Processing Workflow

This workflow (see figure 5.1) is where the majority of archival silences and institutional biases can be brought to light and overcome. As the person responsible for appraising, arranging, and describing these materials, you are responsible for being aware of your personal biases and the perspective you are bringing to the processing of the collection. Involve the creator of the materials, the donor, and the community of origin as much as is possible through direct communication. Have them review the arrangement and description so that unintentional misrepresentations and misunderstandings do not occur due to archivist-created content. Share this power with the creator and community so that the historical record is as inclusive as possible.[1]

Depending on your institutional context, a different team may be processing the collection than those that did the original acquisitioning and accessioning. That is why the documentation generated during those workflows is particularly important. That documentation includes critical contextual clues for processors to follow when they are doing their appraisal and developing their arrangements and descriptions. In most cases, collections that contain digital materials are a hybrid of paper and digital materials. However, as time goes on, we will move toward a situation where most collections will be digital only and the rarities will by hybrid or paper-only collections. Currently, it is rare to have only digital materials, and when this does happen, it is quite often a digitized collection—that is, a digital version of a physical collection created by scanning the original physical materials. In some institutions, digitized collections do not go through the digital preservation workflows because the institution holds the paper originals and considers those originals as the preservation priority. I have found, though, that there are instances where the digitized version is all that you have, or where so much effort and financial resources went into a digitization process that it is a risk management decision to include these files in the digital preservation program.

Hybrid Collection Peculiarities

Hybrid collections that have resided in your institution long enough very likely have already had the physical pieces of the collection processed, including a published description, and are available for researchers to access. In these cases, there may have been no available workflow for processing the digital parts of the collection so there is only a note in the description mentioning that these files exist but are currently inaccessible. In the case where the physical content is already processed, the main decision to make is whether to integrate the digital files into the existing arrangement and description or whether you need a completely new series solely for the digital files. However, if no part of the collection has been processed, you will have the option of creating a plan for the entire collection as a whole, from the beginning.

That being said, in my experience, it is far easier to assess the physical materials and create a processing plan based on the intellectual contents of those materials before ever touching the electronic files. There are some simple reasons for this. The first is that it is much easier and quicker to skim and flick through pieces of paper than it is to access a series of discrete digital files. With paper, all you have to do is turn the page. With digital materials, you have to wait for the software to load the information. No matter how advanced your current computer, there is always a time lag when moving between digital files. Also, paper materials are much easier to lay out and rearrange than digital files. Again, with paper all you have to do is pick it up and move it. With digital files you must copy or move the files and then verify that

the files have not been affected during the process. The larger the file you have to copy or move, the longer the transfer and verification process takes. If you have a general idea of the organization and content of files from the physical records, you will have a much easier time appraising and organizing the digital files.

