Data Rescue Processing Guide

Version 1.0 (August 2020)

A practical guide to processing preservation-ready data from research data collections

Prepared for the National Agricultural Library of the U.S. Department of Agriculture by

Cooper T. Clarke & Hilary Szu Yin Shiue

Data Rescue Fellows at the University of Maryland's College of Information Studies

August 2020

Disclaimer

This document was prepared by Data Rescue Fellows from the University of Maryland, College Park College of Information Studies. This document was reviewed by faculty at the University of Maryland and staff at the National Agricultural Library, and has not been peer reviewed. Any recommendations, standards, or guidance do not necessarily reflect those of the National Agricultural Library, U.S. Department of Agriculture, or the University of Maryland.

Table of Contents

Part 1. Background

This processing guide was developed for use in combination with rapid data appraisal methodology in order to make valuable data available to users as quickly as possible. Data rescue is the process of identifying, intervening, and revitalizing data-rich materials at risk of loss to produce preservation-ready data. Data-rich materials are any medium, either digital and/or analog, that contain data and/or research findings. Preservation-ready data are the final product of processing, stabilized and with sufficient description for preservation and dissemination. Rapid appraisal differs from traditional appraisal in that it applies a uniform framework to determine the information's values as quickly as possible, enabling data rescue.

The considerations include evaluation of data reusability for their designated communities, as well as maintaining data quality and integrity while preserving the data for long-term access. This can be achieved by applying the frameworks and concepts of the Open Archival Information System (OAIS), which considers designated communities for data reuse while the concepts of different information packages assist data curators in assessing data. The OAIS has become an international standard (ISO 14721:2012) and is widely utilized for the long-term preservation of digital information.

Both this processing guide and the Final Report and Recommendations of the Data Rescue Project at the National Agricultural Library were written by Library and Information Science graduate students studying at the University of Maryland for use by the National Agricultural Library (NAL) of the U.S. Department of Agriculture (USDA).

Part 2. The Open Archival Information System (OAIS)

The Open Archival Information System was first initiated for "the long-term storage of digital data generated from space missions" (Lavoie, 2014), later approved as ISO International Standards in 2002 and updated in 2012 as an ISO Standard 14721:2012.

Though focused on digital data, the OAIS reference model can be extended to data, archival materials in any form (Lavoie, 2014). In this section, we introduce the six OAIS requirements and two critical concepts of the reference model, the three information packages and designated community.

A. Foundations

The OAIS establishes six requirements the system is grounded to:[1]

a. Negotiate for and accept appropriate information from information producers.

[1] From "The Open Archival Information System (OAIS) Reference Model: Introductory Guide (2nd Edition)," by B. Lavoie, 2014, p. 8. *Digital Preservation Coalition.* Accessed April 21, 2020. https://www.dpconline.org/docs/technology-watch-reports/1359-dpctw14-02/file.

b. Obtain sufficient control of the information in order to meet long-term preservation objectives.

c. Determine the scope of the archive's user community.

d. Ensure that the preserved information is independently understandable to the user community, in the sense that the information can be understood by users without the assistance of the information producer.

e. Follow documented policies and procedures to ensure the information is preserved against all reasonable contingencies, and that there are no ad hoc deletions.

