Building a Robust Pipeline for ETD Ingestion with Rich Metadata

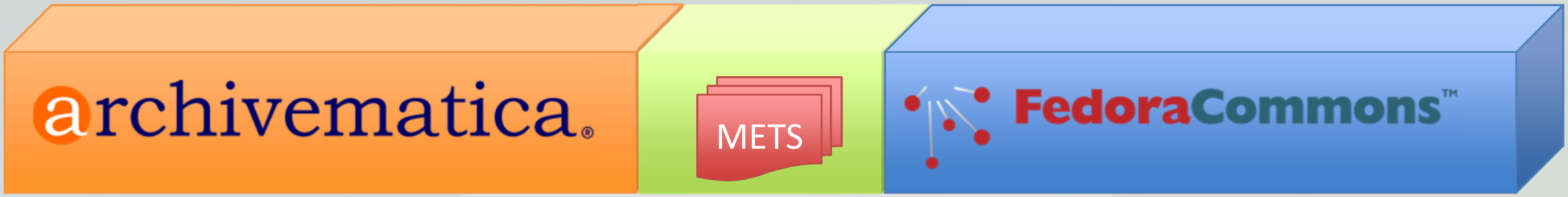
Lucas Mak, Aaron Collie, Devin Higgins
Michigan State University Libraries

Digital Preservation Interest Group, ALA Annual, Chicago, IL, June 30, 2013

Background

- Electronic submission for theses and dissertations at MSU
  - Began in 2010
  - Students submit ETD via ProQuest (www.etdadmin.com/grad.msu)
    - Metadata input by students (author, advisor, committee members, title, degree type, copyrights, embargo, academic unit/program, date, subject category, abstract, etc.)
    - Academic unit/program, and subject category from controlled lists
  - To be approved by Graduate School before releasing to ProQuest
- Agreement with ProQuest
  - Approved but unprocessed submissions sent to MSUL through FTP
    - ETD PDF file and student-supplied XML metadata in a single ZIP file
  - MSU can archive and display the unprocessed version

- Infrastructure
  - Archivematica
    - Micro-service design
      - Suite of tools to extract and generate technical and digital provenance metadata
      - PREMIS in METS
      - Package digital objects and associated metadata in BagIt format
  - Fedora Commons
    - Ingest digital objects and metadata as datastreams
    - Accept METS (Fedora extension) or FOXML

