A Summary of HathiTrust METS File Creation Practice

By Chris Powell, Digital Library Production Service, University of Michigan
Last updated: January 28, 2009

Basically, a HathiTrust METS document has four parts – the metsHdr, the dmdSec, the amdSec, and the fileSec. Most of it is boilerplate structure, filled in with information gleaned from GRIN/GROOVE or the objects in a given directory at the time of loading.

Here are examples of sections of the METS object. Anything supplied but not explained – especially closing elements – is boilerplate that should be supplied by the creation routine.

ROOT ELEMENT:

```
<METS:mets xsi:schemaLocation="http://www.loc.gov/METS/ http://www.loc.gov/standards/mets/mets.xsd
http://purl.org/dc/elements/1.1/" OBJID="mdp.39015000484330" xmlns:METS="http://www.loc.gov/METS/"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xmlns:xlink="http://www.w3.org/1999/xlink"
xmlns:PREMIS="http://www.loc.gov/standards/premis">
```

- OBJID is the name of the directory within the MDP namespace.

metsHdr ELEMENT:

```
<METS:metsHdr ID="mdp.39015000484330" CREATEDATE="2008-04-26T07:14:03"
RECORDSTATUS="NEW">
```

- ID is the name of the directory preceded by mdp namespace. CREATEDATE is the timestamp for the creation of this METS file.

```
<METS:agent ROLE="CREATOR" TYPE="ORGANIZATION">
<METS:name>DLPS</METS:name>
</METS:agent>
</METS:metsHdr>
```

dmdSec ELEMENT AND mdREF SUBELEMENT:

```
<METS:dmdSec ID="DMD1">
  <METS:mdRef MDTYPE="MARC" LOCTYPE="OTHER" OTHERLOCTYPE="Item ID stored as second call
number in item record" XPTR="mdp.39015000484330" />
```

- We have chosen not to replicate the MARC in the METS object, but we could if we need to make DIPs to distribute, for example. The XPTR value is again the namespace and directory name.

```
</METS:dmdSec>
```

admSec ELEMENT AND techMD, digiprovMD, AND mdREF SUBELEMENTS:

```
<METS:amdSec>
<METS:techMD ID="TMD1">
<METS:mdRef LOCTYPE="OTHER" OTHERLOCTYPE="SYSTEM" MDTYPE="OTHER"
OTHERMDTYPE="text" LABEL="production notes" xlink:href="notes.txt"/>
```

- If there is a notes.txt file present, insert this element. If not, don't.

```
<METS:techMD ID="TMD2">
<METS:mdRef LOCTYPE="OTHER" OTHERLOCTYPE="SYSTEM" MDTYPE="OTHER"
OTHERMDTYPE="text" LABEL="md5checksums" xlink:href="checksum.md5"/>
```

- Theoretically, there will always be a checksum.md5 file present. That would make this element be considered boilerplate. However, it would be best to test and insert it only if present. Whether we reject or not based on missing checksums is another issue.

```
<METS:techMD ID="TMD3">
<METS:mdRef LOCTYPE="OTHER" OTHERLOCTYPE="SYSTEM" MDTYPE="OTHER"
OTHERMDTYPE="text" LABEL="page metadata" xlink:href="pagedata.txt"/>
```

- If there is a pagedata.txt file present, insert this element. If not, don't.

```
 <METS:techMD ID="premisobject1">
  <METS:mdWrap MDTYPE="PREMIS">
   <METS:xmlData>
    <PREMIS:object>
     <PREMIS:preservationLevel>1</PREMIS:preservationLevel>
    </PREMIS:object>
   </METS:xmlData>
  </METS:mdWrap>
```

- All our objects are the same preservationLevel, so this is boilerplate.

```
 <METS:digiprovMD ID="premisevent1">
  <METS:mdWrap MDTYPE="PREMIS">
   <METS:xmlData>
    <PREMIS:event>
     <PREMIS:eventIdentifier>
      <PREMIS:eventIdentifierValue>capture1</PREMIS:eventIdentifierValue>
     </PREMIS:eventIdentifier>
     <PREMIS:eventType>capture</PREMIS:eventType>
     <PREMIS:eventDateTime>2007-01-04T00:00:00</PREMIS:eventDateTime>
     <PREMIS:linkingAgentIdentifier>
      <PREMIS:linkingAgentIdentifierType>AgentID</PREMIS:linkingAgentIdentifierType>
      <PREMIS:linkingAgentIdentifierValue>Google, Inc.</PREMIS:linkingAgentIdentifierValue>
     </PREMIS:linkingAgentIdentifier>
    </PREMIS:event>
```

