Chapter 5:
The DAITSS Archiving Process

Topics covered in this chapter:

Overview and glossary

DAITSS archives materials assembled in Submission Information Packages (SIPs), which consist of directories containing content files and a SIP descriptor file that serves essentially as a "shipping manifest" for the SIP. SIPs enter the DAITSS archiving work flow via the Submission function which checks that the SIP is well-formed and valid and, if the SIP is valid, DAITSS sends the package through the archiving process until it is stored in your storage silo(s). Chapter X will describe how to use the DAITSS User Interface to track your packages.

Glossary

Archival Information Package (AIP): Contents:

- AIP descriptor
- All originally submitted content files.

AIP tar file: A single file in tar format consisting of multiple files that comprise an AIP.

Content file: A data object that is the target of preservation. Content files are contained in Submission Information Package directories, and are described in the Descriptor file of the SIP.

Checksum (or Message Digest): A string of characters produced by an algorithm computed on a file. This checksum is recomputed on all files contained in a Submission Information Package after receipt of the package by the FDA, and comparison of the recomputed value against the value provided in the SIP descriptor file assures the FDA that the file has been correctly transmitted.

Descriptor file: An xml document in METS format containing preservation description Information. It serves as a "packing slip" or "manifest" to indicate who is submitting content for archiving and lists details about each of the content files submitted for archiving. The name of the descriptor file must be identical to the name of the folder or directory in which the Submission Information Package is contained. For example, if a SIP directory name is ABC, the descriptor file must be named ABC.xml. In addition, the descriptor file must reside in highest level directory of the Submission Information Package. Please consult the DAITSS METS Document Profile for Submission Information Packages for complete specifications for FDA METS SIP descriptors.

Intellectual Entity: Something that can reasonably be described and used as a unit, and corresponds roughly to what might be described by a bibliographic record: a book, a sound recording, a photograph. (In the case of serial publications, it is recommended that a SIP include only a single issue, not a volume or set of volumes.)

Metadata Encoding and Transmission Standard (METS): A standard for encoding descriptive, administrative, and structural metadata within a digital library, expressed using the XML schema language of the World Wide Web Consortium. The standard is maintained in the Network Development and MARC Standards Office of the Library of Congress, and is being developed as an initiative of the Digital Library Federation. (Website: http://www.loc.gov/standards/mets/)

Package: An information package, usually consisting of one or more data files and a container folder or directory. This is a generic term that can be applied to a SIP, AIP, WIP, or DIP, each of which has a more specific definition.

Producer: The individual or institution who owns the information package submitted for archiving.

Submission Information Package (SIP): A SIP is defined in the Open Archival Information standard (OAIS) as an information package delivered to a repository for archiving. For submission to the Florida Digital Archive, a SIP must follow certain rules, outlined below. The FDA recommends that a SIP contain only one Intellectual Entity.

Workspace Information Package (WIP): A SIP (Submission Information Package) that has successfully passed the Submission/Validation processing and is available for archiving. The internal structure of a WIP is vastly different from that of a SIP, AIP, or DIP.

DAITSS Submission Information Packages (SIPs)

Materials for archiving are transmitted by FDA Affiliates in "packages" called Submission Information Packages (SIPs). Physically, a SIP is a single folder (directory) containing all of the content files that comprise a single Intellectual Entity, as well as a METS SIP descriptor file that serves as a "packing slip" for the contents of the SIP. SIP physical structure and content specifications are described in more detail below.

SIP specifications

Note that the DAITSS Work Flow Interface "submit" function requires that all SIPs, as described below, be submitted as tar files or in .zip format. This means that the SIP single high level folder must be tarred or zipped.

