwer portrait	-Provide fuller description than the folder names (first and last names, role/position of student, professor, etc.) -Avoid abbreviations
Subject *Required, controlled list*	Portrait - Individual Non-JHU	-Will pull from “Lookup_Subjects” table -Assign broad categories to the content -1 subject minimum required, 2 optional
Rights *Required (ideally), controlled list*	Unrestricted Contractual restriction	-Any known restrictions on use/privacy (private client, photos of minors, construction contractor agreement, etc.) -Build off rights use cases you already have
Description *Optional, free text*	Saved as TIFF Baltimore Inner Harbor	-Any additional information about the job content -Could include explanation for any deviation from standards, such as an alternative file format, etc. -Include location information for off campus jobs (if not in “Event” field)
External Drive *Optional, number*	1	-External drive the job is stored on
Folder Name *Optional, 35 char. limit*	20071203_wse_ed Bouwer	-Folder name associated with the job -Consider character limit
Complete?	Yes/No	-Use checkbox to indicate the job and data entry are complete

Recommended Database Subjects

Subject	Use
Academics	Jobs related to classes, tutoring, etc.
Administration	Board of Trustees, Provost and President related events, etc.
Alumni	Reunions, Second Decade Society, etc.
Athletics	Sports games, ceremonies, etc. If using this for non-JHU athletics as well, ensure using the "Non-JHU" subject as well.
Buildings	Campus buildings. Further description of what buildings is appropriate in the "Event" field.
Campus life	General documentation of campus life and can include events such as Spring Fair, etc.
Campus objects	Outdoor sculpture, murals, etc. Does not include copy works.
Commencement	Commencement and graduation events.
Construction	Campus construction and renovations.
Copy works	Special Collections and History of Art photos of books, photos of paintings, etc. Does not include campus objects.
Development	Jobs related to University development, such as donors events.
Faculty	Faculty events, faculty portraits, etc.
Lectures	Lectures and speakers, including series.
Medical	School of Medicine, School of Nursing, Bayview, etc.
Non-JHU	All non-Hopkins jobs.
Peabody	All Peabody Conservatory jobs, does not include the Peabody Library.
Portrait - Group	Posed group photos.
Portrait - Individual	Posed individual photos including headshots.
Publications	Jobs taken for use in publications, such as for the Hub, JHU Press, etc.
Staff	Staff events, staff portraits, etc. Includes all non-faculty employees.
Stock	Use "Johns Hopkins University" as the client.
Students	Student events, student portraits, etc.

Database Entry

- Spell out abbreviations and use the current, official names – this is especially helpful as office and department names change over time, and it can be harder to retroactively figure out what an abbreviation refers to when the official name has changed
    - Changes to official client names can be tracked on the “Lookup_Clients” table in the Access database
- Always populate the client and event fields – ensure the billing system requires this information to be entered when clients book jobs
- Include value-adding information in the event title – such as the subject’s first and last name (when known), “portrait,” etc.

Resources

- “Digital Photography Best Practices and Workflow” American Society of Media Photographers: http://www.dpbestflow.org/
- “Best practices for file naming” Stanford Libraries: https://library.stanford.edu/research/data-management-services/data-best-practices/best-practices-file-naming

For Further Consideration

- How can the billing system be leveraged to further help with description of the client, event, and assigning subject terms from a limited list?
- Are there rights use cases that can serve as a basis for developing a few standard rights statements?
- How to balance logging the client as in the responsible party for billing, versus the client as in the most closely intellectually linked university office, department, or person(s) that would make sense to a user searching the archives for images?
- Training new students on database entry when the fall semester begins
- Storing database in a shared location
- Editing down shoots

3.7 Additional OpenRefine commands

- Reformat a 6 digit date as 8 digit, with leading year:

```
Edit cells -> Transform
'20'+substring(value, length(value)-2,
length(value))+substring(value, 0, length(value)-2)
```

- Transform YYYYMMDD to YYYY-MM-DD

```
Edit cells -> Transform
value.slice(0,4)+'-'+value.slice(4,6)+'-'+value.slice(6,8)

cells['date2 3'].value+' '+cells['date2 1'].value+'
'+cells['date2 2 1'].value
```

- Isolate extra underscores:

```
Custom Facet
if(indexOf(value,'_') >= 1, 'true', 'false')
```

3.8 Access Policy for University Photographs

The access policy is still being finalized as of the end of my residency; this serves as a placeholder and the policy will be added here when it is confirmed.