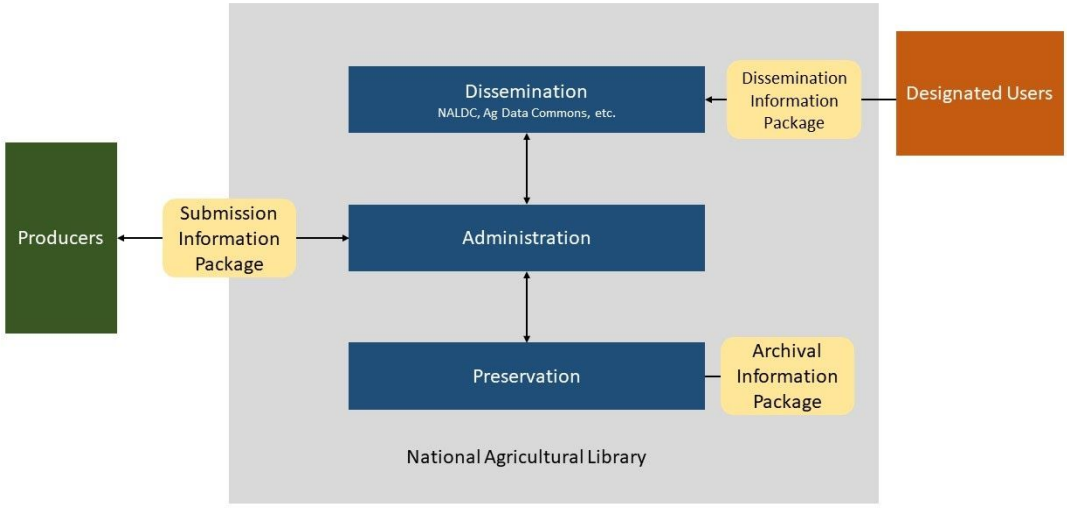
f. Make the preserved information available to the user community, and enable dissemination of authenticated copies of the preserved information in its original form, or in a form traceable to the origin.

B. Information Packages

Data being preserved in an OAIS-compliant repository passes through three information packages, from appraisal, ingest, migration, to dissemination. In the process, the repository preserves the data with sufficient metadata that documents its preservation, access and description.The packages are: the Submission Information Package (SIP), the Archival Information Package (AIP), and the Dissemination Information Package (DIP). Figure 1 presents the relations of the three information packages with the National Agricultural Library as an OAIS-compliant repository.

As a data rescue processing guide, this document is created for rapid data appraisal to efficiently preserve useful information for long-term access with the designated community in mind. To achieve this, we offer questions and topics to consider for each information package in the OAIS framework.

Figure 1: OAIS information model for the National Agricultural Library



a. Submission Information Package (SIP):

The submission information package is negotiated between the data producer, creator or donor and the repository. The preparation work of the SIP focuses on efficient communication with them. It is likely to be a back-and-forth process while gathering thorough metadata for the archival information package. Important information to gather is listed below.

1. Gather sufficient documentation (description, metadata) of data/materials from producers/creator/donor, considering future usability and long-term value.
2. Gather information about the designated community for the data and its reuse.
3. Determine terms of use, release date, and copyright.
4. Confirm authenticity and validity of materials (quality assurance).

The Data Curation Profiles Toolkit, created by Purdue University offered great insights and practical interview worksheets for information professionals.[2] There are a few topics and questions presented below that are worth considering when communicating with the data producer, creator or donor. If they are not reachable, people with domain expertise or have knowledge about the data may be able to contribute, too.

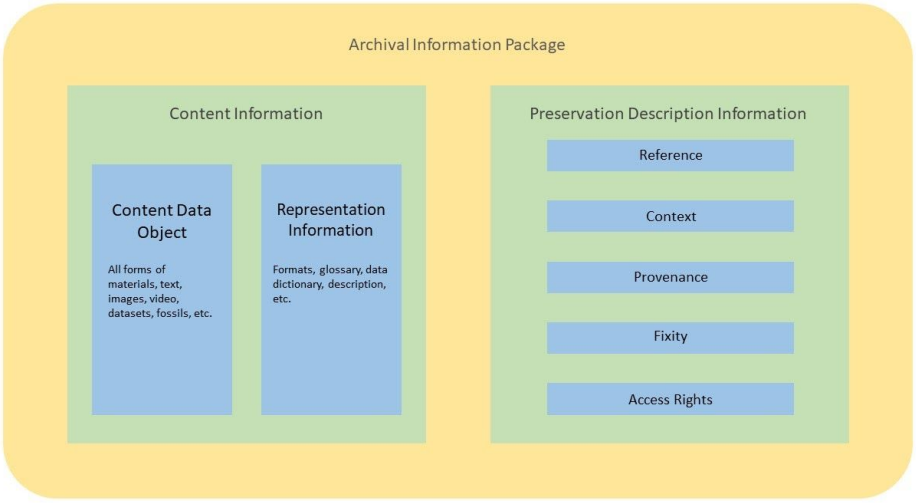
1. **Description and organization of data:** What is the data about? What is the lifecycle of the data? What media and/or formats are the data in? How are the data organized? How are they documented? Is there sufficient information for the designated communities to independently understand them?
2. **Tools:** What software or hardware were used to generate the data? What software or hardware is required to fully access and reuse the data?
3. **Reuse and discovery of data:** Who would be interested in the data? How do you imagine them reusing the data? What formats would be most useful for the designated communities? How would the designated communities discover the data?
4. **Interoperability:** Can the data be linked to other publications and/or data?
5. **Impact of data:** How would you measure the use or impact of the data?
6. **Access and intellectual property:** Are the data held by other repositories? Is there an embargo time? Who owns the data? Are there any obligations attached to the fundings, if applicable? Do the data contain private, confidential, personally identifiable information (PII) or controlled information?