SIP physical structure requirements:

- The SIP must be contained within a single high level folder. (Subdirectories within the high level folder are allowed with some restrictions noted below.
- The size of the SIP must not exceed 100GB.
- The SIP (directory) name can follow any naming system developed by the producer, with the following restrictions:
  - The SIP folder (directory) name is limited to 32 characters, using only the following character set:
  - A-Z, a-z, 0-9, underscore (_), hyphen (-), dot/period (.), exclamation point (!), parentheses (), single space
  - SIP folder (directory) names may not start with dot/period (.)
- The SIP must contain exactly one valid METS descriptor file (an XML file describing the package) and at least one content file. The METS SIP descriptor must conform to the DAITSS METS Specifications.
- SIP content file names are limited to 245 characters, using only the following character set:
  - A-Z,a-z, 0-9, underscore (_), hyphen (-), dot/period (.), exclamation point (!), parentheses (), single space
  - Content file names may not start with dot/period (.)
- The name of the SIP descriptor file must be the same as the name of the SIP directory. That is, if the SIP directory is named A00000123 then the SIP descriptor within that directory must be named A00000123.xml. (Note that the file extension must be ".xml".)
- The SIP descriptor must reference all content files in the SIP that are meant to be archived.
- SIPs may have lower-level directories, with the following restrictions:
  - The descriptor file for the entire package must reside in the highest-level package directory
  - The lower-level directories must contain only content files
  - The relative pathname of the content files must be listed in the in the xlink:href attribute of the <FLocat> element of the file section (<fileSec>) of the descriptor file.

Examples of the physical structure of a Submission Information Package

Example 1, illustrating the minimum required physical structure (at least one content file and a descriptor file):

```
/ A00000123:  (a package directory/folder named A00000123)
        A00000123.xml (the METS descriptor file)
        content_file.pdf (a content file)
```

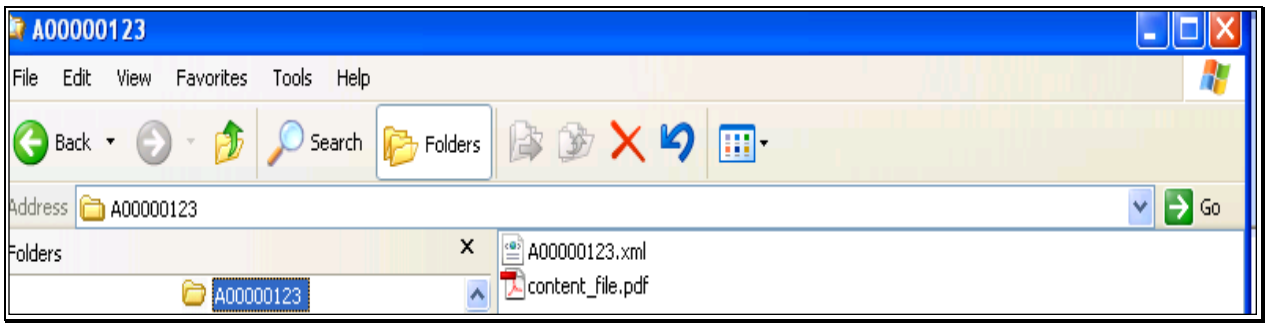
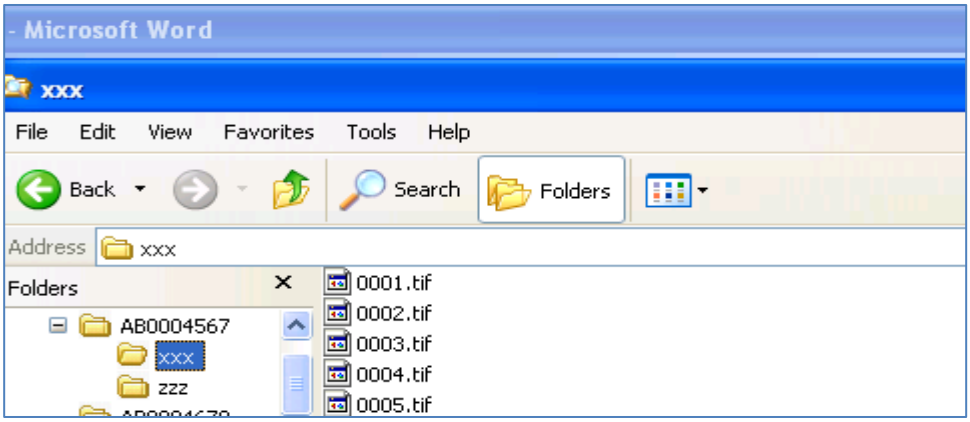


Figure 4-1: SIP with one content file

Example 2, illustrating a multi-directory package containing one descriptor file and multiple content files contained in sub-directories:

/AB0004567 (a package directory/folder named AB0004567)
AB0004567.xml (the METS descriptor file, located in the top-level directory)
/xxx (a lower-level directory)
0001.tiff
0002.tiff
0003.tiff
0004.tiff
0005.tiff
/zzz (a lower-level directory)
0006.tiff
0007.tiff
/AB0004567 (a package directory/folder named AB0004567)
AB0004567.xml (the METS descriptor file, located in the top-level directory)
/xxx (a lower-level directory)
0001.tiff
0002.tiff
0003.tiff
0004.tiff
0005.tiff
/zzz (a lower-level directory)
0006.tiff
0007.tiff

On a Windows PC, the contents of the xxx directory/folder within SIP AB0004567 would appear as:



Note that the METS SIP descriptor for SIP AB0004567 would reference the location of the content files contained in directories "xxx" and "zzz" by including a pathname relative to the SIP directory in the xlink:href attribute of the <FLocat> element of the file section (<fileSec>), as illustrated below:

```
<METS:file ID="file-1" CHECKSUM="5ddb5736a014619bbbb3684bc6ae1613"
     CHECKSUMTYPE="MD5">
  <METS:FLocat LOCTYPE="URL" xlink:href="xxx/0001.tif" />
      </METS:file>
```

SIP requirements and recommendations:

- **Include only one Intellectual Entity per SIP (recommended):** It is recommended practice that a single SIP should include only those files that comprise a single Intellectual Entity. An Intellectual Entity is defined as something that can reasonably be described and used as a unit, and corresponds roughly to what might be described by a bibliographic record: a book, a sound recording, a photograph. (In the case of serial publications, it is recommended that a SIP include only a single issue, not a volume or set of volumes.)
- **Reference all content files in descriptor (Required):** In order to confirm correct transmission of each file contained within a SIP, all files meant to be archived must be referenced in the METS SIP descriptor, as detailed below.
- **Include a valid DAITSS Account and Project Code in the descriptor (Required):** The SIP descriptor file must contain a valid DAITSS Account code and a valid DAITSS Project code associated with that Account in the <amdSec>:

```
<METS:amdSec>
        <METS:digiprovMD ID=”[unique id]”>
                <METS:mdWrap MDTYPE=”OTHER” OTHERMDTYPE=”DAITSS”>
                        <METS:xmlData>
                                <daitss:daitss>
                                        <daitss:AGREEMENT_INFO
ACCOUNT=”[account code]” PROJECT=”[project code]”>
                                </daitss:daitss>
                        <METS:xmlData>
                </METS:mdWrap>
        </METS:digiprovMD>
</METS:amdSec>
```

- **Checksum information in descriptor (strongly recommended):** In order to confirm that the content files received by your repository have not been modified during transmission, it is strongly recommended that the SIP descriptor file contain CHECKSUM information about each content file included in the SIP. This information should be recorded in the CHECKSUM and CHECKSUMTYPE attributes of file element (<file>) of the METS file section (<fileSec>), as illustrated below:

```
<METS:fileSec>
  <METS:fileGrp>
     <METS:file ID="file-1"
     CHECKSUM="5ddb5736a014619bbbb3684bc6ae1613"
     CHECKSUMTYPE="MD5">
     <METS:FLocat LOCTYPE="URL" xlink:href="0001.tif" />
     </METS:file>
  </METS:fileGrp>
    </METS:fileSec>
```

- **Include a title in descriptor <dmdSec> (strongly recommended):** Because DAITSS database stores the Title of the Intellectual Entity as an access point to the package, it is recommended practice to include a title in the descriptive metadata section (<dmdSec>), as illustrated below:

```
    <METS:dmdSec ID="[unique id]">
  <METS:mdWrap xmlns:METS="http://www.loc.gov/METS/" MDTYPE="MODS"
  MIMETYPE="text/xml">
     <METS:xmlData>
        <mods:mods xmlns:mods="http://www.loc.gov/mods/v3">
           <mods:titleInfo>
              <mods:title>Title of intellectual entity</mods:title>
           </mods:titleInfo>
        </mods:mods>
     </METS:xmlData>
       </METS:mdWrap>
  </METS:dmdSec>
```

```
<METS:dmdSec ID="[unique id]">
<METS:mdWrap xmlns:METS="http://www.loc.gov/METS/" MDTYPE="MODS"
MIMETYPE="text/xml">
    <METS:xmlData>
        <mods:mods xmlns:mods="http://www.loc.gov/mods/v3">
            <mods:titleInfo>
                <mods:title>Title of intellectual entity</mods:title>
            </mods:titleInfo>
        </mods:mods>
    </METS:xmlData>
    </METS:mdWrap>
</METS:dmdSec>
```