4.1 Goals & Objectives from NDSR Project Proposal

The following are reflections and summaries of the goals and objectives from the original NDSR Project Proposal.

Goal 1: Select Homewood Photography born-digital images in accordance with University Archives collection development policy

- **Objective 1.1: Development of appraisal criteria that aligns with University Archives collecting strategy**

  Appraisal criteria was developed in consultation with the University Archivist and is captured in 3.1 Content Categories for Weeding and Transfer. The University Archives has retained content from Homewood Photography that represents the activities, people, buildings, and events of the Homewood campus.

- **Objective 1.2: Appraisal of about 50 terabytes of images to determine what selection should be permanently preserved by the Archives (expected yield is 10-15 terabytes, or a possible 30-50% reduction)**

  As of the beginning of the NDSR project, the Archives had accessioned 2,900 DVDs from Homewood Photography, or approximately 13 terabytes. Following the appraisal criteria resulted in approximately a 50% reduction of content, as significant amounts belonged to the Johns Hopkins Peabody Institute Archives or Alan Mason Chesney Medical Archives, or was of low archival value such as passport and visa photos.

  The content that fit within the Archives' collecting strategy was further reduced through automated sampling, detailed in 3.2 Sampling Strategy, to retain 10% of each job shot by Homewood Photography. After normalizing the files, which also reduced the storage implications, a total of 227.4 gigabytes was selected for long-term preservation by the Archives, approximately a 98% reduction from the 13 terabytes accession.

  In July 2017, the Archives received another accession from Homewood Photography of 3,100 DVDs, again approximately 13 terabytes. While 50 terabytes were not appraised during the course of the NDSR project, there was not 50 terabytes in the physical custody of the Archives to be appraised. The accession that pre-dated the NDSR project was completed, and work was begun on the 2017 accession, with the Archives' Processing Archivist being trained on the workflow detailed in this document. Homewood Photography still holds their content from the past few years to the present, and continues to generate content at a rate of about 4 terabytes per year.

Goal 2: Improve the long-term preservation of Homewood Photography's born-digital images

- **Objective 2.1: Creation and application of unique identifiers that clearly map the metadata to the images in a machine-actionable way**

  The images are mapped to metadata through the use of archival and digital objects in ArchivesSpace. The descriptive metadata for each job has a unique archival object ID number and is linked to a corresponding digital object, with a unique digital object ID number. Each digital object record contains a publicly viewable identifier (the folder name), and an internally viewable file URI for the location of the preservation files. In time, there could be publicly viewable file URIs for the location of access files, should the images be available to users in DSpace, for example.

- **Objective 2.2: Migration of in-scope images from CDs, DVDs, servers, and other carrier media to networked storage managed by the Archives for long-term preservation**

  Thus far, all images transferred to the Archives have been on DVDs. The content was captured using a RipStation, transferred to a networked FRED for processing, then transferred to long-term preservation storage managed by the Sheridan Libraries' Digital Research and Curation Center systems administrators. This process has been completed for the accession received prior to the NDSR project, and is in progress for the accession received in July 2017. Homewood Photography decided to move away from storing content on DVDs in 2017, and instead use external hard drives. While there are still thousands of DVDs to transfer to the Archives, eventually the Archives should anticipate receiving content on external hard drives.

- **Objective 2.3: Conversion of all images to target formats for access and preservation**

  All images were converted to .DNG, Adobe Digital Negative, as the target preservation format. Access copies were not created, as the provision of access was out of scope for this project and it was deemed unnecessary to create access copies at this time. Potential ideas discussed for access include making the images available in DSpace, the library's current repository for delivering digital content, or creating access copies in response to researcher requests, and providing access to those on a non-networked reading room computer. In either case, the conversion of images to the target format for access, likely JPEG, will take place when appropriate.

- **Objective 2.4: Creation of functional requirements for the preservation of the images that align with functional requirements for the preservation of other digital assets the library plans to preserve long-term**

  The images have been preserved according to the same standards set for other digital content managed by the Archives. The steps taken in preparation of long-term preservation included converting the images to an open, lossless file format, and following additional steps outlined in the Electronic records accessioning workflow, such as using bagit-python to generate checksums and package content, tar the bags, and use rsync to safely transfer the content to SAM, the secured preservation storage system managed by the Libraries' IT department, where there is one spinning copy and two tape copies, one located off-site.

