Florida Digital Archive (FDA)
SIP Specification

Version 2.2, November, 2012

Superseded versions:

Florida Digital Archive (FDA) SIP Specification Version 2.1, June 2012
Florida Digital Archive (FDA) SIP Specification Version 2.0, May 2011
FCLA Digital Archive (FDA) SIP Specification Version 1.0, February 2006

Sections changed in this version:

The SIP Physical structure requirements section has been modified to include a list of characters specifically disallowed in SIP folder (directory) names and SIP content file names.

© Copyright 2012 by the Florida Virtual Campus.

General Information

This document addresses procedures and specifications for creating Submission Information Packages (SIPs) and transmitting them to the Florida Digital Archive (FDA). It is assumed that Electronic Theses and Dissertations (ETDs) submitted to FCLA should also be archived in the FDA. Please refer to the Support for Electronic Theses and Dissertations page and the METS SPECIFICATION FOR SUL ETDS USING DUBLIN CORE document for details about ETD submissions and ETD descriptor files.

Definitions

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.)

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.

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.

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.

Bitstream: Bitstream objects are subsets of files. A bitstream object is defined as data (bits) within a file that a) have common properties for preservation purposes, and b) cannot stand alone without adding a file header or other structure. .

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.

Submission Information Package (SIP) specifications

SIP Physical structure requirements:

- The SIP must be contained in a single folder.
- The size of the SIP must not exceed 100 GB.
- The SIP (directory) name can follow any naming system developed by the FDA Affiliate, but the name must be unique. (If more than one SIP of the same name is submitted to the FDA, it will be assumed that all such SIPs are intended for archiving and it will be the Affiliate’s responsibility to request withdrawal SIPs with duplicate names from the archive.)
- The SIP folder (directory) name is limited to 32 characters.
- The following characters cannot be used in SIP folder (directory) names because they cause problems in XML documents such as the SIP and AIP descriptors:
    - semi-colon: “;”
    - slash: “/”
    - reverse slash: “\”
    - question mark: “?”
    - colon: “:”
    - at sign: “@”
    - ampersand: “&”
    - equals sign: “=”
    - plus sign: “+”
    - dollar sign: “$”
    - comma: “,”
    - curly brackets: “{“ and “}”
    - vertical line: “|”
    - caret: “^”
    - square brackets: “[“ and “]”
    - multiple spaces
    - SIP directory names may not start with dot/period (.)

  The use of the above characters in SIP names will cause SIPs to be rejected when submission to the archive is attempted.
- Some characters allowed by DAITSS software may cause problems in file transfer other applications, so the FDA strongly recommends using only the following character set:

        - A-Z,a-z, 0-9, underscore (_), hyphen (-), dot/period (.), exclamation point (!), parentheses ()
- 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 220 characters.
- The following characters cannot be used in SIP content file names because they cause problems in XML documents such as the SIP and AIP descriptors:
    - semi-colon: ";"
    - slash: "/"
    - reverse slash: "\"
    - question mark: "?"
    - colon: ":"
    - at sign: "@"
    - ampersand: "&"
    - equals sign: "="
    - plus sign: "+"
    - dollar sign: "$"
    - comma: ","
    - curly brackets: "{" and "}"
    - vertical line: "|"
    - caret: "^"
    - square brackets: "[" and "]"
    - multiple spaces
    - SIP content file names may not start with dot/period (.)

The use of the above characters in SIP content file names will cause SIPs to be rejected when submission to the archive is attempted.

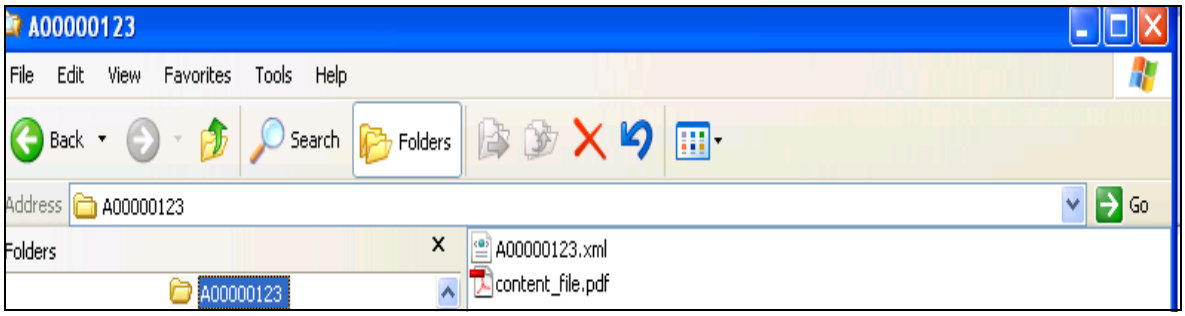
- Some characters allowed by DAITSS software may cause problems in file transfer other applications ,so the FDA strongly recommends using only the following character set:
    - A-Z,a-z, 0-9, underscore (_), hyphen (-), dot/period (.), exclamation point (!), parentheses ()
- The name of the SIP descriptor 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 must be named A00000123.xml. (Note that the file extension must be in lowercase.)
- 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)