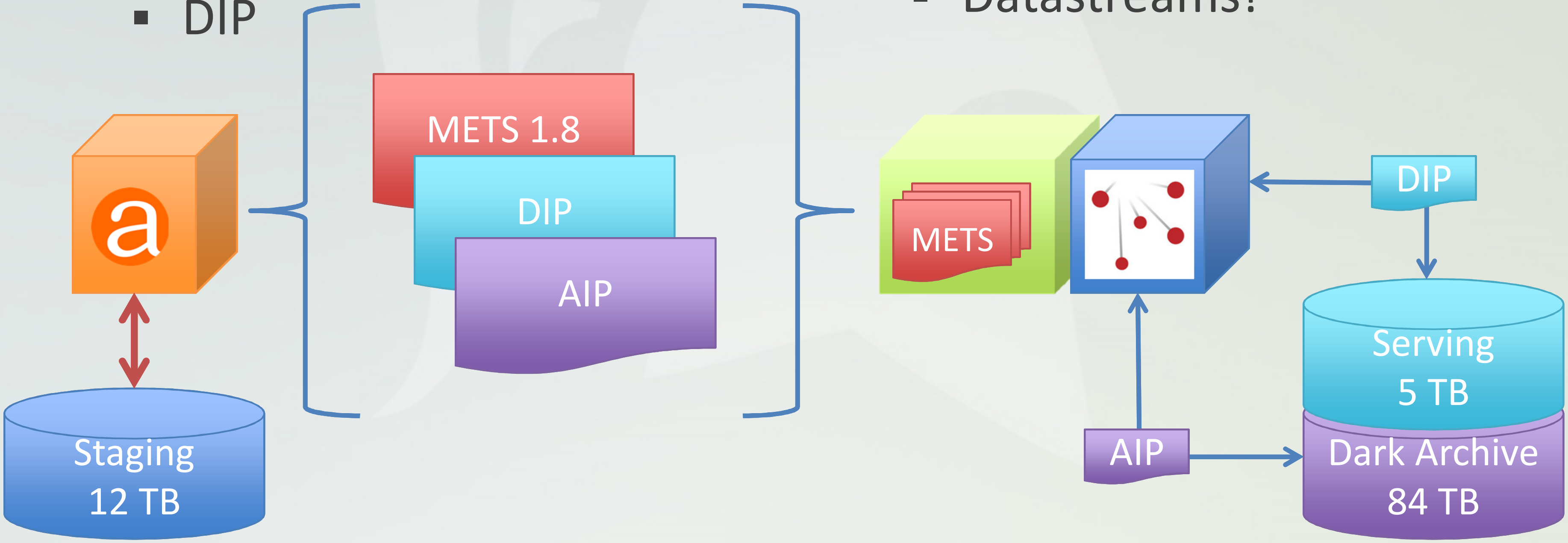


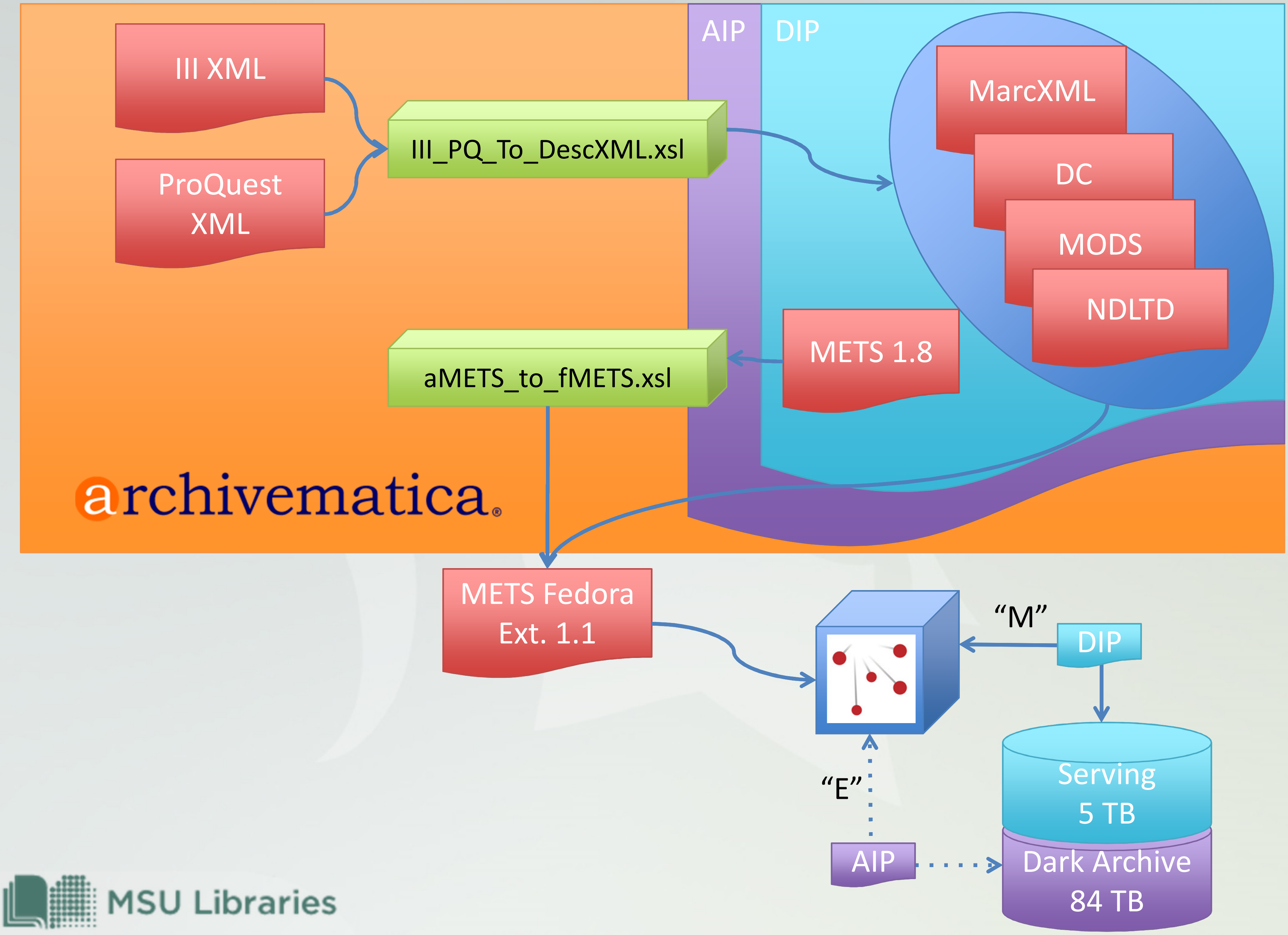
Archivematica Output:

- METS.xml
- AIP
- DIP

Fedora Commons Input:

- METS Fedora Extension
  - fedora-batch-ingest.sh
- Datastreams!