- **Objective 2.5: Evaluation of existing preservation metadata standards (such as PREMIS) and application of standards, as appropriate**

  Working outside of a system that generates PREMIS records, such as Archivematica, it proved challenging to consistently and specifically associate preservation events with the content it pertained to, as the arrangement of the content shifted from arrangement by disc numbers to chronologically. In the earlier stages of processing (pre-arranging the content), the preservation activities of capture and virus checking clearly map back to specific ranges of DVDs. By later processing steps of normalization, message digest calculation, and ingest (post-arranging the content), it was no longer possible to map the preservation activities back to the content using the same tracking methods. For example, while it was possible to record the date of virus checking DVDs numbers 100100-100400, discs from within that range would later be arranged and bagged according to date, thus impossible to record a date for bagging the content from DVDs 100100-100400, if done over a period of a few days.

  Discussions on the SAA Electronic Records listserv, Digital Curation google group, and PREMIS listserv revealed that this is not an issue many people have dealt with, let alone considered, particularly those who work in PREMIS environments. Ultimately, preservation events were tracked on a spreadsheet throughout processing, such as the dates of capture, virus checking, normalization, message digest calculation, and ingest. After the last step of ingest to preservation storage had been completed for all content in the accession, event records were created in the ArchivesSpace accession record. Each event type, such as virus checking, received only one event record and the date range that event type took place across all content in the accession is given in the outcome note field, such as "Virus checked between 2017-01-13 and 2017-07-06."

Goal 3: Improve access to Homewood Photography's born-digital images

- **Objective 3.1: Editing and standardization of descriptive metadata**

  Descriptive metadata about the images, including job titles, clients, dates, and subjects, was edited, standardized, and/or supplied using OpenRefine and referencing Homewood Photography's Access database. Additional work was done to align Homewood Photography's client names with agent names in ArchivesSpace, create new agents for university departments and offices as needed, and link the archival object records for each Homewood Photography job to the appropriate agent records associated with the job in ArchivesSpace. Description was further enhanced by assigning controlled vocabulary subjects to to the jobs in ArchivesSpace; with the intention that these item-level subject assignments will ultimately increase discoverability and improve access to the content.

- **Objective 3.2: Creation of descriptive metadata schema in collaboration with Homewood Photography**

  The descriptive metadata schema provided to Homewood Photography is detailed in 3.6 Recommendations for Homewood Photography. Collaboration with Homewood Photography informed these recommendations, such as the inclusion of a "Complete" field, and the definition of terms used in the controlled subjects list. At present, their Access database has been used solely as a spreadsheet; I presented them a mock-up of the capabilities if it was used as a relational database with tables to query for controlled client and subject lists, required and optional elements, and character limits to certain fields. A key recommendation is to map the metadata to the content through recording the folder name associated with each job in the database, which has not been done in the past.

- **Objective 3.3: Establishment of access restrictions for images that will automatically trigger at appropriate intervals**

  An access policy was agreed upon between Homewood Photography and the University Archives in summer 2017, and is currently undergoing review by the University's General Counsel. Access will be restricted for a period of time from the time of creation, with access to the public being triggered on an annual basis, as detailed in 3.8 Access Policy for University Photographs.

Goal 4: Develop a process for the accessioning of routine and automatic transfers of Homewood Photography's born-digital images to the Ferdinand Hamburger Archives

- **Objective 4.1 Reflection on the feasibility of the Archives' existing born-digital workflow for ingesting and preserving large sets of born-digital images**

  The Archives' feasibility to ingest and preserve large sets of born-digital images is more dependent on the Archives' abilities to ingest and preserve at-scale and from a variety of carrier media, and less about the type of content being ingested and preserved. While the 8 terabyte RAID on the FRED used for processing space may sound large, it couldn't accommodate the whole 13 terabyte accession at once and required iterative processing of the accession: breaking it into two parts, processing the first, bagging it and storing it on SAM, and later pulling it back to integrate with the second half of the accession. Additionally, transferring large amounts of content to SAM can cause the system to choke; described by the systems administrator "like shoving a goat down a snake." Due in part to this, the accession was subdivided by years for bagging, so as to more successfully transfer content to preservation storage (the other reasons for subdividing being organization and ease of retrieval).

  Further, the RAID has been used as temporary storage space for born-digital accessions prior to ingest to SAM. Depending on the amount or size of accessions in progress, this can limit the available space for temporarily storing large sets of content when capturing it from carrier media. As more members of the Archives team become involved in born-digital accessioning and processing, there will be increased demand for processing space and need to access the equipment used for accessioning and processing, namely the FRED. The Libraries' IT recently set up remote access to the FRED, which should help alleviate these constraints.