[2] "The Data Curation Profiles Toolkit." by J. Carlson, 2010, *Purdue e-Pubs*. Accessed April 21, 2020. https://docs.lib.purdue.edu/dcptoolkit/

b. Archival Information Package (AIP):

The archival information package is the preservation package maintained by the repository for long-term access. The creation of the AIP can center around collecting information for its components. Figure 2 presents the structure of the archival information package.

Figure 2: The structure of the Archival Information Package



The AIP contains Content Information and Preservation Description Information. Content Information consists of Content Data Object, that is the preservation master copy for long-term use, and Representation Information, which supports the Content Data Object to be independently understandable by the designated community. Preservation Description Information is separated into five parts: Reference, Context, Provenance, Fixity and Access Rights, some of which may be gathered from the SIP.

1. Content Information:
    i. Content Data Object: It is the data itself, and should be preserved in suitable formats for long-term use and storage. Some actions may be necessary, such as reformatting proprietary file formats, reformatting print materials to machine readable formats (see the National Archives & Records Administration preferred file formats[3]).
    ii. Representation Information: It is the information necessary to make the Content Data Object independently understandable by the designated community. The considerations include description, metadata as well as types of format used. This can vary based on the level of intensity required for processing.
2. Preservation Description is metadata that describes the past and present states of the Content Information. It can be split into five parts:

[3] "Records Management Regulations, Policy, and Guidance Appendix A: Tables of File Formats," U.S. National Archives & Records Administration, (n.d.). Accessed April 21, 2020. https://www.archives.gov/records-mgmt/policy/transfer-guidance-tables.html

i. Reference Information, e.g. ISBN, linked data.
ii. Context Information: It is information relationships to other Content Information objects.
iii. Provenance Information: It provides the history of the Content Information.
iv. Fixity Information: It ensures that the Content Information has not been altered in an undocumented way.
v. Access Rights Information: It provides conditions or restrictions information associated with the Content Information pertaining to both preservation and access.

When creating the AIP, some topics and issues can be considered:

1. **Content Data Object:** What formats/media are sustainable for long-term usability and storage? Are there existing preservation guidelines to refer to?
2. **Representation Information:** What information do the designated communities need to independently understand the content data object? Is there a data dictionary and/or sufficient description? Is the structure information (such as format) independently understandable to the designated communities?
3. **Preservation Description Information:** Do we have enough information for below five items? If not, how can we gather them? Some of them are generated when processing the data, such as reference information, fixity information. Some may be in the SIP, such as access rights information.
    - Reference Information
    - Context Information
    - Provenance Information
    - Fixity Information
    - Access Rights Information

c. Dissemination Information Package (DIP):

Dissemination information package is the package delivered to users. Therefore, the format, description, level of metadata should be based on the information needs of the designated community. It should also include terms of use, copyright notices and method for content authentication, e.g. checksums of content data object files.

1. Dissemination platforms, file formats and metadata schemas should be based on designated community's user information needs.
2. Terms of use and copyright notices
3. Authentication of content

When creating the DIP, some topics and issues can be considered:

1. **Formats:** Are the data in formats that the designated communities are familiar with? Are the data stored in formats that can easily be reused?