III XML

ProQuest XML

III_PQ_To_DescXML.xsl

AIP

DIP

MarcXML

DC

MODS

NDLTD

METS 1.8

aMETS_to_fMETS.xsl

archivematica®

METS Fedora Ext. 1.1

“M”

DIP

“E”

AIP

Serving 5 TB

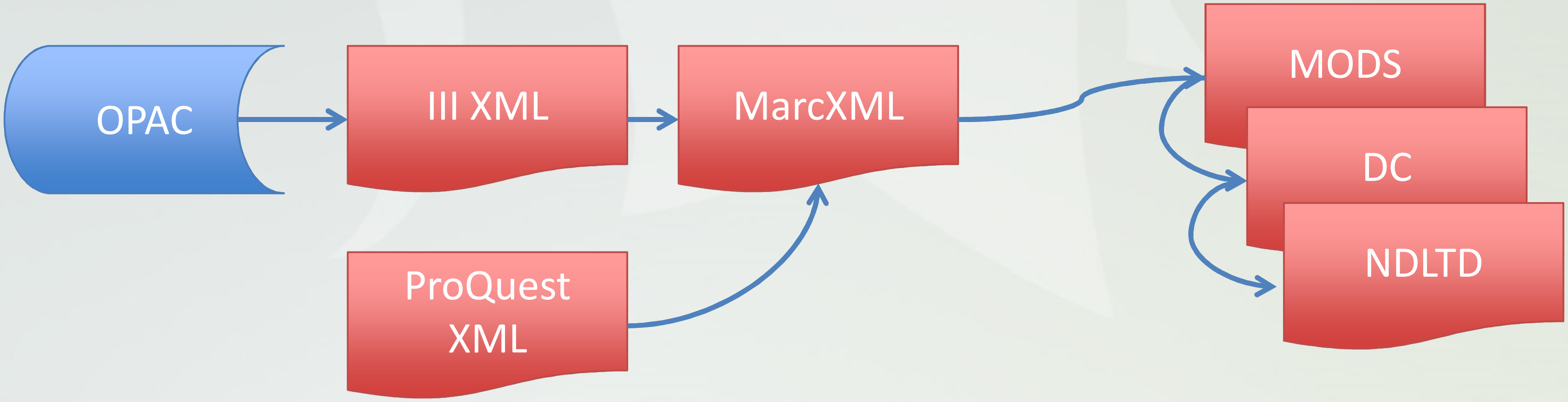
Dark Archive 84 TB

Descriptive Metadata

- Sources
  - MSU OPAC
    - Original cataloging done for MSU ETD
      - Match-and-derive from ProQuest MARC records
      - Enhanced with subjects and controlled names by catalogers
  - Student-supplied metadata (ProQuest XML)
- Targets
  - MarcXML (already exists in OPAC for some ETDs)
  - MODS (MSU-L preferred schema)
  - Dublin Core (required by OAI-PMH and Fedora)
  - NDLTD ETD-MS (international standard for ETD)

- Reconciliation process
  - Reuse OPAC data if available
    - LCSH and controlled names in MARC records
  - Pre-transformation processing
    - Query OPAC using data from ProQuest XML
    - If cataloged, get back III XML and be transformed into MarcXML
    - If not cataloged, create MarcXML directly from ProQuest XML”

- Transformation
  - MarcXML → MODS → DC → NDLTD ETD-MS
  - Unique data captured from ProQuest XML
    - Advisor, committee members, subject categories, copyrights/embargo info, abstract
    - Used to enhance MarcXML retrieved from OPAC

