July 29, 2010

Final Report: Functions and Activities for Acquiring, Preserving and Making Accessible Electronic Records

Harrison W. Inefuku

National Library of Medicine

Archives and Modern Manuscripts Program

Functions and Activities

Acquiring, Preserving and Making Accessible Electronic Records

Prepared by Harrison W. Inefuku
July 2010

Final, v. 1.0 (29 July 2010)

Table of Contents

1	Introduction
3	A. Communicate with Records Creator
5	B. Transfer Electronic Records
7	C. Appraise Records
10	D. Arrange and Describe Electronic Records
12	E. Place Electronic Records in Storage
13	F. Preserve Records
14	G. Provide Access to Records
16	Appendix A: Metadata
19	Appendix B: Preservation Formats

Functions and Activities for Acquiring, Preserving and Making Accessible Electronic Records

Introduction

The purpose of this report is to identify activities necessary to acquire, preserve and make accessible electronic records donated to the National Library of Medicine's Archives and Modern Manuscripts Program (AMMP).

Because the focus of the electronic records project is on private records acquired by the National Library of Medicine, emphasis is placed on pre-ingest activities that seek to determine the context and nature of the records. This report advocates a preventative approach to electronic records preservation—by collecting as much information about electronic records as possible, as soon as possible, so preservation plans can be developed before time and technological change renders such records inaccessible.

Because AMMP primarily acquires personal papers, the functions and activities presented in this report emphasize the capturing of information about the records' contexts (juridical-administrative, provenancial, procedural, documentary and technological). Gathering this information will be beneficial if the types of file formats acquired by AMMP become more complex and/or dynamic (i.e., if AMMP begins to acquire data sets, databases, visualization models, etc.).

The process of preserving and disseminating electronic records is broken down into seven functions, which consist of a number of specific activities. Each function is analyzed, with a list of inputs and outputs, a list of tools that may be required to complete each function and any explanatory notes (these notes usually explain activities that are not currently undertaken by AMMP, or, indicate where policy needs to be further developed).

The functional model presented here draws from a number of sources, including:

- InterPARES *Chain of Preservation Model* and *Preserver Guidelines*
- *Open Archival Information System (OAIS)* (ISO

- Paradigm *Workbook on Digital Private Papers*
- Simon Fraser University *eRecords Research Report 2*
- Tufts/Yale *Fedora and the Preservation of University Records Project*

Functions

Seven functions in the preservation and dissemination of electronic records are presented in this report:

1. Communicate with Records Creator
2. Transfer Electronic Records
3. Appraise Electronic Records
4. Arrange and Describe Electronic Records
5. Place Electronic Records in Storage
6. Preserve Electronic Records
7. Provide Access to Electronic Records

Note that the activities presented in this report are grouped by function, not by chronological order. Thus, activities do not need to be performed sequentially from A1 to G8, but can be performed in the order most appropriate to the electronic records being acquired by the Archives. The Appraisal function (C), for example, will likely be performed at several different times, with each appraisal focusing on a different aspect of the electronic records.

A. Communicate with Records Creator

This function includes all communications with the Records Creator from the point of initial contact to the physical transfer of records to the Archives. It gathers information about the Records Creator and the Electronic Records for appraisal, arrangement and description.

Activities

1. Contact initiated with Records Creator
    a. Determine scope of fonds, including both electronic records and analog records
2. Determine suitability of Archives to acquire records[1]
3. Pre-visit made to Records Creator
    a. Identify records of interest to Archives
    b. Initial inventory
        - Determine preliminary arrangement and extent of fonds[2]
        - Record storage media types and quantity[3]
    c. Interview Records Creator[4]
        - Determine types of electronic records created, context of electronic records[5]
    d. Conduct preliminary appraisal of records[6]
4. Schedule pick-up of records