2. **Metadata:** Are there sufficient metadata for the designated communities to understand the data? Are there metadata provided that are not of interest or use to the designated communities?

C. Designated community of the National Agricultural Library

The designated community is the primary users of the OAIS repository. The scope of the designated community "determines both the contents of the OAIS and the forms in which the contents are preserved" (Lavoie, 2014). The contents and forms provided by the OAIS framework should also "remain available to, and independently understandable by the designated community" (Lavoie, 2014). For repositories that contain discipline-specific information objects, Lavoie (2014) mentions that "the designated community might consist of all individuals possessing a certain level of expertise in that discipline." The designated community of data-rich materials at the National Agricultural Library is most likely scientists who conduct agricultural research and have a certain degree of disciplinary expertise.

The 2017 article, "Supporting the Changing Research Practices of Agricultural Scholars," provides insights about the information needs, issues, and challenges agricultural scholars encounter (Cooper et al., 2017). These scholars are both potential designated communities, donors, producers, and creators for the NAL.

The nature of agricultural research is interdisciplinary. Scholars conduct research with colleagues in other areas of expertise, and sometimes work with people outside their programs and institutions. This also reflects their data management practices and challenges:

a. Data are in a variety of formats.
b. Some data are managed by other people such as student assistants.
c. Multiple data back-up systems create version confusion in data management.
d. Researchers are unsure whether the data they are creating are worth saving.
e. Scholars themselves don't have the knowledge and infrastructure for preserving and sharing the data.
f. The institution does not implement policies for retaining knowledge and data from people who are retiring

Agricultural scholars' information seeking methods:

a. Discovering published materials - subject specific databases (AgEconSearch, AGRICOLA, CAB Abstracts, PubMed), to aggregate platforms in the sciences and even more widely (WoS, Scopus, Google Scholars)
b. Automation and social media - automated key words updates, conference email subscription, following people on social media like Twitter, academic social network sites, and so on.
c. Human discovery engines - rely on students to gather information
d. Discovering and using data created by others

From the perspective of data rescue at the NAL, we can expect to find data in various formats, created by multiple individuals and/or institutions. Our understanding of the designated

community must inform the Dissemination Information Package (DIP). Users of the data may come from a wide variety of professional fields which led us to consider that the DIP requires contextualized metadata and description for use by a diverse designated community. To address the issue of missing or incomplete metadata, our Submission Information Package (SIP), and communication with producers, should be as informative as possible. If ideal, the National Agricultural Library can recommend data management practices to producers to prevent future issues in data preservation.

Part 3. OAIS Data Processing

After a thorough literature review of data processing and rapid appraisal, we concluded that Cornell University Library's "Digital Processing Framework" provided the best guide for processing digital materials.[4] Based on the framework, we developed a work breakdown structure (WBS) for creating the three information packages in the OAIS framework. The original framework and this guide were developed for processing digital materials, however, the general principles can be adapted for other materials. As the National Agricultural Library preserves and receives materials in different formats, we reviewed the Digital Processing Framework, and created a separate WBS in section D. below for processing analog physical materials.

A. Appraisal questions

Addressing the designated community's needs at the National Agricultural Library, we list a variety of questions that should be considered to assess the quality of data. We aggregated appraisal questions and considerations from *Curating Research Data Volume II* (Johnston, 2017), Faundeen & Oleson's article, "Scientific data appraisals: The value driver for preservation efforts" (2007), and our experiences processing collections at the NAL. Some questions are broad that apply to all data rescue cases and others are specific to the data in question and should be further developed with knowledge of the field or subject matter.

1. Is the data used in a published article/study?
2. Potential for reuse in the scientific community?
3. Can this data be reproduced?
4. Is the methodology documented?
5. Has the field's methodology changed? Would that change be documented in this collection?
6. What is/was the intended use?
7. Are there any use limitations (e.g. potential issues that a user removed from the original researcher may not understand)?
8. What format would be preferred by a researcher? Not preferred?
9. Would changing the data's appearance (e.g. formatting, layout, structure) change its meaning or value?
10. How much work is required to transform the data (digitize and/or transcribe)?