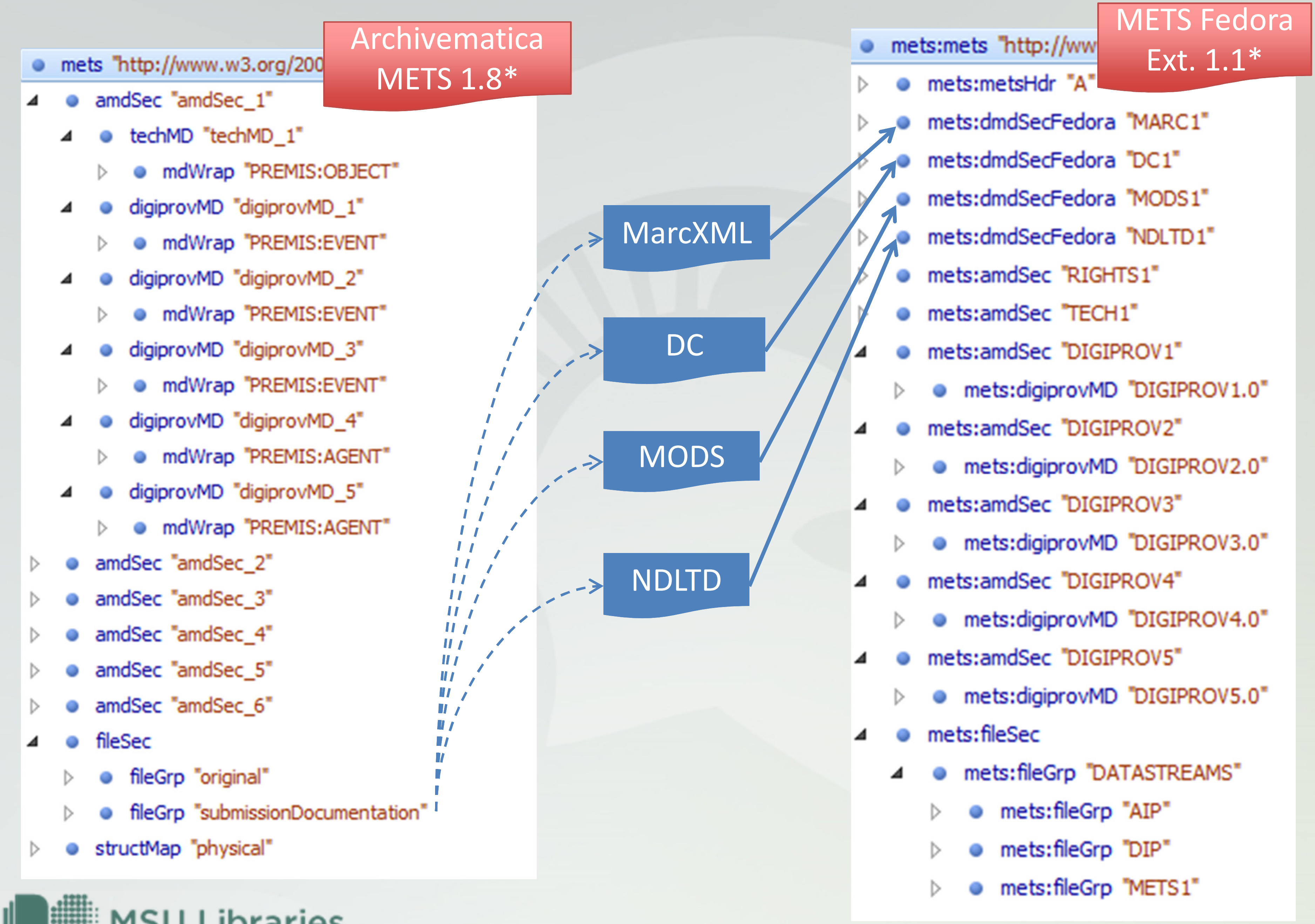


METS Package

- Archivematica METS output
  - PREMIS in METS: <PREMIS:object> (techMD), <PREMIS:event> (digiprovMD), <PREMIS:agent> (digiprovMD)
  - No descriptive or rights metadata
- Arrangement and availability of elements are different between Archivematica and Fedora METS
  - Different METS schema adopted
    - Archivematica: METS v. 1.8
    - Fedora: METS Fedora Extension 1.1
  - XSLT is used to rearrange METS elements and insert descriptive and rights metadata

Incorporation of Descriptive Metadata

- <METS:dmdSec>
  - Archivematica
    - Not used in ETD workflow
      - Only 1 Dublin Core record is allowed to describe the SIP
      - Has to be input through Archivematica GUI
    - MarcXML, MODS, DC, NDLTD ETD-MS records are included in "Submission documentation" folder
  - Fedora
    - Fedora extension element: <dmdSecFedora>
    - Allowed MDTYPE: MARC, EAD, DC, NISOIMG, LC-AV, VRA, TEI Header, DDI, FGDC, & OTHER
    - Copy XML files in "Submission documentation" folder into separate <dmdSecFedora>
      - MODS & NDLTD MS-ETD have to be labeled as "OTHER"
      - Use namespace URI to assign correct "MDTYPE" & "OTHERMDTYPE"



