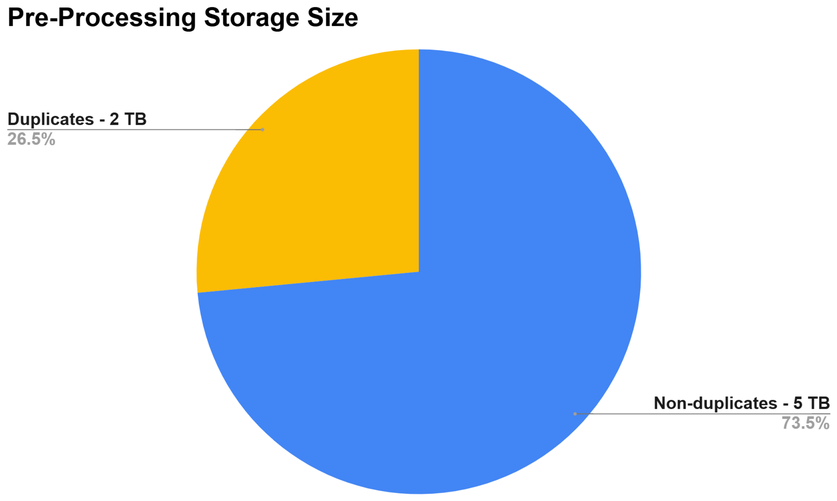
8 TB and 7 Hard Drives: Balancing Original Order and Storage Concerns for Born-Digital Materials

Kate Phillips
North Carolina State University Libraries

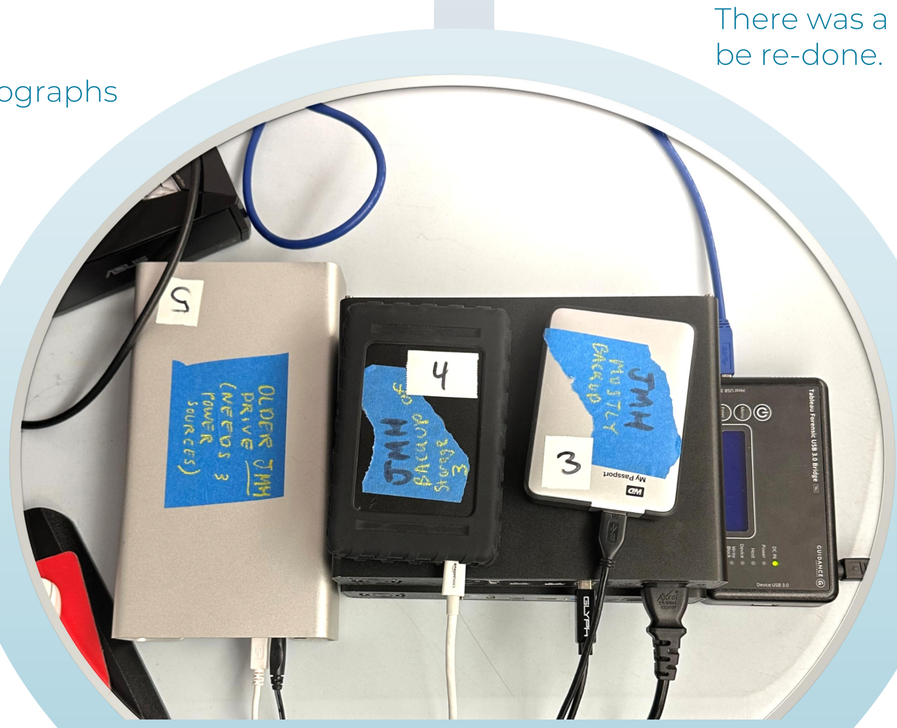
Introduction

The John Mark Hall Photographs Collection at North Carolina State University Special Collections holds photographs taken by the photographer John M. Hall, including seven hard drives that contain backups of his photographs.



Goals and challenges of processing this collection:

- Goals
  - Make it easier to researchers to access photographs on the hard drives
  - Preserve original order where possible
  - Reduce storage size on servers
- Challenges
  - Unclear how many duplicates there were, or how many eventual records
  - Inconsistent labelling across drives
  - Inconsistent organization across drives



Five of the original hard drives; the top three have both the original label and our numbering system visible.