[4] "Digital Processing Framework," E. Faulder et al., 2018, *Cornell University Library. Accessed April 21, 2020.* https://hdl.handle.net/1813/57659.

11. How could this data example be used by a contemporary researcher?
12. How does a field researcher measure data quality?

B. Tiers of processing

Cornell University Library's Digital Processing Framework provides three useful tiers for processing actions, which are Baseline, Moderate, and Intensive. The definitions below are drawn from the Framework with minor revisions to cater to contexts at the National Agricultural Library. Each action in our processing guide is associated with the tier. However, the tier applied to each step was based on our experiences processing a born-digital data set and a legacy analog collection at the NAL, so they may be defined differently from the original Digital Processing Framework.

a. Baseline - the minimum actions that should be taken to process the materials. Usually, the data can be made easily available as is, contains non-sensitive data, may have low reuse value.
b. Moderate - the processing actions that may require using forensic tools, or special skill set. Usually, the data is of a higher reuse value, may contain sensitive data, may require processing software to ensure preservation methods are documented.
c. Intensive - the processing actions that are most time-consuming, resource intensive, and require special tools and skill sets. The data usually requires substantial research to process and describe, may contain sensitive data, requires processing software to ensure preservation methods are documented.

C. OAIS Data Processing Framework for Digital Materials

Each action of the work breakdown structure below is referenced from the Digital Processing Framework, with our interpretation of the tier of processing for the National Agricultural Library. However, in order to more easily conceptualize these actions in the OAIS framework, we decided to map each action to the three information packages, Submission, Archival and Dissemination Information Packages. A condensed top-level guide is also provided in Section d. Condensed Framework for reference.

a. Create Submission Information Package (SIP)

1. Identify collection documentation
    - E.g. donor agreement, accession information [baseline]
2. Run virus scan
    - Run virus scan on materials and the computer system utilized to check for viruses/malware before full ingest to a secure network or system. [baseline]
    - Document results of virus scan and actions taken [baseline]
3. Survey the collection
    - Identify and document scope and content/types of collection materials [baseline]
    - Identify any sensitive information/data [baseline]
    - Determine date range [baseline]

- Review documentation of provenance and accession [baseline]
- Consult with or research creator of collection, if possible [moderate/intensive]
- Establish original order, if possible [moderate]

4. Identify restricted material based on copyright/donor agreement/sensitivity
    - Determine appropriate actions for restricted materials [baseline]
    - Flag files that need to be restricted [baseline]
    - Note if materials are likely to contain sensitive information based on context of research, creators, institutions, etc. [baseline]
5. Manage PII risk
    - Review existing restrictions in documentation and collection file [baseline]
    - Human review of materials for PII, flag and redact or dispose [baseline]
    - Pattern search software for collections with known PII risks [intensive]
6. Establish physical control over removable media
    - Identify physical media and formats [baseline]
    - Assign unique identifier to each piece of physical media [baseline]
    - Create a complete inventory of every file, noting: media type, capacity/size, file system, date, other labels [baseline]
    - Transcribe annotations on media as metadata [baseline]
7. Create submission information package (SIP)
    - Document basic technical metadata [baseline]
    - Document checksums of files (can be done automatically with BagIt compliant software) [baseline]
    - Package contents and metadata as SIP [baseline]
    - Document basic administrative metadata - assign and record persistent IDs (e.g. DOIs) [baseline]
    - Describe the contents of the SIP for collection management system [moderate]
    - Move SIP to non-temporary storage [moderate/intensive]

b. Create Archival Information Package (AIP)

1. Create processing plan
    - Determine access needs and priority of processing [baseline]
    - Establish scope and level of description [baseline]
    - Estimate time and resources required [moderate]
    - Identify relationship between digital and analog materials, if any [moderate/intensive]
    - Consult with creator [intensive]
2. Determine level of description
    - Determine level of description necessary to document the collection for the designated community [baseline]
    - Evaluate anticipated future reuse and research [baseline]