Archivematica METS 1.8*

```
mets "http://www.w3.org/200
◢ amdSec "amdSec_1"
    ◢ techMD "techMD_1"
        ▷ mdWrap "PREMIS:OBJECT"
    ◢ digiprovMD "digiprovMD_1"
        ▷ mdWrap "PREMIS:EVENT"
    ◢ digiprovMD "digiprovMD_2"
        ▷ mdWrap "PREMIS:EVENT"
    ◢ digiprovMD "digiprovMD_3"
        ▷ mdWrap "PREMIS:EVENT"
    ◢ digiprovMD "digiprovMD_4"
        ▷ mdWrap "PREMIS:AGENT"
    ◢ digiprovMD "digiprovMD_5"
        ▷ mdWrap "PREMIS:AGENT"
▷ amdSec "amdSec_2"
▷ amdSec "amdSec_3"
▷ amdSec "amdSec_4"
▷ amdSec "amdSec_5"
▷ amdSec "amdSec_6"
◢ fileSec
    ▷ fileGrp "original"
    ▷ fileGrp "submissionDocumentation"
▷ structMap "physical"
```

MarcXML

DC

MODS

NDLTD

METS Fedora Ext. 1.1*

```
mets:mets "http://ww
▷ mets:metsHdr "A"
▷ mets:dmdSecFedora "MARC1"
▷ mets:dmdSecFedora "DC1"
▷ mets:dmdSecFedora "MODS1"
▷ mets:dmdSecFedora "NDLTD1"
▷ mets:amdSec "RIGHTS1"
▷ mets:amdSec "TECH1"
◢ mets:amdSec "DIGIPROV1"
    ▷ mets:digiprovMD "DIGIPROV1.0"
◢ mets:amdSec "DIGIPROV2"
    ▷ mets:digiprovMD "DIGIPROV2.0"
◢ mets:amdSec "DIGIPROV3"
    ▷ mets:digiprovMD "DIGIPROV3.0"
◢ mets:amdSec "DIGIPROV4"
    ▷ mets:digiprovMD "DIGIPROV4.0"
◢ mets:amdSec "DIGIPROV5"
    ▷ mets:digiprovMD "DIGIPROV5.0"
◢ mets:fileSec
    ◢ mets:fileGrp "DATASTREAMS"
        ▷ mets:fileGrp "AIP"
        ▷ mets:fileGrp "DIP"
        ▷ mets:fileGrp "METS1"
```

