Processing History Note: Born Digital Collection Material[1]

The following text provides a detailed explanation of all 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 processes to ensure that they remain authentic and accessible through the inevitable technological shifts that occur over time.

Born digital files acquired by MSS 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 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 than what is listed here. Information on those actions may be found in Processing History Note in finding aids where they apply.

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 storage media (3.5" floppy disks, optical disks, etc.), each media is photographed, inventoried, and assigned a unique digital ID#. Use the digital ID# listed in the finding aid to request access copies to 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 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. MSS generates MD5 checksums upon completion of digital processing procedures and validates 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*

Legacy Media and Optical Media Captures

Some external storage media (e.g., 5.25 inch floppies) are captured at a bit-for-bit level by creating a disk image. This preserves not only the files, but also any information about the disk from which they were captured. These images are generally not retained long-term, unless there is a preservation or access reason for doing so. Optical media may be captured at the bit-level when there is added value in doing so (e.g., maintaining the relationship between files of multimedia presentations), but otherwise content is transferred using standard MSS file transfer processes and specialized software, primarily Bagger.

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.

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 the 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 or placed under an appropriate level of restriction if redaction is not possible.

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. MSS will remove duplicates on a media or folder level, rather than the file level.

Arrangement

As much as is feasible, the original file structure and file names are maintained as received.