NORMALIZATION FOR PDF 1.X FILES:
A SPECIFICATION OF THE PDF_NORM_! FORMAT

Last Modification Date: 04/17/2003
Author: Andrea Goethals, FCLA

pdf_norm_1 format: A single XML file in METS format with one or more TIFF images

Methodology:
Each PDF page is converted to a separate 24-bit RGB 600dpi TIFF with no compression using a FCLA-written Java wrapper to GNU Ghostscript 7.05.5. A single XML file (“index”) is created that describes how the TIFFs relate structurally to the PDF. The XML file uses the METS 1.2 schema.

Example:
A 5-page PDF file (ABCD1233.pdf) is normalized to the pdf_norm_1 format. Six files are created: 1 XML document and 5 TIFF images.

index (ABCD1234.xml): (created when the PDF file is ingested into the archive, given a name to reflect its ID in the database)

```
<?xml version='1.0' encoding='UTF-8' standalone='no'?>
<METS:mets xmlns:METS="http://www.loc.gov/METS/"
xmlns:xsi=”http://www.w3c.org/2001/XMLSchema-instance”
xmlns:xlink=”http://www.w3c.org/TR/xlink
xsi:schemaLocation=”http://www.loc.gov/METS/
http://www.loc.gov/standards/mets/mets.xsd” OBJID=”ABCD1234”>
   <METS:fileSec>
      <METS:fileGrp>
         <METS:file ID=”ABCD1236” MIMETYPE=”image/tiff” SEQ=”1”>
            <METS:FLocat LOCTYPE=”OTHER” OTHERLOCTYPE=”DatabaseID”
                    xlink:type=”simple” xlink:href=”ABCD1236”/>
         </METS:file>
         <METS:file ID=”ABCD1237” MIMETYPE=”image/tiff” SEQ=”2”>
            <METS:FLocat LOCTYPE=”OTHER” OTHERLOCTYPE=”DatabaseID”
                    xlink:type=”simple” xlink:href=”ABCD1237”/>
         </METS:file>
         <METS:file ID=”ABCD1238” MIMETYPE=”image/tiff” SEQ=”3”>
            <METS:FLocat LOCTYPE=”OTHER” OTHERLOCTYPE=”DatabaseID”
                    xlink:type=”simple” xlink:href=”ABCD1238”/>
         </METS:file>
         <METS:file ID=”ABCD1239” MIMETYPE=”image/tiff” SEQ=”4”>
            <METS:FLocat LOCTYPE=”OTHER” OTHERLOCTYPE=”DatabaseID”
                    xlink:type=”simple” xlink:href=”ABCD1239”/>
         </METS:file>
         <METS:file ID=”ABCD1240” MIMETYPE=”image/tiff” SEQ=”5”>
            <METS:FLocat LOCTYPE=”OTHER” OTHERLOCTYPE=”DatabaseID”
                    xlink:type=”simple” xlink:href=”ABCD1240”/>
         </METS:file>
      </METS:fileGrp>
   </METS:fileSec>
   <METS:structMap TYPE=”PHYSICAL”>
      <METS:div ID=”div1” ORDER=”1” LABEL=”Page 1” TYPE=”page”>
         <METS:fptr FILEID=”ABCD1236”/>
      </METS:div>
      <METS:div ID=”div2” ORDER=”2” LABEL=”Page 2” TYPE=”page”>
         <METS:fptr FILEID=”ABCD1237”/>
      </METS:div>
      <METS:div ID=”div3” ORDER=”3” LABEL=”Page 3” TYPE=”page”>
         <METS:fptr FILEID=”ABCD1238”/>
      </METS:div>
      <METS:div ID=”div4” ORDER=”4” LABEL=”Page 4” TYPE=”page”>
         <METS:fptr FILEID=”ABCD1239”/>
      </METS:div>
      <METS:div ID=”div5” ORDER=”5” LABEL=”Page 5” TYPE=”page”>
         <METS:fptr FILEID=”ABCD1240”/>
      </METS:div>
   </METS:structMap>
</METS:mets>
```

Six TIFF images: (Created during the ingestion of the PDF)
ABCD1236.tif
ABCD1237.tif
ABCD1238.tif
ABCD1239.tif
ABCD1240.tif

schema (ABCD1235.xsd):
The METS 1.2 schema is available at: http://www.loc.gov/standards/mets/mets.xsd

Directory structure (immediately after PDF normalization):
Notes: Files created from the PDF normalization process will be put temporarily into a subdirectory of the package directory. Assume that the normalization directory exists and is empty before nomalization. Assume that the normalization directory is directly below the package directory, even if the PDF file exists in a subdirectory of the package directory. Any of the following results are valid normalization results:

```
[package directory]
                - ABCD1233.pdf (original file)
                - [normalization_directory]
                      - ABCD1234.xml (root object)
                      - ABCD1235.xsd (METS schema)
                      - ABCD1236.tif
                      - ABCD1237.tif
                      - ABCD1238.tif
                      - ABCD1239.tif
                      - ABCD1240.tif
```

```
[package directory]
                - [dir1]
                      - ABCD1233.pdf (original file)
                - [normalization_directory]
                      - ABCD1234.xml (root object)
                      - ABCD1235.xsd (METS schema)
                      - ABCD1236.tif
                      - ABCD1237.tif
```

```
                          - ABCD1238.tif
                          - ABCD1239.tif
                          - ABCD1240.tif

[package directory]
                    - [dir1]
                          - [dir2]
                                - ABCD1233.pdf (original file)
                    - [normalization_directory]
                          - ABCD1234.xml (root object)
                          - ABCD1235.xsd (METS schema)
                          - ABCD1236.tif
                          - ABCD1237.tif
                          - ABCD1238.tif
                          - ABCD1239.tif
                          - ABCD1240.tif
```

References:

Ghostscript Homepage
http://www.cs.wisc.edu/~ghost/

METS Metadata Encoding & Transmission Standard Official Web Site
http://www.loc.gov/standards/mets/