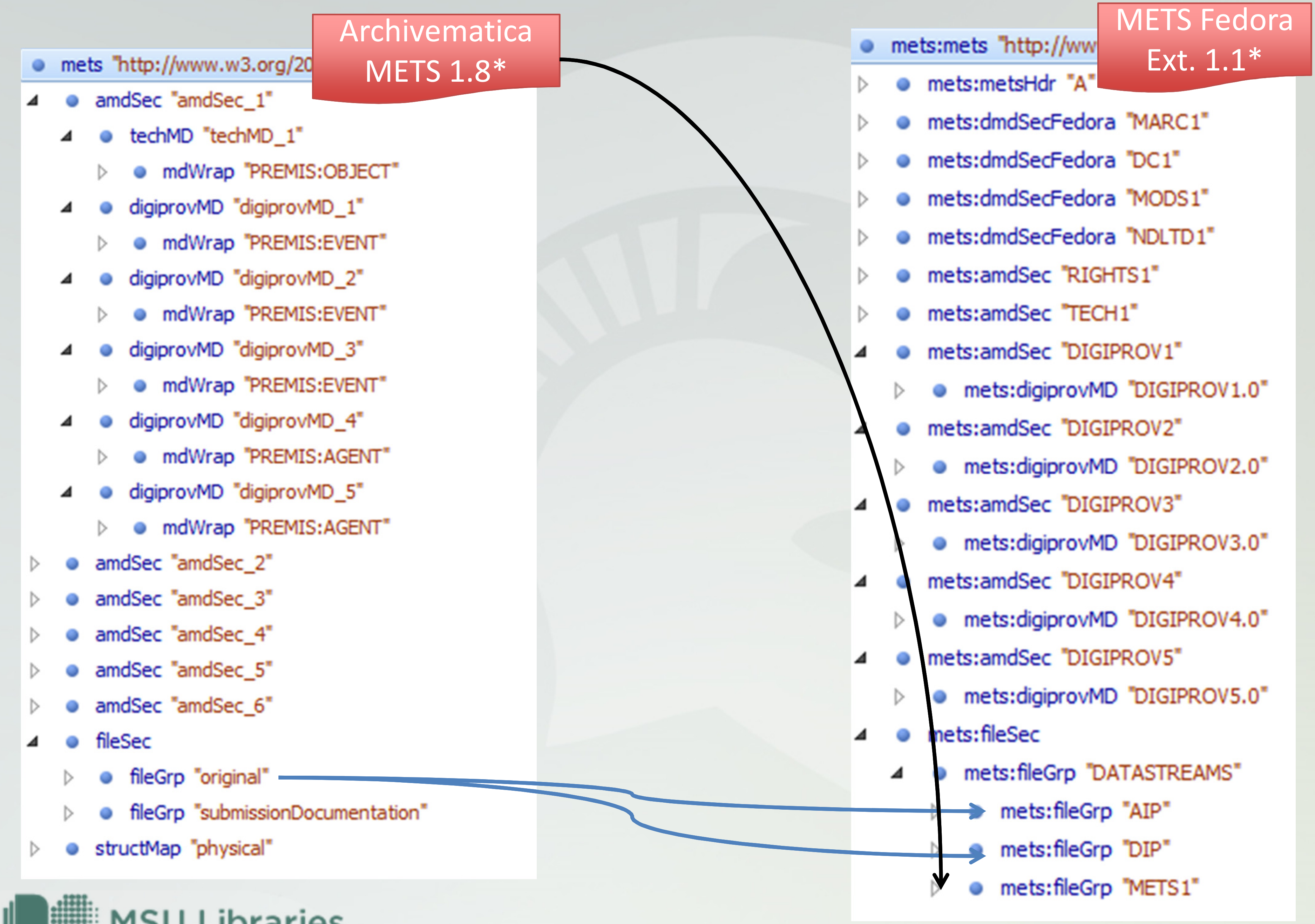
*Not a complete METS record

Incorporation of Rights Metadata

- Copyrights and embargo info available in ProQuest XML
- Read and parsed as variable during XSLT process
  - Create <PREMIS:rights> elements from copyrights and embargo info captured
    - If embargoed, include both start and end dates to enable automatic release of content for public display
  - Wrap <PREMIS:rights> under <METS:rightsMD> and then <METS:amdSec>

Rearrangement of METS Elements

- <METS:fileSec>
  - Archivematica
    - Two file groups: “Original” & “Submission documentation”
      - Original: digital objects
      - Submission documentation: descriptive metadata XML files
  - Fedora
    - Datastreams to be ingested as files
      - Files of digital objects and others (e.g. Archivematica METS)
    - Descriptive metadata XML files are ingested as “inline XML datastreams”
      - Copy all XML files in “Submission documentation” into separate <dmdSecFedora> elements



Archivematica METS 1.8*

```
● mets "http://www.w3.org/20
◢  ● amdSec "amdSec_1"
   ◢  ● techMD "techMD_1"
      ▷  ● mdWrap "PREMIS:OBJECT"
   ◢  ● digiprovMD "digiprovMD_1"
      ▷  ● mdWrap "PREMIS:EVENT"
   ◢  ● digiprovMD "digiprovMD_2"
      ▷  ● mdWrap "PREMIS:EVENT"
   ◢  ● digiprovMD "digiprovMD_3"
      ▷  ● mdWrap "PREMIS:EVENT"
   ◢  ● digiprovMD "digiprovMD_4"
      ▷  ● mdWrap "PREMIS:AGENT"
   ◢  ● digiprovMD "digiprovMD_5"
      ▷  ● mdWrap "PREMIS:AGENT"
▷  ● amdSec "amdSec_2"
▷  ● amdSec "amdSec_3"
▷  ● amdSec "amdSec_4"
▷  ● amdSec "amdSec_5"
▷  ● amdSec "amdSec_6"
◢  ● fileSec
   ▷  ● fileGrp "original"
   ▷  ● fileGrp "submissionDocumentation"
▷  ● structMap "physical"
```

METS Fedora Ext. 1.1*

```
● mets:mets "http://ww
▷  ● mets:metsHdr "A"
▷  ● mets:dmdSecFedora "MARC1"
▷  ● mets:dmdSecFedora "DC1"
▷  ● mets:dmdSecFedora "MODS1"
▷  ● mets:dmdSecFedora "NDLTD1"
▷  ● mets:amdSec "RIGHTS1"
▷  ● mets:amdSec "TECH1"
◢  ● mets:amdSec "DIGIPROV1"
   ▷  ● mets:digiprovMD "DIGIPROV1.0"
◢  ● mets:amdSec "DIGIPROV2"
   ▷  ● mets:digiprovMD "DIGIPROV2.0"
◢  ● mets:amdSec "DIGIPROV3"
   ▷  ● mets:digiprovMD "DIGIPROV3.0"
◢  ● mets:amdSec "DIGIPROV4"
   ▷  ● mets:digiprovMD "DIGIPROV4.0"
◢  ● mets:amdSec "DIGIPROV5"
   ▷  ● mets:digiprovMD "DIGIPROV5.0"
◢  ● mets:fileSec
   ◢  ● mets:fileGrp "DATASTREAMS"
      ▷  ● mets:fileGrp "AIP"
      ▷  ● mets:fileGrp "DIP"
      ▷  ● mets:fileGrp "METS1"
```