Develop Processing Plan

A processing plan could be a formal document that describes the steps you will take in appraising, arranging, and describing the materials with a time line for when each step should be complete. Alternatively, your
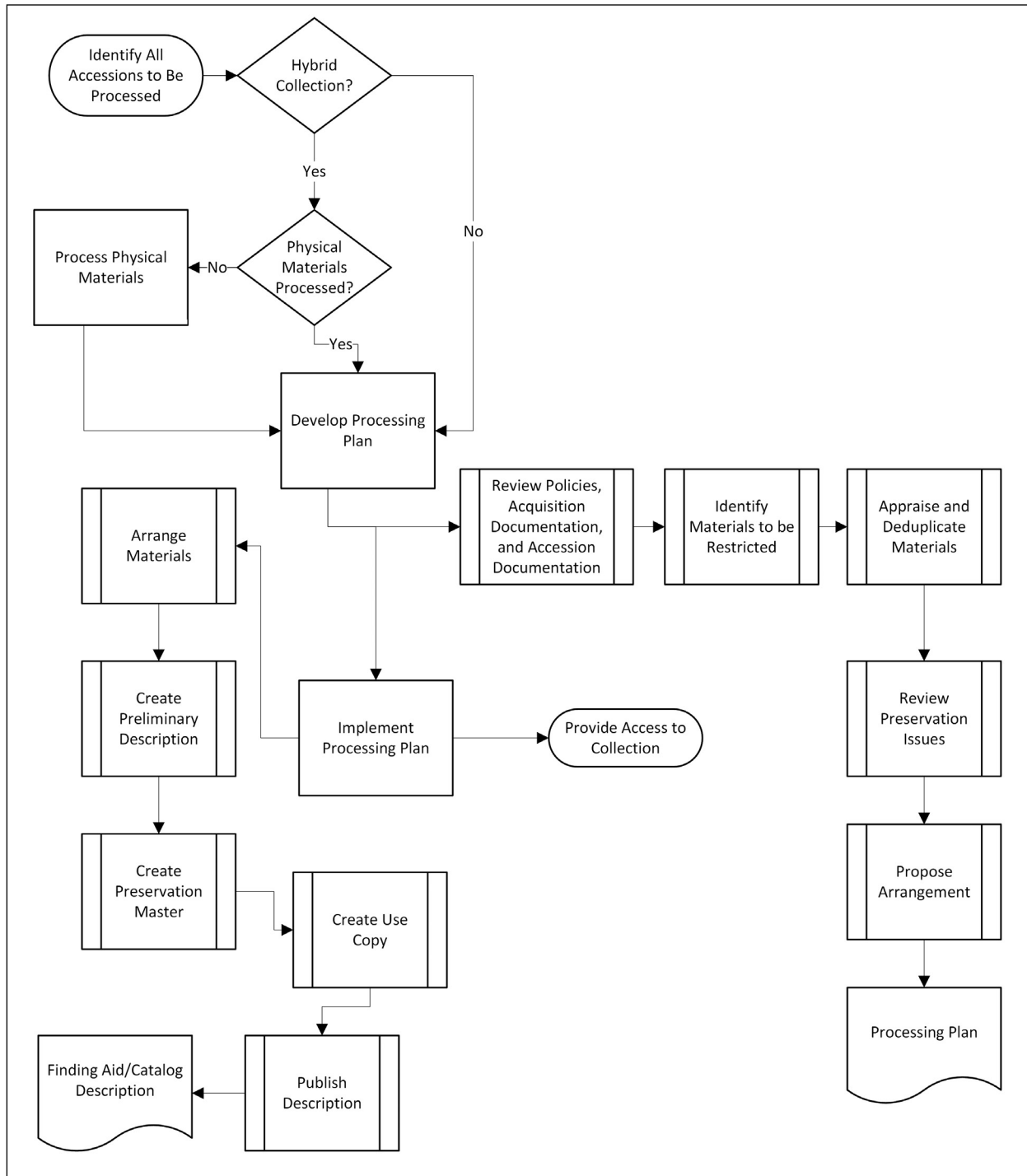

Figure 5.1
Diagram of a high-level processing workflow

processing plan could be an informal set of notes and outlines. This is again dependent on your institutional context. However formal the process, there should be documentation of the decisions you make at each stage and why those decisions were made. This documentation is evidence of the steps you took to make the collection available to researchers and is part of the institutional memory that makes your decisions and justifications transparent to any future archivist.

Review Policies and Donor/Creator Documentation

Before opening a storage box or a digital file, go to the master files for all the accessions that make up the collection. Review the deeds of gift, donor surveys, donor interviews, and any communications with the donor that document your legal obligations in regard to restrictions and what to do with discarded materials, as well as giving a contextual overview of the materials in the collection. Make notes on what to be aware of and where potential private information may be.

Identify Materials to Be Restricted

Using the donor-provided documentation about what materials need to be restricted and where those materials currently live in the collection, if provided, and a standalone tool like Bulk Extractor or the built-in functionality of your digital asset management system, review the flagged materials for potential restriction.[2] Generally, the software tools will flag only personally identifiable information for you to review, items like social security numbers, credit card numbers, phone numbers, and addresses—text that follows a pattern and can be used to steal someone's identity. If the person linked to any of this information is deceased, you often do not have to restrict any of it. Alternatively, if the person is still alive, the individual files that contain this information need to be restricted, redacted, or removed from the collection. For other types of personal information that the donor wants restricted, you are dependent on the donor to give you a map to where this information may be in the collection or distinct keywords to search for in the files to help you find it.

