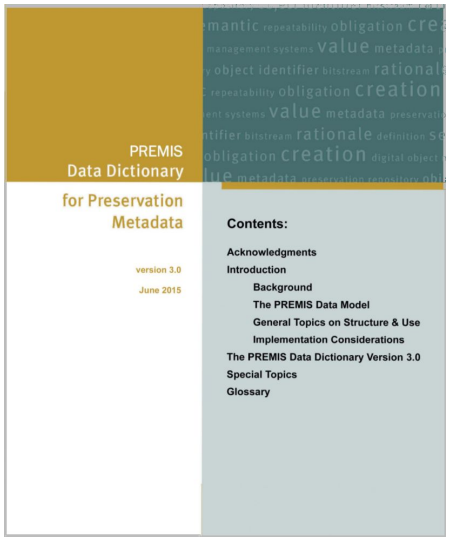
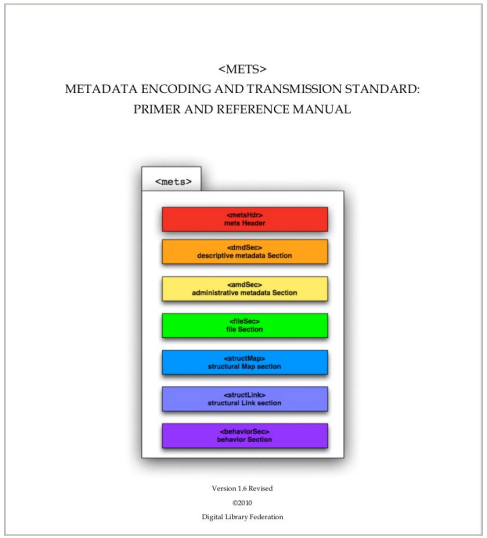
Metadata deep dive



+



=


My_AIP-d937bd1b-acd6-4e0c-b69f-51f1a2ec60e3
bag-info.txt
bagit.txt
manifest-sha512.txt
tagmanifest-md5.txt
data
logs
objects
thumbnails
METS.a934bd1v-sjd6-5f0c-b70f-51g2a2ec60t8.xml

What is PREMIS for?

- PREMIS, or Preservation Metadata Implementation Strategies, is the recognized standard for metadata about objects in a digital preservation system.
  - It captures technical information about an object in order to support the implementation of preservation strategies such as normalization, migration or emulation (PREMIS Object)
  - It describes relationships between digital objects (PREMIS Object)
  - It provides an audit trail of actions taken by the digital preservation repository to preserve the object (PREMIS Event)
  - It names the individuals, organizations and software tools responsible for taking actions to preserve digital objects (PREMIS Agent)
  - It specifies the actions a repository is allowed to take to preserve digital objects (PREMIS Rights)

What is METS for?

- METS, or Metadata Encoding and Transmission Standard, was designed to support inter-repository data exchange.
  - It provides a wrapper for other metadata, such as PREMIS and Dublin Core.
  - It defines relationships between digital objects and other digital objects, and between digital objects and their metadata.
  - It can be used to provide technical metadata about digital objects, although Archivematica doesn't implement it that way (we wrap PREMIS in it instead)

METS sections

<metsHdr> METS header

<dmdSec> Descriptive metadata

<amdSec> Administrative metadata

<fileSec> File section

<structMap> Structural Map

METS header

<metsHdr> METS header

METS dmdSec

<mets:dmdSec>
    <dcterms:dublincore>

METS amdSec

```
<mets:amdSec>
  <mets:techMD>
  <mets:rightsMD>
  <mets:digiprovMD>
  <mets:digiprovMD>
```

```
<mets:amdSec>
  <mets:techMD>
     PREMIS: OBJECT
  <mets:rightsMD>
     PREMIS: RIGHTS
  <mets:digiprovMD>
     PREMIS: EVENT
  <mets:digiprovMD>
     PREMIS: AGENT
```

METS fileSec

```
<mets:fileSec>
  <mets:fileGrp USE=“original”>
    <mets:file>
    <mets:file>
  <mets:fileGrp USE=“submissionDocumentation”>
    <mets:file>
    <mets:file>
  <mets:fileGrp USE=“preservation”>
    <mets:file>
    <mets:file>
```

METS structMap

<mets:structMap>
  <mets:Div> Directory
    <mets:Div> Directory
      <mets:Div> Item
      <mets:Div> Item
    <mets:Div> Directory
      <mets:Div> Item
      <mets:Div> Item

Mets header

<**mets**:metsHdr CREATEDATE="2016-03-28T23:15:51"/>

METS dmdSec

<mets:dmdSec ID="dmdSec_1"> The METS file can have more than one dmdSec

<mets:mdWrap MDTYPE="DC"> mdwrap means the metadata are included in the METS file, not referenced by it

<mets:xmlData>

<dcterms:dublincore xmlns:dcterms="http://purl.org/dc/terms/" xmlns:dc="http://purl.org/dc/elements/1.1/" xsi:schemaLocation="http://purl.org/dc/terms/ http://dublincore.org/schemas/xmls/qdc/2008/02/11/dcterms.xsd">

<dc:title>Pictures at an Exhibition</dc:title>

<dc:creator>Mussorgsky, Maxim<dc:creator>

….etc.