  Accessioning large sets of content from optical media has become more feasible due to the automation of the RipStation. However, the equipment had numerous mechanical issues over the past year which greatly impeded the speed and efficiency for accessioning. Further, it is still labor intensive to arrange and describe the materials, and this was the first born-digital content to actually be processed by the Archives, prior to ingesting it to preservation storage. There are no guidelines in the Archives specifically for processing born-digital content, and more broadly, no best practices and standards exist for processing born-digital materials, as evidenced by the session at SAA's 2017 annual meeting "What We Talk About When We Talk About Processing Born-Digital: Building a Framework for Shared Practice."

  While the Archives' existing born-digital workflow was applicable to accessioning this content, the workflow has been described by the digital archivist as a highly manual "boutique approach," and large collections such as Homewood Photography underscore the need for increased automation in the Archives' workflow in order to feasibly and efficiently ingest and preserve born-digital content.

- **Objective 4.2: Development of a workflow (informed by the above reflection) for setting up the regular transfer of images from Homewood Photography's digital asset management system to the libraries' access and preservation systems**

    Homewood Photography currently uses a web-based digital asset management system, PhotoShelter, for the delivery of jobs to clients. While there is potential for the Archives to accession content using the PhotoShelter API, it was determined to not be useful at this time, primarily because the images in Homewood Photography's PhotoShelter are all JPEGs. As covered in 3.4 File Formats for Preservation, the Archives decided to keep the NEF camera raw files for normalizing to DNG, rather than preserving the JPEG access copies. Additionally, while it was hoped that PhotoShelter could be used to house descriptive metadata about the content, the only metadata schema supported by PhotoShelter is IPTC, which is meant to be item-level description, and the content from Homewood Photography is not described at that granular of a level (although ultimately, each job is considered one archival object in ArchivesSpace). There is no capability in PhotoShelter to lock-down required metadata elements or specify controlled vocabularies.

    As there are still a few thousand DVDs left to transfer from Homewood Photography to the Archives, the workflow developed as part of this residency will continue to be used for some time, with an eventual transition to accessioning from external hard drives, a workflow the Archives is already well-equipped to execute.

- **Objective 4.3: Documentation of all policies and workflows established in course of residency and training of appropriate staff on procedures**

    All workflows and policies established during the residency are contained in this document. The Archives' Processing Archivist has been trained on the workflow for DVDs and can assist in accessioning, processing, and preserving content from Homewood Photography going forward, as there will be more accruals. Additionally, as I am continuing at Johns Hopkins as the Digital Processing Archivist, continued attention to this collection will be within scope of my position.

4.2 Future Directions

Some areas for investigation/further consideration going forward include:

- **Homewood Photography's changing practices**:
    - Homewood Photography intends to no longer save JPEGs, and to save the camera raw files as .DNG instead of .NEF. This would impact the Archives because there would not be a foreseeable need to normalize or migrate the files, and there may be more opportunity to check or generate fixity information when accessioning.
    - Homewood Photography intends to save content to external hard drives and not optical discs. This would reduce a significant amount of labor and expertise required for processing the content once transferred to the Archives.
- **Strengthening cooperation**: In my new position as Digital Processing Archivist in the Sheridan Libraries, I will also serve as the Digital Assets Manager for Homewood Photography one day a week. This presents opportunities to generate more robust, standardized descriptive metadata earlier in the records lifecycle, which would in turn allow for enhanced archival description; weed jobs so only the best photographs are being retained, which would alleviate the Archives' need to heavily sample the content; and ensure greater organization to how the jobs are stored.
- **Assigning university functions**: The Archives has begun assigning university functions to accessions and resource records for university records. Just as subjects were assigned at the file-level for individual Homewood Photography jobs, the opportunity also exists to assign university functions more granularly to this diverse content to aid in resource discovery.
- **Providing access**: There is processed content in this collection that is outside the eight year proposed restriction period. Once a plan for access is developed, whether through an un-networked computer in the Special Collections reading room or unmediated access through a pre-existing repository used by the Sheridan Libraries such as DSpace, content should be made available to researchers.
- **Preservation storage capabilities**: Although there were multiple reasons for dividing the content into multiple bags for preservation storage, this may not be ideal for other large amounts of content. It is worth investigating how capacity might be increased to ease large transfers to SAM so that collections do not have to be arbitrarily divided for storage.