After identifying restricted materials, you have two options. You can immediately remove the material from where it is currently located in the collection and move it to a separate digital folder for restricted material for the entire collection. The second option is to continue with the workflow until you have a proposed arrangement and then restrict the materials in a separate folder that is intellectually associated with where the material belongs in the arrangement. That intellectual association is generally done through the folder name.

Appraise and Deduplicate Materials

If you have a digital asset management system, you will be appraising the files within the system, and the system will automatically remove duplicate files based on the parameters you set when implementing the system. However, if you do not have a digital asset management system that includes processing functionality, there are tools such as TreeSize or WinDirStat that generate a visual overview of the collection and a detailed listing of the types of content included.[3] The visualization breaks the collection out into content types such as video files, audio files, word processing files, and so on. The tools also provide an analysis of how much data and how many files are in the collection, which are key pieces of information for your final description and for determining the best avenues of eventual end user access.

These tools are invaluable during the appraisal process, and some can do double duty of analysis and deduplication. I use TreeSize Professional for this very reason: it allows me to appraise the materials and deduplicate the files in the same step. Using the bird's-eye view of the collection and the more detailed hierarchical view provided by these tools will allow you to determine most of your arrangement without having to review individual digital files. For institutional records, these tools also help you quickly determine if there are personal files that were inadvertently donated alongside the institutional records that were transferred.

Review Preservation Issues

For those institutions that do not have a digital asset management system that will automatically normalize files into standard preservation formats upon transfer into the system, there are several documents created in previous workflows that can help you determine if there are potential preservation issues in the collection. These include the donor survey, the accession report, the more detailed technical documentation generated during the accessioning process, and the collection analysis done in the previous appraisal step. Using all of this information, determine if there are any files your institution does not have the resources to provide end user access to. Document what these files are and determine if they are worth keeping or if they are to be deaccessioned. Be sure to include these decisions in your final description of the collection.

Propose Arrangement

After reviewing all the documentation and the files themselves and making deaccessioning decisions, outline your proposed arrangement. The first decision to make for any type of collection is if you will actually be rearranging the files. For digital-only

collections where the creator-imposed organization is clear enough for users to follow, all you have to do is describe this arrangement in the finding aid. It may also be the case that the collection is so large, regardless of existing organization or lack thereof, that no rearrangement of files is feasible. In this instance, that is what is documented in the finding aid, along with any additional information that can be provided from the initial appraisal steps. For digital-only collections where you are imposing an arrangement, outline the proposed arrangement as you would in a finding aid. Have another archivist review the arrangement, if possible, to see if it makes sense to someone not embedded in the collection, just as you would have a friend review a draft of your journal article before turning it in to a publisher.

For hybrid collections where the physical portion of the collection is already processed, you will need to decide if the digital files will fit well within the existing arrangement or if you need to propose a stand-alone series for the digital files where you can outline an arrangement that better fits the current organization of the digital material or simply describe the creator's organization of their files. For hybrid collections where you are simultaneously processing the physical and digital materials, you will need to determine an arrangement that best fits both sets of materials. Again, you could decide that it would be best to separate out the digital files into their own series, but it is less likely that this will be the case because you are not having to deal with legacy processing decisions.

Implement Processing Plan

Arrange Materials

You should follow your institution's processing workflow for implementing your arrangement on the physical materials. If you have decided to impose a new arrangement on your digital materials, I suggest creating the new folder hierarchy in a staging location first and then moving the digital materials into the folders. That way you can start and stop the process as needed, and you are less likely to accidently delete files or alternatively copy files into multiple unintended locations. Also, if your institution does not have a digital asset management system that will automatically sanitize filenames (remove special characters) or normalize the file formats into a standard preservation format, you will need to do this as you transfer materials into their final arrangement. If you have a digital asset management system, those steps are most often taken care of when the files are transferred into the system.