    - Review existing documentation of materials provenances and accession [baseline]
3. Identify deleted/temporary/system files
    - Use software tools (most file managers can partially complete this) to identify file formats indicating: 1. Temporary files 2. Deleted files 3. System files 4. Hidden files 5. Duplicate files [moderate]
    - Appraise system-generated files [moderate]
    - Appraise found files [moderate]
    - Apply disposition of files as needed [moderate]
4. Capture digital content off physical media
    - Document source media (i.e. what was the data stored on, capacity, file system, manufacturer, etc.) [baseline]
    - Determine capture method: 1. Capture relevant files using OS system tools 2. Copy relevant files using special copy tools 3. Create disk image using tools [baseline]
    - Work copies created [baseline]
    - Disposition of physical media (i.e. destroy or keep drive) [baseline]
5. Address presence of duplicate content
    - Assess contextual and information value of duplicate content to determine if duplicates should be deleted [baseline]
    - Document decisions in collection documentation and in finding aid/catalog entry [baseline]
6. Perform file format analysis
    - Identify file format(s) using readily available resources, e.g. file metadata, file format registry [baseline]
    - Run file format identification and verification tools to determine original file formats [moderate/intensive]
    - Analyze results for preservation risks [baseline]
    - Document formats found at appropriate descriptive level(s) [moderate]
7. Create checksums for transfer, preservation, and access copies
    - Create checksum of files using varying integrity as needed (MD5, SHA-1, SHA-256) throughout the process; before, during, and after processing (can be done automatically with BagIt compliant software) [baseline]
    - Document and store checksum in documentation, SIP/AIP/DIP [baseline]
8. Record technical metadata (preservation description information, PDI)
    - Determine and document any necessary software or tools for viewing or use [baseline]
    - Record date and method of file acquisition or disk imaging [baseline]
    - Review existing documentation of materials provenances and accession [baseline]

- Determine and record information such as: file names, file sizes, file paths, MAC times, checksums, file formats, creating software, file systems [baseline~intensive]
- Reuse technical metadata as descriptive metadata [baseline]

9. Gather metadata for description (representation information and PDI)
    - Describe items that require restrictions, including presence of PII, access, copyright, and donor restrictions. [baseline]
    - Review existing documentation of materials provenances and accession [baseline]
    - Standardize metadata language, aggregate metadata description across collection [baseline]
10. Normalize files (content data object)
    - Document original file formats [baseline]
    - Convert to preferred preservation formats as needed [moderate/intensive]
    - Validate integrity of new files [moderate]
    - Migrate files as needed (including digitization, format migration) [moderate/intensive]
11. Organize electronic files according to intellectual arrangement
    - Determine whether intellectual arrangement or the level of description warrant moving electronic files into new arrangement for preservation and access [baseline]
    - Determine if existing order should be kept, revised, or if archivist/curator should impose new order [baseline]
    - Describe digital content at the appropriate aggregate level (series, folder, item) [baseline]
    - Identify relationship between analog and digital content, if any [moderate/intensive]
    - Create file directory list [baseline]
    - Describe system of arrangement as it exists [baseline]
    - Identify similar content for intellectual arrangement [moderate]
    - Describe physical and digital content under the appropriate level (series, folder, item) [moderate]
    - Move and sort files according to intellectual arrangement [intensive]
12. Create Archival Information Package (AIP)
    - Determine if you will preserve original files and/or only normalized files [baseline]
    - Create checksums for all files held in AIP [baseline]
    - Gather files together and bundle into a container like .tar or .zip BagIt tools are recommended, e.g. BDBag and Bagger. [baseline]
    - Transfer AIP to preservation storage and verify package checksum [baseline]

c. Create Dissemination Information Package (DIP)