METS amdSec: techMD

```
<mets:amdSec ID="amdSec_1">
  <mets:techMD ID="techMD_1">
    <mets:mdWrap MDTYPE="PREMIS:OBJECT"> The PREMIS object entity semantic units go into METS techMD
      <mets:xmlData>
        <premis:object xmlns:premis="info:lc/xmlns/premis-v2" xsi:type="premis:file"
xsi:schemaLocation="info:lc/xmlns/premis-v2 http://www.loc.gov/standards/premis/v2/premis-v2-2.xsd" version="2.2">
          <premis:objectIdentifier>
            <premis:objectIdentifierType>UUID</premis:objectIdentifierType>
            <premis:objectIdentifierValue>bb52e3a0-ac5d-42c5...</premis:objectIdentifierValue>
                ….etc
```

METS amdSec: techMD con’t

```
<premis:objectCharacteristics>
  <premis:compositionLevel>0</premis:compositionLevel>
  <premis:fixity>
    <premis:messageDigestAlgorithm>sha256</premis:messageDigestAlgorithm>
      <premis:messageDigest>a58b87cd1c92881f95f41a35b8b2945970ebf0876b4f…..</premis:messageDigest>
  </premis:fixity>
  <premis:size>186012</premis:size>
  <premis:format>
    <premis:formatDesignation>
      <premis:formatName>Microsoft Word Document</premis:formatName>
      <premis:formatVersion>97-2003</premis:formatVersion>
    </premis:formatDesignation>
    <premis:formatRegistry>
      <premis:formatRegistryName>PRONOM</premis:formatRegistryName>
      <premis:formatRegistryKey>fmt/40</premis:formatRegistryKey>
    </premis:formatRegistry>
  </premis:format>
```

METS amdSec: techMD con't

<premis:objectCharacteristicsExtension>[raw tool output] </premis:objectCharacteristicsExtension> This is where technical metadata from ingested files go, having been extracted by tools like FIDO, Siegfried, Exiftool, MediaInfo, etc.

</premis:objectCharacteristics>

<premis:originalName>%transferDirectory%objects/letter.doc</premis:originalName>

<premis:relationship> This information shows a relationship between an ingested file and its normalized version, along with a relationship to the normalization Event

<premis:relationshipType>derivation</premis:relationshipType>

<premis:relationshipSubType>is source of</premis:relationshipSubType>

<premis:relatedObjectIdentification> This is the relationship to the related normalized file

<premis:relatedObjectIdentifierType>UUID</premis:relatedObjectIdentifierType>

<premis:relatedObjectIdentifierValue>b041d811-879f-4640-8ea5-82…</premis:relatedObjectIdentifierValue>

</premis:relatedObjectIdentification>

<premis:relatedEventIdentification> And this is the relationship to the normalization Event

<premis:relatedEventIdentifierType>UUID</premis:relatedEventIdentifierType>

<premis:relatedEventIdentifierValue>25ccf003-a007-4f12-be…</premis:relatedEventIdentifierValue>

</premis:relatedEventIdentification>

PREMIS Events in Archivematica

- Ingestion
- Message digest calculation
- Fixity check
- Name cleanup
- Unpacking
- Virus scan
- Format identification
- Validation
- Normalization
- Compression

METS amdSec: digivProvMD (Events)

```
<mets:digiprovMD ID="digiprovMD_8">
 <mets:mdWrap MDTYPE="PREMIS:EVENT">
  <mets:xmlData>
   <premis:event xmlns:premis="info:lc/xmlns/premis-v2" xsi:schemaLocation="info:lc/xmlns/premis-v2
http://www.loc.gov/standards/premis/v2/premis-v2-2.xsd" version="2.2">
    <premis:eventIdentifier>
     <premis:eventIdentifierType>UUID</premis:eventIdentifierType>
     <premis:eventIdentifierValue>0e7fb257-ae9c-47f2-b4d9….</premis:eventIdentifierValue>
    </premis:eventIdentifier>
    <premis:eventType>message digest calculation</premis:eventType>
    <premis:eventDateTime>2016-03-28T23:04:31</premis:eventDateTime>
    <premis:eventDetail>program="python"; module="hashlib.sha256()"</premis:eventDetail>
    ….etc.
```

