

Procedures for Accessioning Electronic Records

June 7, 2013

Before any records come to the Archives, they should be appraised while in the office and a retention schedule created. The retention schedule will indicate the length of time the records will exist after creation before either being destroyed or transferred to the Archives. For some electronic records, this length of time will represent when snapshots of a record are taken and transferred, while the active record is still being updated by the agency. This means that over time the Archives will need to manage different versions of the same record and maintain the contextual relationships between them.

Organize Files in Office

Agencies should be taught proper files management techniques for maintaining their electronic records in such a way as to facilitate their local use as well as a future transfer transaction. Electronic records are best managed within the context of record series and retention schedules. Sets of electronic records should either be contained within an automated electronic recordkeeping system that will apply retention schedule metadata to the files, or within a file structure within a local server where folders contain files by series. This way, files that have met their disposition date can be easily sorted.

If files are to be transferred, the agency should separate these records from those that have not met retention. An agency may choose to keep a local copy or not. The agency may also choose to transform the original version to a normalized format prior to transfer. The format management rules identified as part of the retention schedule should identify what is being transferred, including multiple representations (Word, PDF) of the same record. For validation purposes, it is preferable for the agency to send us both the original format and its transformed version. The pattern being used to maintain relationship links between original and normalized versions should be documented. It could be a case where the file names are the same, stored in the same folder, but with different extensions; or each version is stored in separate folders and folder names provide context.

Prepare Records for Transfer

Just as an agency is expected to prepare its paper records for transfer to the Archives by purchasing the appropriate boxes, labeling them, and documenting the contents through a Records Transfer Sheet (RTS), similar steps are required for electronic records.

1. Before an agency moves any files out of its environment, they should run a virus check against the records. In their transfer paperwork, they should express how and when this virus check took place, and any remediation that took place if a virus was found. Agencies that send us

records which contain a virus will have their transfer rejected and sent back to them. They may resubmit the transfer after the records are clean.

- a. An exception would be for defunct agencies that are sending us their records. In that instance, the Archives will need to resolve any viruses found.
2. Ideally, a checksum of the records as they exist while still at the agency should be taken. Multiple tools are available for this function, but BagIt is our tool of choice. It was created specifically to support the transfer process of records from one institution to another, and maintains the internal integrity of the records being transferred. The output from BagIt is also natively recognized by our AXAEM ingest tool, so records can be validated upon receipt.
 - a. BagIt is a tool designed to be installed on the desktop and run on the command line. Extensive documentation for using BagIt is available from North Carolina: <http://www.records.ncdcr.gov/erecords/>.
 - b. Agencies should be trained on the use of BagIt prior to transferring records.
 - c. While a bag could contain an almost infinite number of records, a better practice is for bags to contain a reasonable number of records, especially so that a complete bag can be stored as a unit on the transfer media (required for validation). During the ingest process, the bag will be moved around and stored in different locations, including M-disc, which has a limitation of 4 GB, so smaller bags are better if possible. Some individual records, of course, such as databases or video may exceed 4 GB. These should still be bagged, but will be managed differently upon ingest.
 - i. For ingest purposes, it is best that a bag contain only one series at a time. Multiple series could be stored on the transfer media, however.
3. After an agency's records have been "bagged", they should either be written to media such as a flash drive, external hard drive, or DVD, or the agency should make arrangement to upload their bag to our SFTP server. The agency should be warned that large data transfers that take place over a network connection can take a very long time to complete.
4. The agency should complete the RTS and submit the transfer request as normal. Transfer sheets for electronic records should probably be modified to allow the following:
 - a. List each series contained within the transfer
 - b. Number of files within the transfer (lots of tools can help provide this number automatically)
 - c. Certification that files have been virus checked (when, tool used, results)
 - d. Format types, if known
 - e. Description of procedures done to normalize any file formats, including date normalized, software used, versions from and to, etc.
 - f. Any documentation such as file layouts that will help us understand and interpret the records, which is especially important with databases.
5. Records that have been transferred to the Archives, including all transfer documentation are now known as the Submission Information Package, or SIP