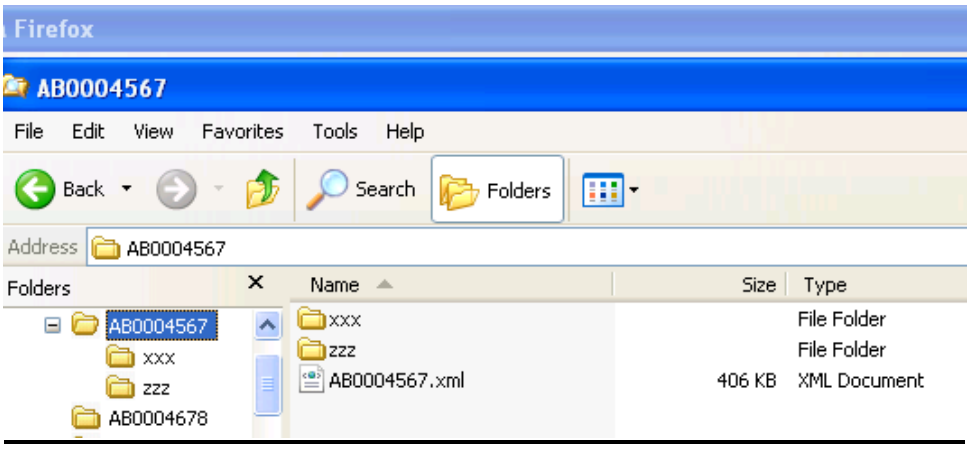
On a Windows PC, this SIP would appear as follows:



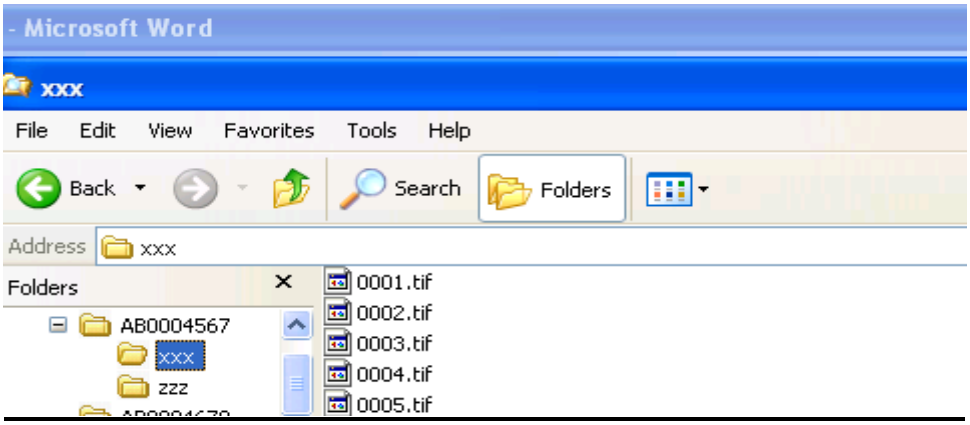
Example 2, illustrating a multi-directory package containing one descriptor file and multiple content files contained in sub-directories:

/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

On a Windows PC, this SIP would appear as follows:



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 content 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.)
- **Include Processing Instructions in the descriptor (Required for SIPs deposited on FCLA's FTP server):** Because SIPs can be deposited on the FCLA FTP server for loading into the PALMM project database and/or archiving in the FDA, the processing instruction(s) must be added as the second line of the METS descriptor file, after the XML declaration node. (An exception to this rule is made in the case of ETDs, which are assumed to be destined for the FDA in addition to FCLA's ETD server):

```
<?xml version="1.0" encoding="UTF-8"?>
<?fcla fda="yes"?>
<?fcla dl="yes"?>
```

FDA actions for ftp-ed SIPs based on Processing Instructions:

<?fcla fda="yes"?>	Send to FDA
<?fcla dl="yes"?>	Send to PALMM or ETD servers (see note below regarding special handling of ETDs)
<METS:note>projects=ETD</METS:note>	Send to ETD server and to the FDA
No processing instructions and no "projects=ETD" flag	Send to BAD directory and notify institution

.

- **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 FDA Account and Project Codes in descriptor (Required):** The SIP descriptor file must contain the Affiliate's FDA account code and the FDA project code associated with the package. These codes are specified in Appendix A of the "FCLA—Library Agreement". The Account and Project codes must be contained in a daitss:daitss block in the administrative metadata section (amdSec) of the descriptor, as follows:

```
<METS:amdSec>
    <METS:digiprovMD ID="[unique id]">
        <METS:mdWrap MDTYPE="OTHER" OTHERMDTYPE="DAITSS">
            <METS:xmlData>
                <daitss:daitss>
                    <daitss:AGREEMENT_INFO ACCOUNT="[required FDA
                    account code]" PROJECT="[required FDA 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 the FDA 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 the FDA database stores the Title of the Intellectual Entity if provided, it is recommended practice to include a title in the descriptive metadata section (<dmdSec>). Title information can be stored in using

MARC, MODs, or Dublin Core schemas. An example of title information using the MODS schema is 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>
```

- **For Serials, include the ISSN, Volume and Issue information in the descriptor (recommended):** In addition to any serial volume and issue information provided in descriptive metadata, it is recommended that the ISSN of the serial be included in the OBJID attribute of the METS root element, starting with a two character "SN" prefix (<METS:mets> OBJID=SN[*issn number]*). 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:

```
<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 also 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 in SIP (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".

Rejected SIPs:

Submission Information Packages (SIPs) will be rejected by the FDA's 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 file names contain any illegal characters.
- If the SIP does not contain any content files (the SIP descriptor file is not considered a content file).

Note that rejected Submission Information Packages are deleted and are not retained by the FDA. FDA Affiliates must retain a copy of SIPs until they are successfully archived in the FDA repository.