METS amdSec: digivProvMD (Events) con’t

```
<premis:linkingAgentIdentifier>
  <premis:linkingAgentIdentifierType>Archivematica user pk</premis:linkingAgentIdentifierType>
  <premis:linkingAgentIdentifierValue>1</premis:linkingAgentIdentifierValue>
</premis:linkingAgentIdentifier>
<premis:linkingAgentIdentifier>
 <premis:linkingAgentIdentifierType>preservation system</premis:linkingAgentIdentifierType>
 <premis:linkingAgentIdentifierValue>Archivematica-1.5</premis:linkingAgentIdentifierValue>
</premis:linkingAgentIdentifier>
<premis:linkingAgentIdentifier>
    <premis:linkingAgentIdentifierType>repository code</premis:linkingAgentIdentifierType>
    <premis:linkingAgentIdentifierValue>NRI</premis:linkingAgentIdentifierValue>
</premis:linkingAgentIdentifier>
```

PREMIS Agents in Archivematica

- Institution
- Logged-in user
- Digital preservation system

METS amdSec: digivProvMD (Agents)

```
<mets:digiprovMD ID="digiprovMD_58">
  <mets:mdWrap MDTYPE="PREMIS:AGENT">
    <mets:xmlData>
      <premis:agent xmlns:premis="info:lc/xmlns/premis-v2" xsi:schemaLocation="info:lc/xmlns/premis-v2
http://www.loc.gov/standards/premis/v2/premis-v2-2.xsd" version="2.2">
        <premis:agentIdentifier>
          <premis:agentIdentifierType>repository code</premis:agentIdentifierType>
          <premis:agentIdentifierValue>NRI</premis:agentIdentifierValue>
        </premis:agentIdentifier>
        <premis:agentName>Not a Real Institution</premis:agentName>
        <premis:agentType>organization</premis:agentType>
      </premis:agent>
```

METS fileSec

<mets:fileSec>
<mets:fileGrp USE="original"> Identifies what the role of the file is in the context of this AIP
<mets:file GROUPID="Group-b041d811-879f-4640-8ea5-821920a81cf9" This is used to link this file to related files within the METS fileSec.
ID="file-b041d811-879f-4640-8ea5-821920a81cf9" ADMID="amdSec_2"> Note the link to the related amdSec, which has all of the PREMIS data in it
<mets:FLocat xlink:href="letter.doc" LOCTYPE="OTHER" OTHERLOCTYPE="SYSTEM"/> This shows where the file is located in relation to other files within the AIP
</mets:file>
<mets:file GROUPID="Group-002db941-78e1-4cbf-9bd9-afe7ef9c7466"
ID="file-002db941-78e1-4cbf-9bd9-afe7ef9c7466" ADMID="amdSec_4">
<mets:FLocat xlink:href="report.doc" LOCTYPE="OTHER" OTHERLOCTYPE="SYSTEM"/>
</mets:file>

METS fileSec con’t

```
<mets:fileGrp USE="preservation">
  <mets:file GROUPID="Group-b041d811-879f-4640-8ea5-821920a81cf9"
ID="file-ty41d811-879f-4640-8vj5-848573a84jf9" ADMID="amdSec_3">
    <mets:FLocat xlink:href="letter.pdf" LOCTYPE="OTHER" OTHERLOCTYPE="SYSTEM"/>
  </mets:file>
  <mets:file GROUPID="Group-002db941-78e1-4cbf-9bd9-afe7ef9c7466"
ID="file-002db941-78e1-4cbf-9bd9-afe7ef9c7466" ADMID="amdSec_5">
    <mets:FLocat xlink:href="report-f3d84155-3df1-427e-9ff8-5b480895372a.pdf" LOCTYPE="OTHER"
OTHERLOCTYPE="SYSTEM"/>
  </mets:file>
```

Remember the GROUPID on the previous slide?

METS fileSec con’t

<mets:fileGrp USE="submissionDocumentation"> Submission documentation gets its own fileGrp
<mets:file "Group-c01feb1d-cbcc-4a4b-a27a-3c1ba7618e99" ID="file-c01feb1d-cbcc-4a4b-a27a-3c1ba7618e99" ADMID="amdSec_7">
<mets:Flocat xlink:href="objects/submissionDocumentation/Transfer1-8e3003b8-a136-4bfc-b2f7-75d4fb637136/METS.xml” LOCTYPE="OTHER" OTHERLOCTYPE="SYSTEM"/>
</mets:file>
<mets:file GROUPID="Group-612db941-78e1-4cbf-9bd9-afe7ef9c7466" ID="file-612db941-78e1-4cbf-9bd9-afe7ef9c7466" ADMID="amdSec_9">
<mets:FLocat xlink:href="objects/submissionDocumentation/Transfer1-8e3003b8-a136-4bfc-b2f7-75d4fb637136/DeedOfGift.pdf” LOCTYPE="OTHER" OTHERLOCTYPE="SYSTEM"/>
</mets:file>