- EventIdentifierValues and eventTypes increment if they repeat. eventDateTime comes from GRIN date scanned. linkingAgentIdentifierValue is extracted from Artist tag in XMP metadata.

```
<PREMIS:event>
 <PREMIS:eventIdentifier>
  <PREMIS:eventIdentifierValue>compression1</PREMIS:eventIdentifierValue>
 </PREMIS:eventIdentifier>
 <PREMIS:eventType>compression</PREMIS:eventType>
 <PREMIS:eventDateTime>2007-02-07T16:12:00</PREMIS:eventDateTime>
 <PREMIS:linkingAgentIdentifier>
  <PREMIS:linkingAgentIdentifierType>AgentID</PREMIS:linkingAgentIdentifierType>
  <PREMIS:linkingAgentIdentifierValue>Google, Inc.</PREMIS:linkingAgentIdentifierValue>
 </PREMIS:linkingAgentIdentifier>
</PREMIS:event>
```

- EventIdentifierValues and eventTypes increment if they repeat. eventDateTime comes from GRIN date converted. linkingAgentIdentifierValue is extracted from Artist tag in XMP metadata.

```
<PREMIS:event>
 <PREMIS:eventIdentifier>
  <PREMIS:eventIdentifierValue>decryption1</PREMIS:eventIdentifierValue>
 </PREMIS:eventIdentifier>
 <PREMIS:eventType>decryption</PREMIS:eventType>
 <PREMIS:eventDateTime>2007-02-08T22:21:25</PREMIS:eventDateTime>
 <PREMIS:linkingAgentIdentifier>
  <PREMIS:linkingAgentIdentifierType>AgentID</PREMIS:linkingAgentIdentifierType>
  <PREMIS:linkingAgentIdentifierValue>UM</PREMIS:linkingAgentIdentifierValue>
 </PREMIS:linkingAgentIdentifier>
</PREMIS:event>
```

- eventDateTime comes from GROOVE.

```
<PREMIS:event>
 <PREMIS:eventIdentifier>
  <PREMIS:eventIdentifierValue>fixity check1</PREMIS:eventIdentifierValue>
 </PREMIS:eventIdentifier>
 <PREMIS:eventType>fixity check</PREMIS:eventType>
 <PREMIS:eventDateTime>2007-02-08T22:21:25</PREMIS:eventDateTime>
 <PREMIS:eventOutcomeInformation>
  <PREMIS:eventOutcomeDetail>pass</PREMIS:eventOutcomeDetail>
 </PREMIS:eventOutcomeInformation>
 <PREMIS:linkingAgentIdentifier>
  <PREMIS:linkingAgentIdentifierType>AgentID</PREMIS:linkingAgentIdentifierType>
  <PREMIS:linkingAgentIdentifierValue>UM</PREMIS:linkingAgentIdentifierValue>
 </PREMIS:linkingAgentIdentifier>
</PREMIS:event>
```

- There should be multiple occurrences of this event over time. eventDateTime and eventOutcomeDetail comes from GROOVE.

```
<PREMIS:event>
```

```
     <PREMIS:eventIdentifier>
      <PREMIS:eventIdentifierValue>ingestion1</PREMIS:eventIdentifierValue>
     </PREMIS:eventIdentifier>
     <PREMIS:eventType>ingestion</PREMIS:eventType>
     <PREMIS:eventDateTime>2007-02-08T22:21:25</PREMIS:eventDateTime>
     <PREMIS:linkingAgentIdentifier>
      <PREMIS:linkingAgentIdentifierType>AgentID</PREMIS:linkingAgentIdentifierType>
      <PREMIS:linkingAgentIdentifierValue>UM</PREMIS:linkingAgentIdentifierValue>
     </PREMIS:linkingAgentIdentifier>
    </PREMIS:event>
```

- eventDateTime comes from GROOVE.

```
    <PREMIS:event>
     <PREMIS:eventIdentifier>
      <PREMIS:eventIdentifierValue>message digest calculation1</PREMIS:eventIdentifierValue>
     </PREMIS:eventIdentifier>
     <PREMIS:eventType>message digest calculation</PREMIS:eventType>
     <PREMIS:eventDateTime>2007-02-07T16:12:00</PREMIS:eventDateTime>
     <PREMIS:eventDetail>jhove1_1e</PREMIS:eventDetail>
     <PREMIS:linkingAgentIdentifier>
      <PREMIS:linkingAgentIdentifierType>AgentID</PREMIS:linkingAgentIdentifierType>
      <PREMIS:linkingAgentIdentifierValue>Google, Inc.</PREMIS:linkingAgentIdentifierValue>
     </PREMIS:linkingAgentIdentifier>
    </PREMIS:event>
```