---

[1] i.e., Does the Creator's records fit into the Archives' mandate and scope? If no, refer Creator to more appropriate repository.
[2] Conduct a high-level overview to identify potential series structure, etc.
[3] In part, this identification is to determine if the Archives has the capability to accept transfer of, and, preserve electronic records. Important facts to note: (1) operating system used; (2) software applications used frequently; (3) approximate age of electronic records; (4) presence of obsolete storage media. Conducting the inventory can be as simple as making screen captures of the Creator's file structure and records.
[4] If the Archives expects a continuing relationship with the Records Creator, it might be useful to provide the Records Creator with advice on the maintenance of electronic records. Otherwise, this interview should be used to gather information about the Records Creator and the electronic records' contexts.
[5] Obtaining information about the context of electronic records will be useful during appraisal, arrangement and description, especially if the file structure of the Records Creator's electronic records to not mirror the file structure for the Records Creator's paper records.
[6] Based on information about the electronic records generated in the inventory and interview, the Archives should conduct a preliminary assessment on the feasibility of preservation for the electronic records in question. These assessments will be confirmed/adjusted during the Appraisal function following the transfer of records from the Records Creator to the Archives.

a. Prepare Deed of Gift
    - Transfer of physical property and intellectual rights
    - Negotiate restrictions on access

Inputs

- Information about Records Creator
- Information about Records
- Information about Archives
- Records

Outputs

- Analysis of electronic records
- Analysis of Records Creator
- Preliminary inventory of records
- Interview with Records Creator
- Preliminary appraisal decisions
- Records Access Policy
- Deed of Gift
- Records

Tools

- Interview instrument
- Survey form
- Inventory form
- Deed of Gift form
- Advice sheets

Notes

The period of communication with the Records Creator is the ideal time to gather information needed to develop preservation plans for electronic records. The Records Creator serves as a valuable primary source to determine the technological context of the electronic records and should be able to provide information on

B. Transfer Electronic Records

This function transfers physical and legal custody of the electronic records from the Records Creator to the Archives. The electronic records are quarantined to ensure they are free from viruses and other malware, and, to ensure that their contents have been verified.

Activities

1. Register transfer
    a. Confirm receipt of records
2. Copy electronic records from storage media onto Archives server
3. Quarantine records[7]
    a. Confirm authorization for the transfer
    b. Confirm authenticity of electronic records in transfer
    c. Verify content of electronic records in transfer[8]
    d. Scan for viruses and other malware
    e. Remove prohibited characters from file names
4. Monitor performance of transfer system[9]

Inputs

- Electronic records transfer

Outputs

- Notification of receipt of transfer
- Results of content verification and virus scan
- Log of changes to file names
- Records to be ingested
- Information for preservation

Tools

[7] Electronic records should be quarantined for a sufficient amount of time to ensure that electronic records are checked against virus and malware definitions that cover the time period up until point of transfer.
[8] This can be done by applying hash values to the Records on the transfer medium, then verifying hash values on archival server.
[9] Ensure that the transfer system is up-to-date and working properly. Ensure that virus definitions are kept updated.

- File transfer software
- Anti-virus software, with registry updates
- Md5sum Scripts, *checksum verification scripts*
- Hashdeep, *compute, match and audit hashsets during transfers*

Notes

Because of the fragility of digital storage media, it is recommended that electronic records transferred on all storage media be copied on the Archives server.

Ideally, the electronic records transferred to the archives would first be moved to a stand-alone workstation—thus, if the records contain and viruses or other malware, they will remain separate from NLM's main storage network.

C. Appraise Electronic Records

This function determines the capability of the Archives to effectively manage, store and preserve the electronic records at question. It also determines the archival value of the records.

Activities

1. Conduct Records Survey
    a. Identify electronic records':
        - Principle creator(s)
        - Directory and folder names
        - File Format
        - Extent of materials
        - Approximate dates of creation
        - Subject areas covered
        - Presence of any passwords/encryption
    b. Analyze information about electronic records' context (juridical-administrative, provenancial, procedural, documentary and technological)
2. Assess authenticity of electronic records
3. Determine continuing value of electronic records
4. Determine record elements and digital components of records to preserve[10]
5. Assess file formats of electronic records
    a. Identify file formats of electronic records
    b. Assess feasibility of preservation
    c. Determine preservation and access formats for electronic records
6. Make appraisal decision
    a. Record appraisal decision
    b. Conduct destruction for non-archival electronic records
    c. Monitor appraisal decision[11]

Inputs

- Information about electronic records

[10] These record elements may include the textual content of the file; layout and "look and feel" of a document; functionality (of databases, spreadsheets, etc.).
[11] Monitoring appraisal decisions will be of greater importance for donors with whom NLM maintains an ongoing relationship. Initial appraisal decisions may be changed depending on development of new technology; changes in the Archives' preservation system; etc.

