

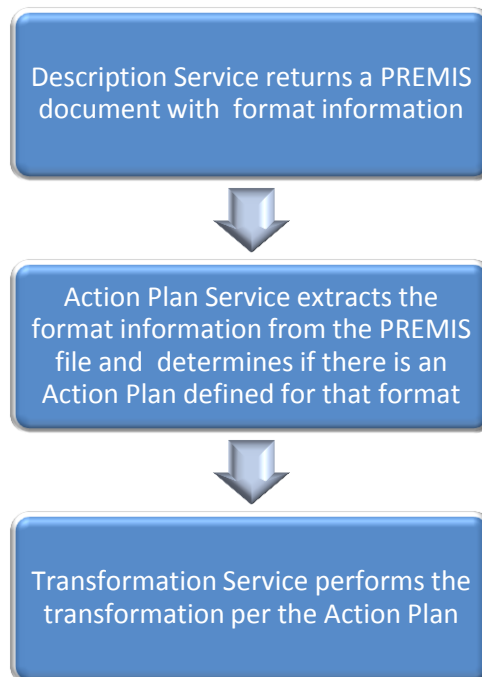
# Chapter 4: DAITSS Preservation Services

---

**Topics covered in this chapter:**

- ✓ [DAITSS Description Service](#)
- ✓ [DAITSS Action Plan Service](#)
- ✓ [DAITSS Format Transformation Service](#)
- ✓ [DAITSS XML Resolution Service](#)

## Preservation Diagram/Overview



Note that if the Transformation Service is used to create a normalized or migrated version of the file, that newly created file is sent through the Description Service, Action Plan Service, and Transformation Service in turn, just as the original file was.

Detailed information about each service is discussed below.

### **DAITSS Description Service**

The DAITSS Description Service performs format identification, validation and characterization on a given file.


Identification is the process of identifying the format and version of a file based on its internal signature. The Description Service uses the tool Droid to perform format identification.

Validation is the process of comparing the characteristics of a given file with the requirements of the file format specification. Characterization is the process of extracting technical metadata from the file for use as metadata. The Description Service uses the tool JHOVE to perform validation and characterization, but technical metadata returned by JHOVE is transformed to standard formats by a stylesheet translation.

The results of the identification, validation and characterization steps are recorded in a [PREMIS](#) document containing a PREMIS object with the identification and characterization information, a PREMIS event describing the validation status (along with any anomalies found), and a PREMIS agent identifying the software generating the PREMIS output, in this case, the Description Service.

### **Using the stand-alone DAITSS Description Service**

The DAITSS Description Service can also be used as a stand-alone service. The URL for the stand-alone service is <http://description.fcla.edu/>.



**Description Service**  
identify, validate and extract

Description Service performs format identification, validation and characterization on a given file. The result is expressed in [PREMIS](#) schema containing a PREMIS object for the identification and characterization result, a PREMIS event describing the validation status with applicable anomaly, and a PREMIS agent identifying the software agent that generates PREMIS output, in this case, the description service.

[Submit](#) [Information](#)

**Submit :** file uri

Submit a file to the description service. The result of format description may be saved locally using the browser "Save Page As" function.

File:

Alternatively, users can use HTTP clients such as curl to submit local files directly to the description service via HTTP protocol. For example, to use curl to upload a local file named 00001.pdf to the description service and save the result to premis.xml,

```
curl -F "document=@00001.pdf" -F "extension=pdf" http://description.fcla.edu/description > premis.xml
```

Address:

Alternatively, users can use HTTP clients such as curl to submit URI address directly to the description service via HTTP protocol. For example, to use curl to submit the resource located at <http://www.fcla.edu/daitss-test/files/00004.txt> to the description service and save the result to premis.xml,

```
curl http://description.fcla.edu/describe?location=http://www.fcla.edu/daitss-test/files/00004.txt > premis.xml
```

## DAITSS Action Plan Service

The Action Plan Service oversees all format action plans. It receives from the Description Service a PREMIS document describing a file, and extracts from that the file format information which it needs to look up the appropriate Action Plan for the format. If the format Action Plan dictates that a format transformation is needed, the Action Plan Service returns a transformation identifier which will be used by the Transformation Service to create a derivative version of the file.

## Sample action plan

Below is a sample Action Plan for the JPEG2000 format.

```
<?xml version="1.0" encoding="UTF-8"?>
<action-plan format="JPEG 2000" format-version="1.0">
  <implementation-date>2005.11.07</implementation-date>
  <revision-date>2010.09.16</revision-date>
  <review-date>2008.01.07</review-date>
  <next-review>2012.01</next-review>
  <ingest-processing>
    <identification>Yes</identification>
    <validation>Yes</validation>
    <characterization>Yes</characterization>
    <migration>No</migration>
  </ingest-processing>
</action-plan>
```

## Chapter 4: DAITSS Preservation Services

```
        <normalization>No</normalization>
        <xmlresolution>No</xmlresolution>
</ingest-processing>
<significant-properties>
    <content>TBD</content>
    <context>TBD</context>
    <behavior>TBD</behavior>
    <structure>TBD</structure>
    <appearance>TBD</appearance>
</significant-properties>
<long-term-strategy>
    <original>
Migrate to newer JP2 versions or to a format that shares the
essential and
distinguishing characteristics of the JP2 specification.
    </original>
</long-term-strategy>
<short-term-actions>
    <action>None</action>
</short-term-actions>
</action-plan>
```

