

IFP Digital Materials Weeding Policy

(9/17/2014)

In accordance with the Transfer Agreement and Deed of Gift, the complete copy of the original transfer will be preserved in the Long-Term Archive in a form of a .RAR compressed archive and will never be deaccessioned. However, not all of the digital assets will be subject to long-term preservation.

The following file categories will not be ingested into the long-term archive:

1. Confidential, transitional and sensitive materials (e.g. copies of passports and visas, insurance and banking information)
2. Disk setup and fileinventory files, except for fileInventory.xml
3. System files (e.g. .32, .db, .exe, .lnk, .rel, .rels, .ini, .inf, .utf8)
4. Files related to web-pages (e.g. .js, .htm, .html, .css, .ico, .php, .thm)
5. Files associated with applications that are not actual documents (e.g. .p65, .tmp, .xml, .wbk)
6. Autorecovery temporary files (filenames starting with “~\$”)
7. Database-related files: lock (.ldb) and login (.exe)
8. Files that fail to open (due to unknown file type, password-protection, corruption, etc.)
9. Duplicate files (files with identical checksums): if duplicate files are discovered, the following rules apply:
 - Files located in directoryName_zip and directoryName_rar subdirectories within their original parent directory are considered duplicate and should be deleted
 - If duplicate files are located in different directories, weeding is performed on a case-by-case basis (e.g. such files found in “Cohort 2006” and “Cohort 2007” directories are not considered duplicates since they reflect different context)
 - Files named “Copy of <filename>” are considered duplicate and should be deleted
 - Files located in Restricted directory can be deleted if they have duplicates in Unrestricted directory.
 - For audiovisual materials that are identified as duplicates in terms of their content, but exist in multiple formats, the largest file should be taken for transcoding to preservation format. The other copies should be deleted.