- Information about Records Creator
- Information about Archives' technology

Outputs

- Record survey
- Preservation and access format specifications
- Feasibility Reports
- Appraisal decisions
- Information about digital components to be preserved
- Updated file format registry

Tools

- DROID, *file format identification*
- File Information Tool Set (FITS), *file format identification*
- JHOVE, *file format identification*
- NLNZ Metadata Extractor, *file format identification*

Notes

The appraisal of electronic records needs to take a number of factors into consideration: the appropriateness of the fonds within the acquisition mandate of the Archives, the selection of electronic records at the file or item level and the ability of the archives to preserve the electronic records are examples. Appraisal based on each consideration requires differing sets of information and will occur at different points in the lifecycle of the record.

It will be important to note the software application used to create electronic records. As AMMP begins to receive more complex electronic records, the capacity of AMMP to view and manage these records needs to be considered. Upon (or, more preferably, before) the acquisition of complex records, the following questions need to be addressed: (1) Does NLM have the software applications necessary to view complex electronic records in question? (2) Does NIH have the software applications necessary to view complex electronic records in question? If yes, does the license cover NLM? (3) If software applications are contained on hard drives used to transfer records, is it a violation of copyright/license to utilize software for preservation purposes by the

Archives? (4) If Archives cannot view electronic records, should it be retained in its original format only? How will content be determined? How will access be provided?

The file format registry should list all the file formats held by AMMP. Information captured in this registry should include (1) file format; (2) creating software application; (3) operating system; (4) file format version; (5) creating software application version.[12]

[12] For example, for hypothetical GenericRecord.pdf, this information may be: (1) PDF, (2) Adobe Illustrator, (3) Mac OS 10.4.1, (4) PDF 1.3; (5) CS2 and would need to be captured in the registry.

D. Arrange and Describe Electronic Records

The purpose of this function is to conduct an intellectual arrangement of the electronic records and develop appropriate description tools to facilitate access.[13]

Activities

1. Determine arrangement of electronic records[14]
2. Ensure appropriate metadata is attached to electronic records
3. Develop appropriate descriptive tools for electronic records
    a. Folder structure/file list
    b. Catalog entry in LocatorPlus
    c. DACS-compliant finding aid[15]

Inputs

- Information about the context of the electronic records
- Information about the Records Creator
- Metadata

Outputs

- File structure of electronic records, with list of files within folders
- Entries in LocatorPlus and other relevant OPACs
- DACS-compliant finding aid
- Metadata

Tools

- Metadata Standards (i.e., PREMIS)
- Descriptive Standards (DACS, EAD)

Notes

[13] As the National Library of Medicine follows the More Product, Less Process principle, descriptive tools may be temporarily limited to file-lists, with more detailed instruments created when time allows for more thorough processing.
[14] The intellectual arrangement of electronic records should fit into the overall intellectual arrangement of the fonds, and not placed in a separate series based on medium. Although this activity focuses largely on intellectual arrangement, it could entail the movement of records within the fonds' file structure (e.g., moving a record that is clearly misfiled in an application folder to, say, the Documents folder).
[15] *DACS (Describing Archives: A Content Standard)* is the American standard for archival description.

Although, to ensure quick access to recently acquired records, full descriptions will not be written before the electronic records will be made accessible, the volatile nature of electronic records requires a minimum set of processing that is greater than its paper (or other traditional media)-based counterparts. This entails noting software applications (and its versions) used to create electronic records, the file formats of the records and the approximate date of creation. This will allow the Archives to flag records at risk of obsolescence, even if the fonds has not been fully processed and described.