Accession the SIPs

6. Once received, the Archives should run its own virus check against the SIP. Any viruses found will be quarantined and the agency informed.
7. The Archives will validate the files in the SIP by comparing the Records Transfer Sheet to the bagged records. Any anomalies found will be communicated to the agency. Such anomalies might include records of a different series than what was declared on the RTS, records of a different date span, or missing elements required to render or understand the record.
8. A Transfer record should be added in AXAEM with the RTS attached.
9. If the SIP contains a BagIt bag, it will then be run through the VerifyValid process using the BagIt tool. AXAEM can be used to run the command as a graphical interface provided BagIt is installed on your desktop. The files do not need to be ingested onto the server to run this.
10. If digital forensics software is available, this would be the stage at which it is used. The purpose of using digital forensics is to get a good bird's eye view of what is contained on the media, and run statistics of file sizes, format types, folder layouts, etc.
11. File validation will then be run through both automated processes and manual processes. The automated processes include using tools such as FITS, Droid, JHOVE2, and others.
 - a. These tools can be run on the desktop, or alternatively the SIP can be ingested into AXAEM and the user can pick which tool to run as part of that process. Note that the place to ingest a SIP vs. an AIP in AXAEM is on a different menu accessed via the Electronic Records Menu (look for the Manage SIPs link).
 - i. If ingesting a SIP into AXAEM, be sure and include the Transfer key as part of the information requested on the ingest screen. AXAEM will then recreate the file structure of the SIP and record metadata. While not a current feature, we could also provide a button for AXAEM to keep the metadata of the SIP, but delete the physical files if those are better stored offline.
 - b. The output from automatic tools will produce a best guess as to the format of the record based upon some internal clues inside the file as well as the file extension. Often these tools will disagree with each other, make tentative decisions, or miss some obvious and common file types. Their format decisions will be expressed as a Pronom ID, aka PUID. The Pronom database, which was created by the National Archives in the UK, maintains a list of all formats and provides a unique identifier, e.g. fmt/18 for PDF 1.4 files, or x-fmt/88 for a PowerPoint presentation. The PUIDs will be recorded as metadata within AXAEM. To update the accuracy of these tools, Pronom publishes something called a "signature file" which has instructions for identifying various file types. Some tools use older signature files, and so their conclusions are not as accurate as those that use later signature files. The version of the tools and the version of the signature files operate independently of each other. An older tool can use newer signature files, and a newer tool can be told to use older signature files (using Droid 6.0 with a signature file above 45 causes it to hang, even though the latest signature files now are in the upper 60s). Often these tools can be configured to automatically download and use the latest signature file. Parameter settings can also impact how the

tools work (Droid 6.1 can be made to run a simplified output that doesn't have as much detail but also doesn't hang, yet it uses the latest signature files). FITS and JHOVE2 by default use Droid version 3.0 and a signature file somewhere around 30.

- i. As an archivist, it is up to you to reconcile the decisions made by the tools and record the correct format in AXAEM. The original XML output produced by these tools is attached to the database so it can be viewed and acted upon.
 - ii. If the format cannot be determined automatically, then the file should be opened using software appropriate for its extension and looked at by the archivist. If the format extension is unknown, try Notepad first.
 - iii. Files should always be at least spot-checked manually to make sure they open and render properly. Perhaps these could be opened at a rate of one document per folder, provided a folder contains many similar files of the same file type. If the agency transformed the format from an original version to a new file type, then more frequent spot-checking is required to make sure the normalized files contain the intended information and do not have missing elements such as fonts, watermarks, or graphics. The original and the transformed version would need to be compared side-by-side if possible.
 - iv. The purpose behind correctly identifying file types is so that the preservation rules we develop for a given format can kick in. If we decide that every time a WordPerfect file is ingested it gets transformed to Open Office, then we need to correctly identify when something is a WordPerfect file so we can run an update process against all such files in the database and transform them at the same time. This likely would happen years in the future when we need to migrate a whole bunch of records to a newer version all at the same time (then validate the results, yikes!).
12. After the files have been virus checked, validated, and the checksum verified, they should be stored on M-disc to await further processing. The M-disc should be accessioned into Versatile and given a barcode. Each series contained on the transfer media should be stored on its own M-disc, which could be a waste of M-discs if the series only contains one small file, but it's easier to manage both physically and intellectually if they are separate.

Prepare the AIP

13. When processing begins, a copy of the files on M-disc is made and stored on a local drive (internal or external) available to the processor's PC. The processor will then arrange the records in an accessible way and adjust folder names and file names to best describe the contents of the files. This includes adding missing extensions or changing wrong ones where necessary.
- a. **Warning!** Changing file names could have unexpected consequences if there are file dependencies (one file that calls or references another either through direct links where one launches the other, or by naming it inside of text as a footnote or something).
 - b. **Warning!** Sometimes folder names have extensions that have meaning to software, as is the case with geospatial file geodatabases. Do not rename folders of these types.