*Not a complete METS record

- <METS:amdSec>
  - Archivematica: Hierarchical structure

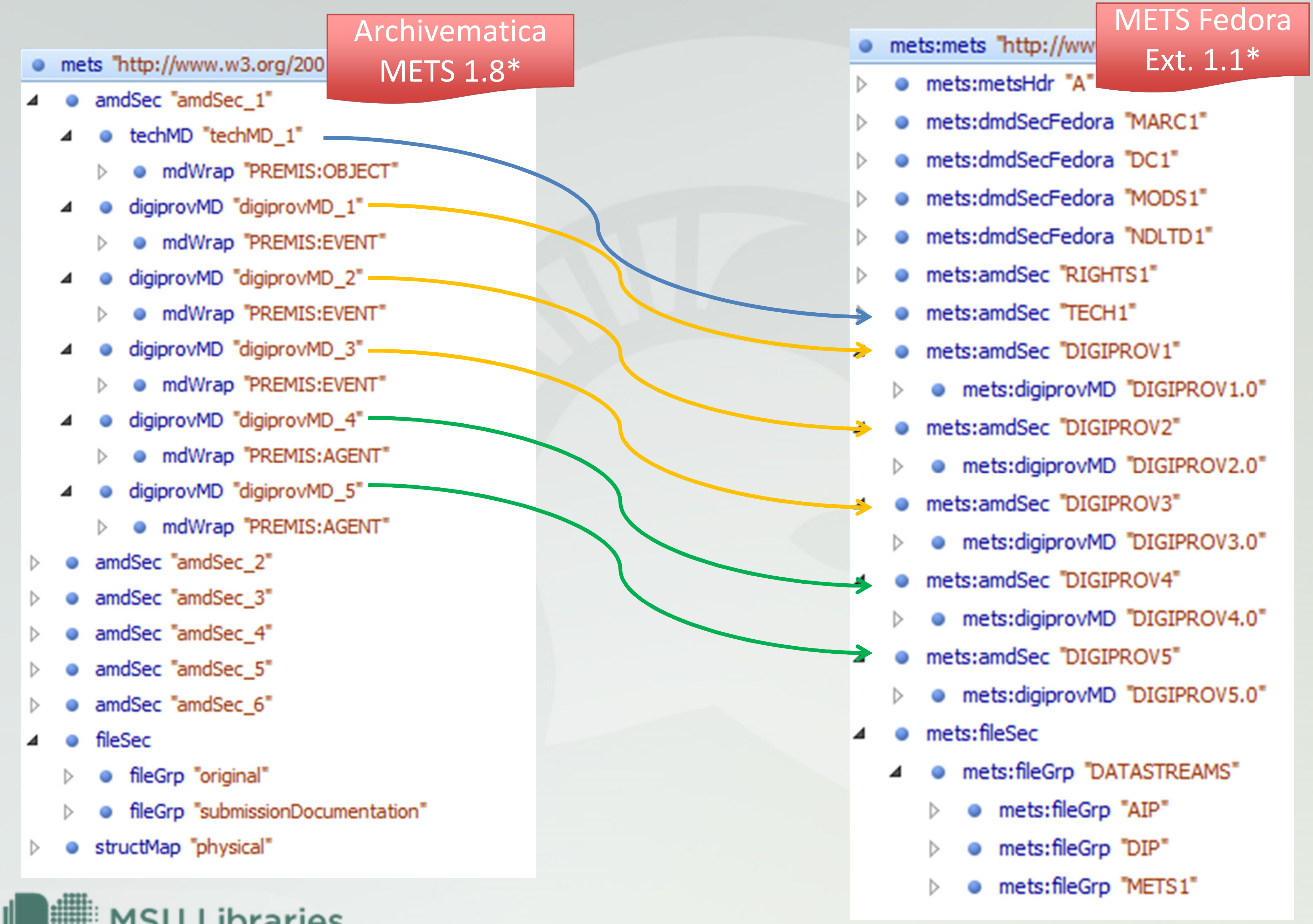
```
<amdSec ID=“amdSec1”>
        <techMD ID=“techMD1”/>
        …
        <digiProvMD ID=“digiProvMD1”/>
</amdSec>
<amdSec ID=“amdSec2”>
        <techMD ID=“techMD2”/>
        …
        <digiProvMD ID=“digiProvMD2”/>
</amdSec>
```