- **For Serials, include the Volume and Issue information in the descriptor (recommended):** In addition to any serial volume and issue information provided in descriptive metadata, Volume and Issue information should be included in LABEL, ORDERLABEL and TYPE attributes in the division ( ) element of the <structMap> section of the descriptor, as in the following example. These are also indexed by DAITSS for providing access to the package.

```
<METS:structMap>
    <METS:div DMDID="DMD1" LABEL="Volume 25 (2005-2006)" ORDERLABEL="25"
    TYPE="volume">
    <METS:div DMDID="DMD2" LABEL="Number 3" ORDERLABEL="3" TYPE="issue">
        <METS:fptr FILEID="[unique file ID]"
        ...
    </METS:div>
    </METS:div>
</METS:structMap>
```

- **Include all Intellectual Entity content files (recommended):** It is recommended practice that the SIP include all of the files needed to render at least one version of the Intellectual Entity.
- **Include a content file containing descriptive metadata (recommended):** Because archived packages are intended for long-term preservation, it is recommended that a file containing detailed descriptive metadata be included as one of the content files if detailed descriptive metadata is not provided in the

SIP descriptor file. Descriptive metadata files can be in any format; their contents will not be indexed or directly accessible from the repository, but a detailed descriptive metadata content file can enhance the understandability and usability of an information package after dissemination.

Bitstreams within SIP content files:

The FDA will extract and store within its preservation database the technical metadata from only the first 1,000 bitstreams contained within any given SIP content file. (Content files with more than 1,000 bitstreams are very likely to be malformed.) Such content files will be archived with an anomaly indicating "excessive number of … bitstreams".

Reasons for Rejected SIPs:

Submission Information Packages (SIPs) will be rejected by DAITSS software and will not be archived under the following circumstances:

- If the SIP does not contain a descriptor file at the highest directory level or if the descriptor file is misnamed.
- If the SIP descriptor file is not a valid METS file.
- If the SIP descriptor file does not contain both a valid ACCOUNT code and a valid PROJECT code for that account.
- If the SIP descriptor references a file that is not included in the SIP directory/folder. (Note that files contained in the SIP directory/folder but not referenced in the SIP descriptor will be deleted and will not be archived.)
- If the contents of the CHECKSUM attribute of any file referenced in the SIP descriptor file does not match the checksum of the submitted file. (The DAITSS software computes the checksum value of each submitted file during the Ingest process, and compares that value against the value provided in the CHECKSUM attribute.)
- If the SIP directory name or content files contain any illegal characters.
- If the SIP does not contain any content files (the SIP descriptor file is not considered a content file).

Using the SobekCM METS Editor to create SIP descriptors

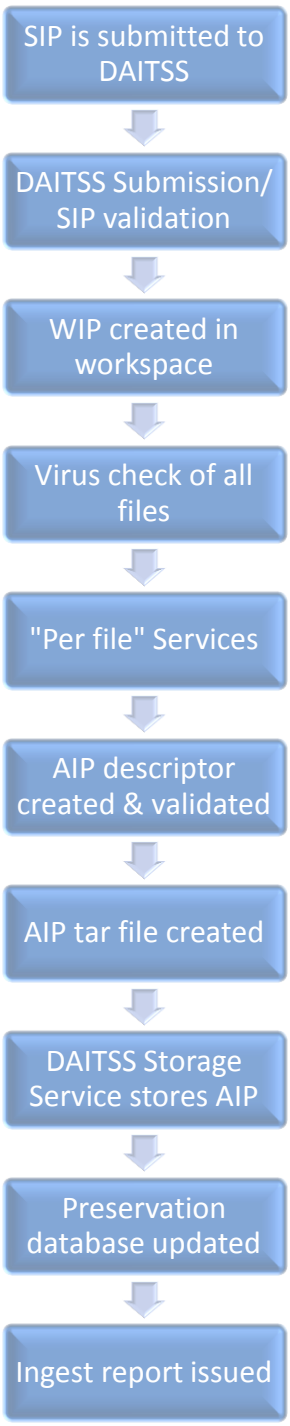
The SobekCM METS Editor, an Open Source Windows client developed by the University of Florida, can be a useful tool for creating SIP descriptors that meet the DAITSS SIP specifications. You must start with a high level folder containing all of the files required for your SIP. The METS Editor will create a SIP descriptor based on the contents of that folder. The SIP descriptor must be saved to this folder and renamed with a .xml extension.