Changes to Workflow

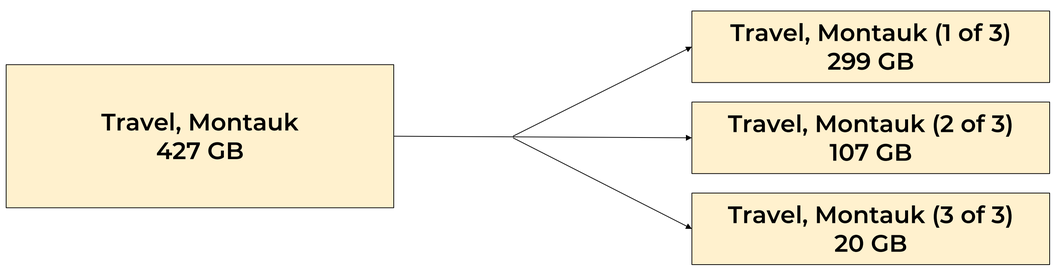
The standard workflow for born-digital materials at NC State Special Collections already includes

- Finding total size of package
- Itemizing file types
- Determining whether there are duplicates

We created a spread sheet inventory to track folders in each drive, which ones had been processed, and within which record they ended up. This helped us keep track of the arrangement that we completed as we processed the drives.

Due to the age of the drives, we also made disk images of each that we could pull files from instead of the physical media. This reduced repetitive steps like adding write blockers, while protecting the integrity of the physical media itself. We added three main steps to the processing workflow:

1. File Transfer – Adding an arrangement step
   - We initially began with top-level folder names, but several drives had large sets of nested folders, and our soft limit was 100 GB per package. Large record names needed to be split for researcher ease of use
   - Records were either split by name when possible or into multiple smaller packages
   - Within the package, folders remained nested within a folder representing each drive.



Changes to Workflow Continued

2. De-duplication
   - After transferring and arranging files, we used the tool Jdupes to create symbolic links for each duplicate and generate a report of where link pointed. This was placed with other processing reports to be packagsed with the digital files.
   - We also modified the step to calculate the size of the package – adding the command “du –sh” – in order to correctly account for the symbolic links.
3. Description
   - We added extra metadata into the finding aid to record the arrangement and processing steps unique to this collection.

Results

At the end of the project, we ended up with an estimated 41% reduction of total storage size for images.

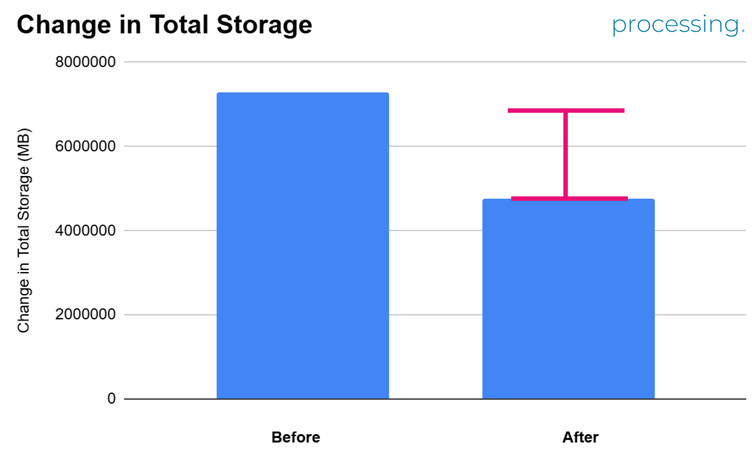
There was a known error rate of 8% for records that needed to be re-done.

There are also 9 known records that we do not have accurate storage data for due to the late discovery that the Brunnhilde tool would count symbolic links as having the same size as the file they link to – rather than the much smaller size of the link itself.

Our reduction estimate has an error of 15%, which keeps the low end of potential sizes at the estimated reduction of 26% total storage based on the original amount of duplicate images.

That’s equivalent to approximately 50 thousand songs

The red bar represents 2518996.112 MB or a 41.89% difference after processing.



Conclusion

The de-duplication process successfully reduced the total storage that the digital collection takes up for archival storage.

The largest change was to add an arrangement step to the workflow; it was time extensive, had a relatively significantly larger error rate, and reduced our ability to maintain original order of the material.

We had to balance maintaining the integrity of the collection while keeping it useable for researchers. De-duplicating made such a large collection more manageable by reducing the size of archival packages available to researchers. Ideally, our thoughtful process and intentional records of the processing and arrangement we completed will be a benefit to the researcher, making it easier to locate records of interest that are related to each other while being able to recreate Hall’s original folder structure.

Moving forward, we now have a tested workflow for large digital collections with large numbers of duplicates as well as confirmation that it does reduce storage and retrieval speeds for records.