- eventDateTime comes from GRIN date converted. linkingAgentIdentifierValue is extracted from Artist tag in XMP metadata.

```
    <PREMIS:event>
     <PREMIS:eventIdentifier>
      <PREMIS:eventIdentifierValue>validation1</PREMIS:eventIdentifierValue>
     </PREMIS:eventIdentifier>
     <PREMIS:eventType>validation</PREMIS:eventType>
     <PREMIS:eventDateTime>2007-02-08T22:21:25</PREMIS:eventDateTime>
     <PREMIS:linkingAgentIdentifier>
      <PREMIS:linkingAgentIdentifierType>AgentID</PREMIS:linkingAgentIdentifierType>
      <PREMIS:linkingAgentIdentifierValue>UM</PREMIS:linkingAgentIdentifierValue>
     </PREMIS:linkingAgentIdentifier>
    </PREMIS:event>
```

- eventDateTime comes from GROOVE.

```
   </METS:xmlData>
  </METS:mdWrap>
</METS:amdSec>
```

fileSec ELEMENT AND fileGrp SUBELEMENTS:

```
<METS:fileSec>
  <METS:fileGrp ID="FG1" USE="zip archive">

  <METS:fileGrp ID="FG2" USE="image">
```

<METS:fileGrp ID="FG3" USE="ocr">

- The fileSec lists the files that comprise the digital object. HathiTrust METS objects have three fileGrps – one for the zip archive, one for the images, one for the OCR.

file ELEMENT and FLocat SUBELEMENTS:

<METS:file ID="ZIP00000001" MIMETYPE="application/zip" SEQ="00000001" CREATED="2008-04-26T07:14:03" SIZE="63248715" CHECKSUM="edb221de821042e5f96a629ac8e87970" CHECKSUMTYPE="MD5">
<METS:FLocat LOCTYPE="OTHER" OTHERLOCTYPE="SYSTEM" xlink:href="39015000484330.zip" />

- The METS:file ID, MIMETYPE, and SEQ are boilerplate for the zip file. CREATED is the timestamp off the zip file. SIZE is the file size of the zip file. CHECKSUM is pulled from the line in the checksum.md5 file pertaining to the zip file.

- The METS:FLocat xlink:href value is the zip file name.

<METS:file ID="IMG00000001" MIMETYPE="image/jpeg" SEQ="00000001" CREATED="2007-02-07T19:06:14" SIZE="191842" CHECKSUM="cfc5caa8039e756e0e0310da7d253190" CHECKSUMTYPE="MD5">
<METS:FLocat LOCTYPE="OTHER" OTHERLOCTYPE="SYSTEM" xlink:href="00000001.jpg"/>
</METS:file>

- The METS:file ID is the image name stripped of extension and preceded by IMG. The MIMETYPE is the extension preceded by image/ and adapted for syntactic correctness (tiff is another possibility). SEQ is the image name stripped of extension. CREATED is the timestamp off the image. SIZE is the file size of the image. CHECKSUM is pulled from the line in the checksum.md5 file pertaining to the particular image.

- The METS:FLocat xlink:href value is the image name.

<METS:file ID="TXT00000001" MIMETYPE="text/plain" SEQ="00000001" CREATED="2005-11-22T14:51:00" SIZE="0" CHECKSUM="cfc5caa8039e756e0e0310da7d25319" CHECKSUMTYPE="MD5">
<METS:FLocat LOCTYPE="OTHER" OTHERLOCTYPE="SYSTEM" xlink:href="00000001.txt"/>
</METS:file>

- The METS:file ID is the ocr file name stripped of extension and preceded by TXT. The MIMETYPE is text/plain . SEQ is the ocr file name stripped of extension. CREATED is the timestamp off the ocr file. SIZE is the file size of the ocr file. CHECKSUM is pulled from the checksum.md5 file.

- The METS:FLocat xlink:href value is the ocr file name.

</METS:fileGrp>
</METS:fileSec>

structMap ELEMENT and div SUBELEMENTS:

<METS:structMap ID="SM1" TYPE="physical">
<METS:div TYPE="volume">

- This is essentially a boilerplate wrapper, containing a div for the volume. The volume div contains a div subelement for each page with two fptr elements apiece – one each for the image and the OCR.

<METS:div ORDER="35" TYPE="page" ORDERLABEL="15" LABEL="CHAPTER_START, IMAGE_ON_PAGE, UNTYPICAL_PAGE">

- The ORDER is derived in the same manner as the SEQ, above. The TYPE is page (boilerplate). The ORDERLABEL is present if there is a number in the third position of the comma-delimited file pagedata.txt on the line whose first number corresponds to the ORDER value. The LABEL is present if that is a value in the fourth and following positions of the comma-delimited file pagedata.txt on the line whose first number corresponds to the ORDER value. For example, this is the pagedata.txt line for the div above:

35,13510798884466800,15,CHAPTER_START,IMAGE_ON_PAGE,UNTYPICAL_PAGE

<METS:fptr FILEID="IMG00000029" />
<METS:fptr FILEID="TXT00000029" />
</METS:div>