E. Place Electronic Records in Storage

The purpose of this function is to ingest the electronic records into the National Library of Medicine's Fedora-based repository.

Activities

1. Create preservation copies of electronic records[16]
2. Create AIPs[17]
    - AIPs should contain:
        - Electronic records in original format
        - Electronic records in preservation format
        - List of files contained in AIP
        - Information on any activities performed on the records
        - Metadata and any other information for arrangement and description
3. Ingest AIPs into Fedora-based repository

Inputs

- SIPs
- Electronic Records in original format
- Metadata

Outputs

- Electronic Records in original format
- Electronic Records in preservation format
- AIPs

Tools

- BagIt, *content packaging*
- Xena, *transformation*
- PAWN, *bulk ingest*

[16] See Appendix B for a list of recommended preservation formats.
[17] AIPs should include Records in original format; Records in preservation format; and accompanying metadata.

F. Preserve Electronic Records

The purpose of this function is to monitor the preserved records in storage to ensure continuing accessibility to the content of the records.

Activities

1. Monitor storage system
    a. Check for media deterioration
    b. Back-up storage system
    c. Refresh storage media
2. Monitor electronic records in storage
    a. Ensure continuing verification of electronic records in storage[18]
    b. Check for file format obsolescence
    c. Correct problems with electronic records
    d. Update preserved electronic records
3. Document actions taken on records

Inputs

- Electronic records nearing obsolescence
- Storage Media nearing deterioration and/or obsolescence

Outputs

- Audit reports
- Migrated/Refreshed Electronic Records and/or Storage Media
- Preservation Event metadata

Tools

- Md5sum Scripts, *checksum verification scripts*

Notes

AMMP should provide regularly scheduled tests to ensure that the electronic records in permanent storage remain authentic over the long-term.

[18] Run routine checks on electronic records to ensure that records have not corrupted while in storage.

G. Provide Access to Electronic Records

The purpose of this function is to facilitate access to the Records in permanent storage.

Activities

1. Facilitate access of electronic records
    a. Provide search and other discovery capabilities for users to locate electronic records of interest
    b. Generate retrieval requests to preservation and access system
    c. Register access requests
2. Determine access status of electronic records
3. Create access copies of electronic records[19]
4. Verify access copies of electronic records
5. Create DIPs[20]
6. Ingest DIPs to Fedora-based repository
7. Receive request for 508-compliant access copies of electronic records[21]
    a. Generate 508-compliant access copies of electronic records
        - OCR/text-only
        - Oral delivery (via phone)
    b. Provide 508-compliant access copies of electronic records to User
    c. Ingest 508-compliant access copies
8. Monitor access system

Inputs

- Access requests
- Preservation copies of electronic records

Outputs

- Access copies of electronic records
- 508-compliant access copies of electronic records

[19] Creation of access copies may entail the redaction of information and/or other measures to protect privacy.
[20] DIPs should contain the access copies of electronic records, metadata and other information necessary to understand the electronic records, as well as a Certificate of Authenticity, if required.
[21] The issue of whether electronic records acquired by AMMP need to be 508-compliant needs further investigation. This report is written from the viewpoint that 508-compliant access copies of electronic records will not need to be produced automatically. Instead, the generation of 508-compliant access copies will stem from user request.

- DIPs

Tools

Notes

As the electronic records acquired by AMMP become more complex, so to will the determination of access formats. Major questions to consider when determining access formats include: (1) How important is the functionality of electronic records? Must records retain their functionality, or, can a static view be presented instead? (2) How will users be able to view the electronic records? Will we expect them to already have the software necessary to view the electronic records? Are viewers available that we can point users to (think Adobe Reader v. Adobe Acrobat Pro)?

If functionality is vital to the understanding of electronic records, the venue of delivery must be taken into consideration. If we cannot expect users to have the necessary software to view complex electronic records, perhaps a static access copy can be placed into the Fedora repository while the complex, functional access copy will be available for in-house viewing.