- c. Any time you make any changes to records, it is important that the change and the reasons be documented in notes. A generic note could be added at the series level, and more specific notes added at the item level.
- 14. After records have been arranged, it is time to ingest them into AXAEM as the AIP. Go to the Electronic Records Menu | Data Ingest tab, and run Desktop File and Folder Ingest. If the set of files being ingested is expected to take a very long time, you may optionally walk a copy of the files up to the Capitol and have DTS load them onto a temporary location on the server, then you can ingest them from the server directory, which will shorten the ingest process considerably.
- 15. AXAEM will display a directory tree of all the files you indicated from the folder location you chose. At this point you can choose to ingest everything or just some things. Uncheck the checkboxes next to any files you do not want to upload. If the number of files being ingested is large, note that it could take AXAEM a little bit of time to display the initial directory tree because it is building that whole list into memory.
- 16. The ingest screen gives you several options of different tools to run. You may do another virus check at this point or not. You may choose to run a metadata extractor tool or not. Using a metadata extractor lengthens the time of ingest. There is also a way to run these extractors after the files have been ingested if you want to separate those steps. If your batch includes a Bagit bag or other container files (e.g. .zip), then other choices will pop up on screen that allow you to save the container as a separate object from its contents.
 - a. If you have already ingested the SIP into AXAEM, be sure to include the SIP Batch ID as metadata to the AIP, so that the two will forever be linked in the database and you can compare “before” and “after” processing more easily.
- 17. When records are ingested, the following file structure is created:
 - a. Object Group (generally a description of the folder, although it could be a subset of a folder)
 - i. Electronic Record (generally a description of a single intellectual item, which may contain one or more physical files but they are always treated as a unit, e.g. geospatial shapefile)
 - 1. Representation (a specific instance of the Electronic Record, e.g. .tiff vs. .jpg)
 - a. File (the actual physical file(s) tied to the representation)
- 18. Different metadata fields are available for editing at the different description levels. For instance, at the Representation level, about the only thing you can identify is whether the representation is the preservation copy or access copy. At the file level, you can see the extent of the one file, its format, and things related to its physical characteristics. But at the object group or electronic record level, you can describe the whole purpose of the record.
- 19. To allow for maximum flexibility of preserving a record over time, the original record, in whatever format it is, should be preserved just as is. Another copy could be normalized and migrated forward in time to be renderable with the latest software improvements.
- 20. At some point we will add a feature that lets you export the AIP as an object with an XML wrapper. This is done by converting the file into its base64 equivalent, and then adding METS

metadata around it. All of the unique identifiers, technical metadata, and descriptive information would need to already be in place in the database so the METS export would have data to pull from and include in the wrapper. The METS object could then be stored locally as an additional security copy or distributed to other locations for safekeeping. If metadata is changed in the database, the METS object would need to be recreated.

Create the DIP

21. After the AIP is described, you may create the Dissemination Information Package (DIP), aka the access copy. The AIP and DIP are always separate physical copies. For some file types, the original format will be used as the DIP. For others, the files are better transformed to a normalized format that is easy to migrate forward, especially one that uses widely-accepted software that the public generally knows about. It can be posted to a web server and linked to the intellectual content of the AIP, or written to M-disc and made available for local use in the Research Room. Non-public records could also have DIPs created, but they would need to be stored securely.
 - a. If files are to be normalized, often the best place to do this is on the desktop. You will always get better results when transforming a Word document to a PDF using Word and Acrobat, than you would uploading the Word file, then using AXAEM's feature of creating the PDF by first converting the Word file to Open Office (which will mangle fonts and graphics), then saving the Open Office version to PDF. On the other hand, transforming images from one format to another works pretty well using the ImageMagick tool. So transforming a .tiff to a .jpg might be done easier on the server. Having the server do it means that you don't have to upload individual DIPs and then manually create the link from the AIP to the DIP.
 - b. If an access copy has already been created by another institution and is perpetually available at the same URL, then you may link to that within the AIP metadata.

When Actively Acquiring Records

In some cases, records may come to the Archives not because an agency transferred them to us, but because we went out and harvested them, especially if we take content directly off an agency's web server without their knowledge.

22. To harvest a web server's contents, you can use a tool such as HTTrack, which runs on the desktop, or you can use a Linux server command line command such as wget. See <http://en.wikipedia.org/wiki/Wget> for further information. We could also use Archive-IT, since we share a license with State Library. Soon an interface directly in AXAEM will allow you to run the wget command without actually needing to remember all the parameters of that command. The resulting files would be stored on the server using the same storage device where AIPs are stored. AXAEM could then bag the contents of the harvest using BagIT, and provide a link to download a copy of the bag to the desktop for further management of the SIP.

23. Since the records would not be coming in with a Records Transfer Sheet, one should be created by the archivist. Additionally, a Transfer record should be created in AXAEM that identifies the contents of the transfer and other details about how the acquisition occurred.
24. The SIP should then be written to M-disc and follow the same processing and ingest procedures that are written above.