The most recent version of the METS Editor can be downloaded from http://sourceforge.net/projects/metseditor/files/. A Users Guide is available on: http://ufdc.ufl.edu/metseditor

Note: To use the SobekCM METS Editor with DAITSS, select the "I will use this in support of the Florida Digital Archive and/or the PALMM digital library" option in the "Initialization: General Use" screen, and in the "Initialization: Select Template Add-Ons" screen select "FCLA Add-on for FCLA-destined packages". Then set the "Initialization: FCLA Add-On Settings" by checking the "FDA" box and filling in valid DAITSS Account and Project codes in the "FDA Account" and "FDA Project" sections.

DAITSS Archiving Workflow

The diagram below provides a brief overview of the DAITSS archiving process. A detailed description of key processes follows the diagram.



Key processes:

Submitting SIPs to DAITSS: SIPs can be submitted to DAITSS either one at a time via the User Interface or in batch via the command-line submission client.

DAITSS Submission and SIP validation service: the DAITSS Submission Service validates all Submission Information Packages to ensure that they are valid and well-formed. Valid SIPs are parsed and moved into a Workspace Information Package (WIP) for ingest processing. Invalid SIPs are deleted from DAITSS so source copies should be kept until the SIPs have been accepted. A Reject Report is produced for each rejected SIP. Events relating to the submission of the SIP are permanently retained in the DAITSS database. Note that any files included in the SIP but not correctly described in the SIP Descriptor will be deleted and their deletion will be noted in the DAITSS database during this process.

WIP created in workspace: If a SIP passes Submission/validation, it is parsed into a Workspace Information Package (WIP) for ease of processing. The format of the WIP is substantially different from the SIP.

Virus checking: After a WIP is created, all content files in the Workspace Information Package (WIP) are checked for viruses. If a virus is found in any files the WIP is flagged with a "SNAFU" Operations Event. See Chapter 5 for information about managing WIPs.

DAITSS "per file" services: Each file in the SIP is processed through the "per file" services in turn until all submitted files have been processed. (Note that files in the SIP not referenced by the SIP Descriptor file are not processed and are not retained in the AIP.) Processing includes file format identification, format validation, extraction of format-specific technical metadata, creation of derivative (normalized and/or migrated) versions if required, and special handling for certain file formats (e.g., XML resolution).

Creation of an AIP descriptor: Each step of the archiving process produces a PREMIS-based xml snippet. At the completion of the "per file" processing, the individual snippets are assembled into an AIP descriptor and the descriptor is validated. A sample AIP descriptor is attached in Appendix A.

Creation and storage of the AIP tar file: The entire DAITSS Archival Information Package (AIP) is assembled into a single UNIX tar file for storage. DAITSS computes an md5 and sha1 checksum on each AIP tar file before it is forwarded to the DIATSS Storage Service for deposit into archival storage. Details of the DAITSS Storage Service can be found in Chapter 6.

DAITSS preservation database: The DAITSS database contains both records of the movement of packages through processing as well as PREMIS event and object information for each AIP and its content files. The complete AIP descriptor XML file is stored in the preservation database. A detailed description of the DAITSS preservation database can be found in (Data dictionary Appendix).

DAITSS storage database(s): Each DAITSS silo pool is managed and monitored via a storage database located on that silo pool. Details of the DAITSS storage databases can be found in (Data Dictionary Appendix).