Appendix A: Metadata

Activity	COP Metadata[22]
B1. Register transfer	• Name of the person responsible for effecting the transfer • Transfer registration number • Date/time the transfer was received • Name of the person registering the transfer • Indication of records and other transfer documentation received • Name of person(s) to whom a notification of receipt of transfer was issued • Name of the person who issued the notification • Date/time the notification was sent

B2. Copy electronic records from storage media onto Archives server	• The records placed into server, including:
	  ○ Name of the juridical or physical person that created the records
	  ○ Name of the juridical or physical person that transferred, donated or sold the records
	  ○ Quantity and characteristics of the records
	• Original state of the records in the transfer when received
	• Indication of the security and control procedures used for the transfer
	• Indication of any modifications made to the records since their receipt
	• Indication of the post-modification state of the records

B3a. Confirm authorization for transfer	• Date/time the transfer was accepted/rejected as being authorized • Name of the person confirming/rejecting the authorization of transfer • Transfer authorization number • Terms and conditions of transfer number • Name of person(s) to whom a notification of rejection of transfer was issued • Name of person who issued the rejection notification • Indication of the reason for the rejection
B3b. Confirm authenticity of electronic records in transfer	• Date/time the transfer was accepted/rejection as containing authentic records • Indication of the measures used to assess authenticity of the records in the transfer • Name of the person confirming/rejecting the authenticity • Authenticity assessment report number • Transfer authenticity verification number • Terms and conditions of transfer number • Name of the person(s) to whom a notification of rejection of transfer was issued • Name of the person who issued the rejection notification • Date/time the rejection notification was sentIndication of the

[22] Identified metadata for each activity have been derived from the InterPARES 2 Chain of Preservation Model. To determine specific metadata elements that should be generated during each activity, the metadata identified in this table should be crosswalked to the standard selected for the repository (e.g., PREMIS).

Activity	COP Metadata[22]
	reason(s) for the rejection
B3c. Verify content of electronic records in transfer	• Date/time the transfer was accepted as verified • Indication of the measures used to verify the transfer • Name of the person verifying the transfer • Transfer content verification number
B3d. Scan for viruses and other malware	• Date/time the transfer was accepted as verified • Name of the person conducting the virus scan • Results of the virus scan • Indication of any corrective actions taken
B3e. Remove prohibited characters from file name	• Date/time the filename was changed • Original file name • Corrected file name
C2. Assess authenticity of electronic records	• Date/time the transfer was accepted/rejection as containing authentic records • Indication of the measures used to assess authenticity of the records in the transfer • Name of the person confirming/rejecting the authenticity • Authenticity assessment report number • Transfer authenticity verification number • Terms and conditions of transfer number • Name of the person(s) to whom a notification of rejection of transfer was issued • Name of the person who issued the rejection notification • Date/time the rejection notification was sentIndication of the reason(s) for the rejection

C5b. Assess feasibility of preservation	• Date/time the feasibility of preservation was confirmed/rejected
	• Name of the person confirming/rejecting the feasibility
	• Feasibility report number
	• Feasibility verification/rejection number
	• If records are determined unfeasible:
	  ○ Indication of the measures used to confirm the feasibility of preservation
	  ○ Indication of the reason(s) for rejection

D3. Develop appropriate descriptive tools for electronic records	• Date of the description • Name of the person responsible for the description • Indication of the descriptive rules used
E1. Create preservation copies of electronic records	• Indication of the original file format of the record(s) • Indication of any modifications made to the record(s) in preparation for storage (e.g., normalization) • Indication of the file format of the record(s) after modification • Indication of the reason/authorization for the modification(s) • Date/time of the modification(s) • Name of person responsible for the modification(s)
E3. Ingest AIPs into Fedora-based repository	• Date/time the record(s) place in storage • Location of the record(s) in storage
F1. Monitor storage system	• Indication of the reason/authorization for the backup • Indication of the type of backup • Indication of the extent or content of the backup • Name of the person creating the backup • Date/time of the backup • Indication of the software application (including version number) used to create the backup

