Processing History Note: Born Digital Collection Material[1]

The following text provides an overview of the standard processes the Manuscript Division of the Library of Congress undertakes when preparing born digital files for preservation and access.

Born digital files are those that were created in digital form and, unlike digital files from digitization workflows, were not created as surrogates for physical materials. They include everything from Microsoft Word documents to Outlook email archives. These records require specialized equipment and processing to ensure that they remain authentic and accessible through the inevitable technological shifts that occur over time.

Born digital files acquired by the Manuscript Division undergo a number of standard processes in order to prepare them for preservation and access. These processes are based on industry best practices to ensure long-term preservation and authenticity of the records, laying the groundwork to ensure persistent, reliable access to content. The following steps are taken on all born digital records, but some collections may require further action beyond what is listed here. Information on those actions may be found in the finding aid's Processing History note when applicable.

Secure transfer

Born digital materials are acquired in a number of ways, depending on where the digital files are being stored or kept. Regardless of original location, steps are taken to ensure that no data is lost during transfer -- usually by generating checksum values and an inventory for the files before and after they are moved, which allows the library to validate that the files are complete and unchanged.

Inventory

When a collection contains digital storage media (3.5" floppy disks, optical discs, etc.), each piece of media is photographed, inventoried, and assigned a unique digital ID#. Researchers should use the digital ID# listed in the finding aid to request access copies of the files associated with each media.

Checksum Creation and Validation

Ensuring the authenticity of born digital files over time is of utmost importance to long-term digital preservation programs. It is easy to assume that born digital files are safe and stable when stored on a server, but in reality the data that makes up these files can degrade over time in a process called bit-rot. Part of ensuring the authenticity and integrity of these records includes creating a checksum -- a numerical value that allows us to validate a file over time to see if content has changed. Manuscript Division staff use a variety of tools to generate an inventory and MD5 checksums for all files at the beginning and upon completion of digital processing procedures and then to validate the checksum during transfer to long-term storage.

---

[1] Adapted from Princeton University's *Born-Digital Processing Information Note*, *https://library.princeton.edu/special-collections/workflows/born-digital/adapt-born-digital-policies*

Disk Imaging
In order to rescue data from corrupt media or access obsolete media (e.g. 5.25" floppy disks), staff may create a disk image. Disk images may be created if Manuscript Division staff are unable to safely copy files from any type of storage media. A disk image is a bit-for-bit copy that preserves not only the files (i.e. allocated space), but also the unused space (i.e. unallocated space), which can include deleted files and file slack. These images are generally not retained long-term, unless there is a preservation or access reason for doing so. For example, optical media may be captured at the bit-level and retained as disk images to maintain the relationship between files of multimedia presentations. In most cases, however, the disk image's content is exported using specialized software and the disk image is not retained long-term.

Virus Scan
All records are scanned for viruses on transfer to the processing workstation and again before transfer to long-term server storage. Files that are found to contain viruses are quarantined and evaluated to determine further action.

Identify and Extract Archive Files
Content that is wrapped in archive files like .zip and .tar are unzipped before any file analysis reporting and content appraisal. The unzipped files are saved to a folder named with the original file name followed by _UNZIPPED. The original .zip file is always retained long-term.

Identify File Extension Mismatches
All incoming files are scanned to locate and identify missing or incorrect file extensions. Because missing and incorrect extensions can prevent access to content, an effort is made with the help of various file format identification tools to identify and document the correct file formats.

Personally Identifiable Information (PII) Scan
PII, or Personally Identifiable Information, refers to data such as Social Security Numbers, credit card numbers, bank account numbers, healthcare and medical information, and other highly sensitive data that could be used to identify an individual. All incoming born digital content is scanned to identify as much of this information as possible. When this information can be identified and located, the content may be redacted, placed under an appropriate level of restriction if redaction is not possible, or deleted if not of historical value.

Identify Duplicates and Empty Directories
Duplicate files take up space and in large numbers can encumber and confuse research efforts. When duplicate files and empty directories are located, they may be appraised for potential removal. In general, the Manuscript Division staff will remove duplicates at the media or folder level, rather than the file level.

Arrangement
Whenever feasible, the original file structure and file names are maintained as received.