1. Delete work copies of files
    - Confirm that preservation and access copies are stored in the appropriate locations [baseline]
    - Securely delete working copies from workstations [baseline]
    - Document deletion [intensive]
2. Publish finding aid
    - Create or edit Encoded Archival Description (EAD) [baseline]
    - Publish EAD to publicly available environment [baseline]
    - Add description of electronic material to finding aid [moderate]
        - Determine to what level of description information about electronic material will be added; 1. Collection 2. Series 3. File 4. Item
        - Add: access statement, dates, extent (byte size), processing note, scope and content, use statement, arrangement note, formats of born-digital material
3. Publish catalog record
    - Create or update collection level MARC record [baseline]
4. Create Dissemination Information Package (DIP)
    - Create access copies from AIP [baseline]
    - Document access and use conditions [baseline]
    - Review existing documentation of materials provenance and accession [baseline]
    - Capture file-system metadata associated with the directory tree [intensive]
    - Create file directory list [baseline]
    - Create final access file list [baseline]
    - Determine access formats and compress if necessary [baseline]
    - Package access files as DIP [baseline]
    - Transfer files to delivery mechanism (online platform) [moderate]

d. Condensed framework

1. Survey the collection
2. Identify restricted material based on copyright/donor agreement
3. Manage personally identifiable information (PII) risk – repository side
4. Create processing plan
5. Determine level of description
6. Run virus scan
7. Establish physical control over removable media
8. Create SIP
9. Identify deleted/temporary/system files
10. Address presence of duplicate content
11. Capture digital content off physical media
12. Perform file format analysis

13. Create checksums for transfer, preservation, and access copies
14. Record technical metadata
15. Gather metadata for description
16. Normalize files
17. Organize electronic (physical) files according to intellectual arrangement
18. Create AIP
19. Create DIP for access
20. Delete work copies of files
21. Add description about electronic material to finding aid
22. Publish finding aid
23. Publish catalog record

D. OAIS Data Processing Framework for Analog Materials

The National Agricultural Library receives and preserves records in various formats. The processing steps below are created with the vocabularies for physical analog materials, though the overall data rescue concepts and goals are shareable across formats. Notably, it is often iterative to process analog materials, because additional information objects may be generated during processing work, and require to be included in the information package(s), such as new files created because of format migration from physical documents to electronic files, etc. The order of the steps is flexible and can be moved around to reflect actual workflows.

a. Create Submission Information Package (SIP):

1. Identify collection documentation
    - E.g. donor agreement, accession information [baseline]
2. Survey the collection
    - Identify and document scope and content/types of collection materials [baseline]
    - Identify any sensitive information/data [baseline]
    - Determine date range [baseline]
    - Review documentation of provenance and accession [baseline]
    - Consult with or research creator of collection, if possible [moderate/intensive]
    - Establish original order, if possible [baseline]
3. Establish physical control over the media
    - Identify types of physical media [baseline]
    - Create a complete inventory, such as a container list or folder list [baseline]
4. Identify restricted material based on copyright/donor agreement/sensitivity
    - Determine actions to take with content containing restricted material [baseline]
    - Flag files that need to be restricted [baseline]
    - Note if materials are likely to contain sensitive information based on context of research, creators, institutions, etc.

5. Manage PII risk
    - Review existing restrictions in documentation and collection file [baseline]
    - Review materials for PII, flag and redact or dispose [baseline]
6. Create SIP
    - Package contents and metadata as SIP [baseline]
    - Document basic administrative metadata - assign and record unique identifiers [baseline]
    - Describe the contents of the SIP for collection management system [moderate]
    - Move SIP to non-temporary storage [moderate]

b. Create Archival Information Package (AIP):

1. Create processing plan
    - Determine access needs and priority of processing [baseline]
    - Establish scope and level of description [baseline]
    - Estimate time and resources required [moderate]
    - Identify relationship between digital and analog materials, if any [moderate/intensive]
    - Consult with creator [intensive]
2. Determine level of description
    - Determine level of description necessary to document the collection for the designated community [baseline]
    - Evaluate anticipated future reuse and research [baseline]
    - Review existing documentation of materials provenances and accession [baseline]
3. Address presence of duplicate content
    - Assess contextual and information value of duplicate content to determine if duplicates should be discarded [baseline]
    - Document decision in collection documentation and in finding aid/catalog entry [baseline]
4. Normalize files (for content data object) if digital objects are generated
    - Document original media formats [baseline]
    - Convert to preferred preservation formats [moderate/intensive]
    - Validate integrity of new files [baseline]
    - Migrate files as needed (including digitization, format migration) [moderate/intensive]
5. Analyze media for preservation risks [baseline]
6. Gather metadata for description
    - Describe items that require restrictions, including presence of PII, access, copyright, and donor restrictions. [baseline]
    - Review existing documentation of materials provenances and accession [baseline]
    - Standardize metadata language, aggregate metadata description across collection [baseline]