Activity	COP Metadata[22]
	• Location of the backup • Backup identification number
F2. Monitor electronic records in storage	• Indication of the file format of the record(s) prior to modification • Indication of modification process(es) used • Indication of the file format of the record(s) after the modification • Indication of the reason/authorization of the modification • Name of person responsible for the modification • Date/time of the modification • Modification identification number
G1c. Register access requests	• Name of the person requesting the records/information • Access privileges of the requestor • Indication of the records and/or information requested • Date/time the request was received/registered • Name of the person registering the request • Access request registration number
G3. Create access copies of electronic records	• Indication of the original file format of the record(s) • Indication of any modifications made to the record(s) in preparation for storage (e.g., normalization) • Indication of the file format of the record(s) after modification • Indication of the reason/authorization for the modification(s) • Date/time of the modification(s) • Name of person responsible for the modification(s) • Indication of any problems encountered in generating access copies • Indication of required maintenance action(s) • Indication of any redaction for privacy or copyright reasons • Date of the redaction • Name of the person responsible for handling/executing the redaction
G4. Verify access copies of electronic records	• Date/time access copy was accepted/rejected as verified • Indication of the measures used to verify the access copy • Name of the person verifying the access copy
G5 Create DIPs	• Indication of the record(s) and/or information presented • Indication of a Certificate of Authenticity, if issued • Name of the person(s) to whom the record(s) and/or information were presented • Date when the record(s) and/or information were presented • Name of the person responsible for handling/effecting the access request
G7. Receive request for 508-compliant access copies of electronic records	*Metadata generated during G7 will be similar to metadata generated in previously listed activities in G. Provide Access to Records.*

Appendix B: Preservation Formats

As stated by Evelyn McLellan, the selection of preservation format should be based on several factors:

1. The specification must be freely available.
2. There must be no patents or licenses on the format.
3. Other established digital repositories should be using or have endorsed the format.
4. There should be a variety of writing and rendering tools available for the format.[23]

The following table includes selected preservation formats or recommended storage formats for common file types that have been published by four digital repositories: Repositório de Objectos Digitais Autênticos (RODA), Florida Digital Archive, Smithsonian Institution Archives and Simon Fraser University (SFU).

File Type	RODA[24]	Florida Digital Archive[25]	Smithsonian[26]	SFU[27]
Text (Plain)	PDF/A	Plain Text (ASCII)	No change	PDF
Text (Structured)	PDF/A	XML, PDF/A	PDF/A	PDF
Images (Raster)	METS+TIFF	TIFF, JPEG 2000, PNG	TIFF	TIFF
Images (Vector)	METS+TIFF	SVG	TIFF	TIFF
Database	DBML	CSV, TXT, SQL DDL	XML, SQL	Not listed
Audio	WAV	AIFF, WAV	WAV	Not listed
Video	MPEG-2	Motion JPEG 2000, AVI, MOV	AVI	Not listed

[23] Evelyn Peters McLellan, "Selecting Formats for Digital Preservation: Lessons Learned During the Archivematica Project" *Information Standards Quarterly* 22, no. 2 (Spring 2010): 31.
[24] RODA, "Suported File Formats," (c. 2009) <http://redmine.keep.pt/wiki/roda-public/SuportedFormats>
[25] The formats listed here are identified by the Florida Digital Archive as having a "High Confidence Level" for preservation. Florida Digital Archive, "Recommended Data Formats for Preservation Purposes in the Florida Digital Archive" (2008) <http://www.fcla.edu/digitalArchive/pdfs/recFormats.pdf>
[26] Riccardo Ferrante, "Conservation of Digital Records: A Collaborative Electronic Records Project of the Rockefeller Archive Center and the Smithsonian Institution Archives." (2006) <http://siarchives.si.edu/pdf/SIA_EREC_06_06.pdf>
[27] Richard Dancy. "F-172 TeleLearning Network Inc. fonds: Electronic Records Preservation and Access Strategy," (2005) <http://www.sfu.ca/archives2/indx/F172PrsAcesscStrat-2005.pdf>