## Action plan components

Action Plans contain the following components:

- **Meta Information:** information about the Action Plan itself, including the date it is due to be reviewed again.
- **Ingest Processing:** information about the actions that will be programmatically performed on the file format during the Ingest and Refresh tasks. The ingest processing section indicates whether these services should be performed:
  - **Identification:** determine the file format, based on published/known format signatures. The FDA uses the magic number (internal signature) embedded in the file.
  - **Validation:** validate the file's format against that format's specifications by parsing the digital object. Any anomalies in the object identified in this step are recorded in the DAITSS database, and are also reported in the Ingest Report as file warnings.
  - **Characterization:** extract the technical metadata from the file.
  - **Normalization:** transform the format into one that is perceived to be more stable and easier to preserve.
  - **Migration:** convert a file format subject to obsolescence to a successor format.

## Chapter 4: DAITSS Preservation Services

- **Significant Properties:** metadata about the properties of the file that will be extracted and verified to ensure the significant properties of the digital object are preserved during format migration
- **Long-term Preservation Strategy:** the current strategy for the long-term preservation of the file format should the file format be at risk of obsolescence. This may include migration to a newer version of the format or migration to a different format.
- **Timetable of Anticipated Short-term Actions:** short-term actions to implement the prescribed preservation strategy.
- **Timetable of Action Plan Review and Revisal:** the next anticipated review of this file format and revisal of the Action Plan, if necessary.

If a transformation (normalization or migration) is required, the Action Plan ingest processing section must contain the transformation identifier code that will allow the Transformation Service to look up the appropriate transformation processing instructions.

```
<normalization>Yes
    <transformation id="avi_norm"/>
</normalization>
```

If XML Resolution is required (see "DAITSS XML Resolution Service" below) the address of the Resolution Service must be provided.

```
<xmlresolution>
    http://localhost:7000/xmlresolution
</xmlresolution>
```

The Action Plans used by the FDA are distributed as part of DAITSS. Any institution implementing DAITSS can prepare Action Plans for the formats of interest to that repository, implementing the preservation strategies preferred by the institution. The preservation decisions made in the Action Plans used by the FDA are based on detailed analyses of the characteristics of each file format. These analyses are documented as "background reports" which are linked to from the FDA Formats Table on the [FDA website](#).

### DAITSS Format Transformation Service

The DAITSS Transformation Service executes a format transformation on a given file based on the transformation identifier provided by the Action Plan Service. The transformation identifier is used to look up the transformation instruction to perform the format transformation. Transformation instructions are stored in the configuration file, in the /opt/web-services/sites/transform/current/config/transform.xml on the DAITSS demonstration installation.

For additional details, please refer to the [DAITSS Transform Project](#)

## An example of the connection between the Action Plan and Transformation Services

```
<transformation ID='AVI_NORM'>
  <instruction>mencoder #INPUT_FILE# -oac pcm -ovc lavc -lavcopts vcodec=mjpeg -o #OUTPUT_FILE#</instruction>
  <extension>.avi</extension>
  <identifier>avi/norm/v0.1.1</identifier>
  <software>
    MEncoder SVN-r28728-snapshot-4.1.2 (C) 2000-2009 MPlayer Team
  </software>
</transformation>
```

The Transformation Service configuration file contains instructions on how to process the DAITSS Action Plan Service transformation identifier, as follows:

**<transformation ID='AVI\_NORM'>**: defines the AVI\_NORM transformation identifier.

**<instruction>**: is the command line that will be used to carry out the format transformation.

For example, "mencoder #INPUT\_FILE# -oac pcm -ovc lavc -lavcopts vcodec=mjpeg -o #OUTPUT\_FILE#" means the program mencoder will be used to perform the actual format transformation using the specified command line arguments.

It is assumed that the referenced tool, mencoder, is installed properly on the machine where the transformation service is running. INPUT\_FILE is used to specify the input of the transformation instruction and OUTPUT\_FILE is used to store the output of the format transformation.

**<extension>** : defines the file extension that will be used for the output file. This is to ensure the output file format can be properly identified and validated.

**<identifier>** : is the agent identifier for the transformation service, to be used to generate proper versioned PREMIS agent.

**<software>** : describes the software that is used for the format transformation. The software description is then used to generate PREMIS agent detail.

### DAITSS XML Resolution Service

The XML Resolution Service downloads any XML schema referred to in XML content files and creates a tarfile of all schema files used in the package. Retaining copies of the schemas in an AIP can aid in the long term understanding of the XML file as well as facilitate its future validation.

For example, the DAITSS SIP Descriptor is itself an XML document that is archived. As a METS document, it contains:

## Chapter 4: DAITSS Preservation Services

<METS:mets

xmlns:METS="<http://www.loc.gov/METS/> ..."

xsi:schemaLocation=<http://www.loc.gov/METS/> <http://www.loc.gov/standards/mets/mets.xsd>  
... .. >

The XML Resolution Service will attempt to download the METS schema (xsd file) and any other schema referenced as a namespace. It processes recursively, checking the downloaded files for internal schema references and resolving them as well.

Using a caching proxy such as Squid with this service is recommended to improve delivery speed and prevent the DAITSS implementation from overwhelming the server hosting the schema.