Processing Information Note: Born-Digital Records

Born-digital records are those that were originally created as a digital or electronic file: anything that was created on a computer and not transferred from an analog version. They include everything from Microsoft Word documents to Outlook email archives. These records require specialized equipment and procedures to ensure that they remain authentic and accessible through the inevitable technological shifts that occur over time.

Born-digital records acquired by PUL undergo a number of standard procedures in order to prepare them for preservation and access. These procedures 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 Information notes in finding aids where they apply.

Secure transfer

Born-digital materials are acquired in a number of ways, depending on where the records 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 for the records before and after they are moved, which allows the library to validate that the files are complete.

Legacy Media and Optical Media Captures

All legacy media (3.5 inch floppies, 5.25 inch floppies, zip disks, etc.) 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. Optical media may be captured at the bit-level when there is added value in doing so, but otherwise content is transferred using standard PUL file transfer procedures, including scripts and other specialized software like Exact Audio Copy.

Virus Scan

All records are scanned for viruses before transfer to the processing workstation. 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 is hidden from PII scans, file extension mismatch scans, and correct data size estimation. These files are identified and extracted using the DROID[1] file profiling tool in order to be appropriately run through the rest of the born-digital processing workflow.

---

1 http://www.nationalarchives.gov.uk/information-management/manage-information/policy-process/digital-continuity/file-profiling-tool-droid/

Identify File Extension Mismatches
All incoming files are scanned to locate and identify missing or incorrect file extensions. Because missing and incorrect extensions prevent access to content, an effort is made with the help of the PRONOM registry[2] to identify and apply correct extensions to affected files.

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 records are scanned to locate and identify as much of this information as possible. This 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. Duplicate files and empty directories are located and can be appraised for potential removal.

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[3]. 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. PUL generates SHA256 checksums upon completion of digital processing procedures and before ingest into long-term storage.

Tools and Workflows
For more information about the tools that are used in born-digital processing procedures, please view the Glossary for Born-Digital Processing. For a more step-by-step picture of how born-digital materials are processed, please visit the PUL Born-Digital Workflows page.

[2] http://www.nationalarchives.gov.uk/PRONOM/Default.aspx
[3] https://en.wikipedia.org/wiki/Data_degradation