The implementation process I use is as follows:

- Move files into a folder structure that mirrors the layout of the finding aid, using file transfer software that verifies the files were not changed during the move, such as TeraCopy.[4] For example:
  - Mss###_CollectionTitle
    - Series_I_Personal
      - Subseries_1_Finances
        - Put all the files and folders that belong in the Finances subseries into this folder.
- If there are restricted materials as part of the collection, create a "RESTRICTED_ Mss###_CollectionTitle" folder hierarchy that mirrors the finding aid. Have that folder hierarchy's access limited by username to the head of the archives, the digital archivist, and the processing archivist.
- If there are files to be normalized, save the new versions of the file into the destination folder instead of moving the original file.
- Using a file renaming protocol, such as ReNamer, sanitize filenames in the folder.[5]

After the files have been arranged, delete your working files copy so there is no confusion over what version of the collection to carry forward into the rest of the workflows.

Create Preliminary Description for Materials

After you have arranged the materials, draft the finding aid text relating to the digital materials. There are guidelines for this in *Describing Archives: A Content Standard (DACS).*[6] Part of creating this draft will be deciding if you will include direct links from the finding aid to the digital materials or if users will have to request access. It is not an all-or-nothing decision. It could be that there are direct links to some of the materials in the collection, while other require mediated access. Have at least one other person review the description, preferably someone who was not involved in the processing of the collection, for readability and usability. Ideally, the donor or creator would also be able to review the draft description before it is published. If that is not possible, I would recommend that part of the engagement with the donor include an emphasis on the fact that they have the ability to request that changes be made to the description as needed.

Create Preservation Master

After you are completely satisfied with your arrangement, create a preservation master of the complete collection. This could be done automatically through your digital asset management system. Alternatively, this could be the point where you transfer the materials to a system such as Archivematica.[7] I have found this system works best on a fully arranged collection; it will automate the process of creating a preservation

master copy and a use copy. Your preservation master could simply be a copy of the fully arranged files placed in your dark archive with their administrative and technical metadata generated during accessioning and stabilization.

Create Use Copy

With the preservation master carefully tucked away, you are now ready to create your use copy. The use copy of the collection is what you provide to your researchers. The major difference between the preservation master and the use copy is for content such as videos, audio files, and images. The file types will be different, and the file sizes will be smaller. For example, the preservation master of an audio file could be a WAV at close to a 500 megabytes. The use copy of that same audio file would be an MP3 at close to 160 megabytes. Generally, if you do not have a system to automate the creation of use copies, you would focus on creating use copies only of very large files that would be difficult for users to access over the web because the bandwidth needed to stream or download them is beyond what most researchers at home reliably have access to.

Integrate Description into Finding Aid/Catalog Description

Only after you have created the use copy for the collection should you create or modify the finding aid.

In this way, if you are creating direct links from the finding aid to the digital materials, you will have to do so only once. This is incredibly important if you are hand coding your finding aid versus using a tool such as ArchivesSpace.[8] Either way, having drafted your description already, it should be a matter of copying that draft into the tool you use to generate finding aids for the final published document.

Notes

1. Archives for Black Lives home page, https://archivesforblacklives.wordpress.com/.
2. Simson Garfinkel, "bulk_extractor," GitHub, https://github.com/simsong/bulk_extractor.
3. "TreeSize," JAM Software, https://www.jam-software.com/treesize/; WinDirStat home page, last updated November 12, 2018, https://windirstat.net/.
4. "TeraCopy for Windows," Code Sector, https://www.codesector.com/teracopy.
5. "ReNamer," den4b, https://www.den4b.com/products/renamer.
6. Society of American Archivists, Describing Archives: A Content Standard (DACS) (Chicago: Society of American Archivists, 2004, 2013), https://www2.archivists.org/groups/technical-subcommittee-on-describing-archives-a-content-standard-dacs/describing-archives-a-content-standard-dacs-second-.
7. Archivematica home page, https://www.archivematica.org/en/.
8. ArchivesSpace home page, https://archivesspace.org/.