7. Create AIP
    - If digital objects are generated, create checksum of AIP package [baseline]
    - If digital objects are generated, gather files together and bundle into a container like .tar or .zip. BagIt tools are recommended, e.g. BDBag and Bagger. [baseline]
    - Transfer AIP to preservation storage and verify package checksum, if applicable [baseline]

c. Create Dissemination Information Package (DIP):

1. Create DIP
    - Create access copies from AIP [baseline]
    - Document access and use conditions [baseline]
    - Review existing documentation of materials provenance and accession [baseline]
    - Create container list, and/or file directory list, if applicable [baseline]
    - Create final access container list, and/or file directory list, if applicable [baseline]
    - Determine access formats and compress digital files, if necessary [baseline]
    - Package access files as DIP [baseline]
    - Transfer digital files to delivery mechanism (online platform) [moderate]
2. Publish finding aid
    - Create or edit Encoded Archival Description (EAD) [baseline]
    - Publish EAD to publicly available environment [baseline]
    - Add description of electronic material, if applicable, to finding aid [moderate]
        - Determine to what level of description information about electronic material will be added; 1. Collection 2. Series 3. File 4. Item
        - Add: access statement, dates, extent (byte size), processing note, scope and content, use statement, arrangement note, formats of born-digital material
3. Publish catalog record
    - Create or update collection level MARC record [baseline]

d. Condensed framework

1. Identify collection documentation
2. Survey the collection
3. Establish physical control over the media
4. Identify restricted material based on copyright/donor agreement/sensitivity
5. Manage PII risk

6. Create SIP
7. Create processing plan
8. Determine level of description
9. Address presence of duplicate content
10. Normalize file (for content data object) if digital objects are generated
11. Analyze media for preservation risks
12. Gather metadata for description
13. Create AIP
14. Create DIP
15. Publish finding aid
16. Publish catalog record

References

Carlson, J. (2010). The Data Curation Profiles Toolkit Interviewer's Manual. doi:10.5703/1288284315651

Carlson, J. (2010). The Data Curation Profiles Toolkit Interview Worksheet. doi:10.5703/1288284315652

Carlson, J. (2010). The Data Curation Profiles Toolkit User Guide. doi:10.5703/1288284315650

Carlson, J. (2010). The Data Curation Profile Toolkit: The Profile Template. doi:10.5703/1288284315653

Cooper, D., Bankston, S., Bracke, M. S., Callahan, B., Chang, H., Delserone, L. M., & Diekmann, F. (2017). Supporting the changing research practices of agriculture scholars. Ithaka S+R. https://doi.org/10.18665/sr.303663

Faulder, E. et al. (2018, August 01). Digital Processing Framework. Retrieved August 18, 2020, from https://hdl.handle.net/1813/57659

Faundeen, J. L., & Oleson, L. R. (2007). Scientific data appraisals: The value driver for preservation efforts. In *Proceedings of PV 2007 International Conference*. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.94.8035&rep=rep1&type=pdf.

Johnston, L. R. (Ed.). (2017). *Curating research data: Volume two: A handbook of current practice* [ebook]. Association of College and Research Libraries, American Library Association. http://www.ala.org/acrl/sites/ala.org.acrl/files/content/publications/booksanddigitalresources/digital/9780838988633_crd_v2_OA.pdf

Lavoie, B. (2014). *The Open Archival Information System (OAIS) reference model: Introductory guide* (2nd ed.). Digital Preservation Coalition. https://doi.org/10.7207/twr14-02

U.S. National Archives and Records Administration (n.d.). Records Management Regulations, Policy, and Guidance Appendix A: Tables of File Formats. Retrieved August 18, 2020, from https://www.archives.gov/records-mgmt/policy/transfer-guidance-tables.html