METS structMap

<mets:structMap ID="structMap_1" LABEL="Archivematica default" TYPE="physical"> The default Archivematica METS structMap provides a simple physical listing of the AIP's contents

<mets:div LABEL="Images-298af460-fdf4-4c78-ac8b-2f9266495f77" TYPE="Directory">

<mets:div LABEL="objects" TYPE="Directory" DMDID="dmdSec_1"> This is a link to the AIP's descriptive metadata

<mets:div LABEL="letter.doc" TYPE="Item">

<mets:fptr FILEID="file-b041d811-879f-4640-8ea5-821920a81cf9"/> fptr = file pointer, a link to the relevant entry in the fileSec

</mets:div>

<mets:div LABEL="letter-f3d84155-3df1-427e-9ff8-5b480895372a.pdf" TYPE="Item">

<mets:fptr FILEID="file-eb4a2422-93e2-4c70-ab4c-a56b4eeadab0"/>

</mets:div>

….etc.

</mets:div>

<mets:div LABEL="submissionDocumentation" TYPE="Directory">

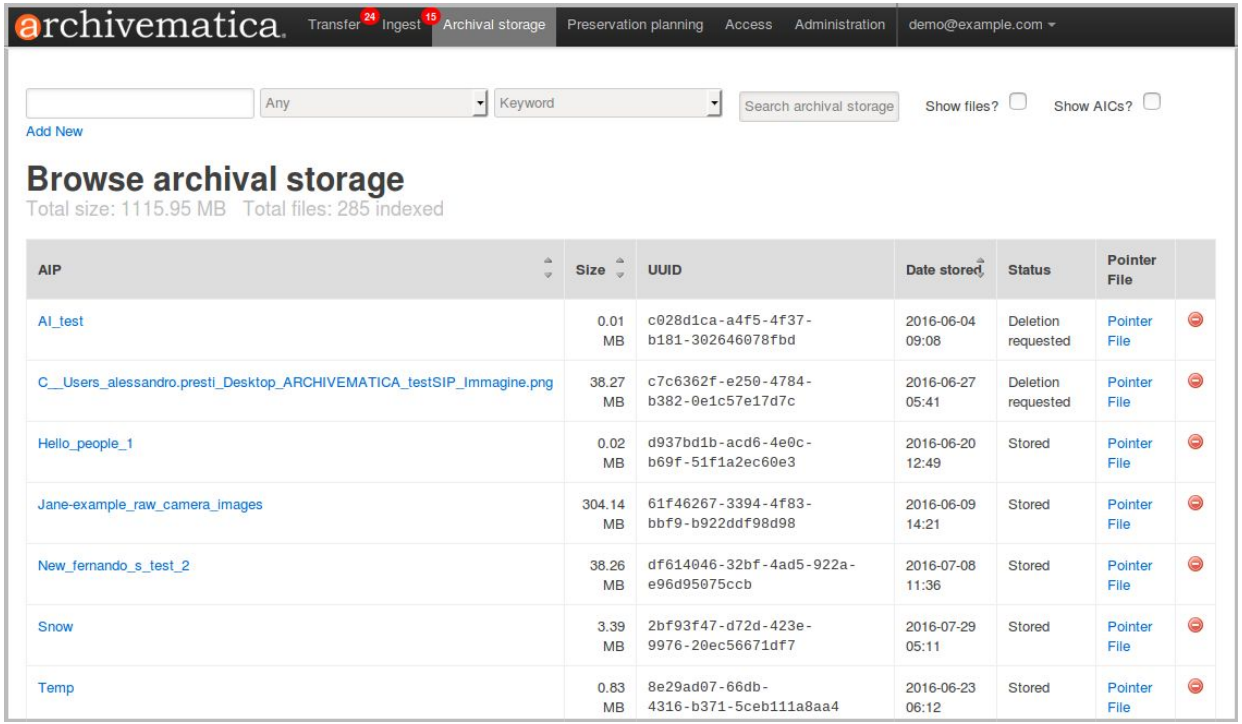
<mets:div LABEL="transfer1-540aec2f-9b01-463e-bf16-f12d6b58680c" TYPE="Directory">

<mets:div LABEL="DeedOfGift.pdf" TYPE="Item">

<mets:fptr FILEID="file-612db941-78e1-4cbf-9bd9-afe7ef9c7466"/>

….etc.

Another METS file - the AIP pointer file



What does the pointer file describe?

- The format of the AIP (e.g. 7zip format) (**PREMIS: OBJECT**)
- Information about the compression event (**PREMIS: EVENT**)
- The institution and preservation system that performed the compression (**PREMIS: AGENT**)
- The location of the AIP (**METS: fileGrp**)