- 1 digital file has 1 <amdSec>
- All <techMD>, <rightsMD>, <sourceMD> and <digiProvMD> pertaining to the same file are nested under the same <amdSec>

- Fedora: Flat structure

```
<amdSec ID=“tech1”>
  <techMD ID=“tech1.0”/>
</amdSec>
<amdSec ID=“digiProv1”>
  <digiProvMD ID=“digiProv1.0”/>
</amdSec>
<amdSec ID=“tech2”>
  <techMD ID=“tech2.0”/>
</amdSec>
```

  - 1 digital file has mulitple <amdSec>
  - Each <techMD>, <rightsMD>, <sourceMD>, or <digiProvMD> is under its own <amdSec> to allow datastream versioning



Archivematica METS 1.8*

```
mets "http://www.w3.org/200
◢ amdSec "amdSec_1"
    ◢ techMD "techMD_1"
        ▷ mdWrap "PREMIS:OBJECT"
    ◢ digiprovMD "digiprovMD_1"
        ▷ mdWrap "PREMIS:EVENT"
    ◢ digiprovMD "digiprovMD_2"
        ▷ mdWrap "PREMIS:EVENT"
    ◢ digiprovMD "digiprovMD_3"
        ▷ mdWrap "PREMIS:EVENT"
    ◢ digiprovMD "digiprovMD_4"
        ▷ mdWrap "PREMIS:AGENT"
    ◢ digiprovMD "digiprovMD_5"
        ▷ mdWrap "PREMIS:AGENT"
▷ amdSec "amdSec_2"
▷ amdSec "amdSec_3"
▷ amdSec "amdSec_4"
▷ amdSec "amdSec_5"
▷ amdSec "amdSec_6"
◢ fileSec
    ▷ fileGrp "original"
    ▷ fileGrp "submissionDocumentation"
▷ structMap "physical"
```

METS Fedora Ext. 1.1*

```
mets:mets "http://ww
▷ mets:metsHdr "A"
▷ mets:dmdSecFedora "MARC1"
▷ mets:dmdSecFedora "DC1"
▷ mets:dmdSecFedora "MODS1"
▷ mets:dmdSecFedora "NDLTD1"
▷ mets:amdSec "RIGHTS1"
▷ mets:amdSec "TECH1"
◢ mets:amdSec "DIGIPROV1"
    ▷ mets:digiprovMD "DIGIPROV1.0"
◢ mets:amdSec "DIGIPROV2"
    ▷ mets:digiprovMD "DIGIPROV2.0"
◢ mets:amdSec "DIGIPROV3"
    ▷ mets:digiprovMD "DIGIPROV3.0"
◢ mets:amdSec "DIGIPROV4"
    ▷ mets:digiprovMD "DIGIPROV4.0"
◢ mets:amdSec "DIGIPROV5"
    ▷ mets:digiprovMD "DIGIPROV5.0"
◢ mets:fileSec
    ◢ mets:fileGrp "DATASTREAMS"
        ▷ mets:fileGrp "AIP"
        ▷ mets:fileGrp "DIP"
        ▷ mets:fileGrp "METS1"
```

*Not a complete METS record

- <AMDID> attribute in <mets:file>
  - Archivematica
    - Pointing to one <amdSec>, which has <techMD>, <rightsMD>*, <sourceMD>^, and <digiProvMD> nested within, per file
      - <mets:file ID="file1" AMDID="amdSec1"/>
  - Fedora
    - Pointing to multiple <amdSec>, each of which contains <techMD>, <rightsMD>, <sourceMD>^, or <digiProvMD>, per file
      - <mets:file ID="file1" AMDID= "tech1 rights1 digiProv1"/>

*<rightsMD> to be input in post-Archivematica process
^<sourceMD> is not used

Hardcoding Missing METS Elements

- <METS:metsHdr>
  - Archivematica
    - Does not use (optional in METS schema)
  - Fedora
    - <RECORDSTATUS> attribute to indicate whether the object is “active”, “inactive” or “deleted”
    - Hard-coding in with constant data

```
<mets:metsHdr RECORDSTATUS="A">
   <mets:agent ROLE="IPOWNER" TYPE="ORGANIZATION">
       <mets:name>MSU Libraries Digital and Multimedia Center</mets:name>
   </mets:agent>
 </mets:metsHdr>
```

- <OWNERID> attribute in <METS:file>
  - Archivematica
    - Does not use (optional in METS schema)
  - Fedora
    - To indicate whether the file is "managed by Fedora internally", "externally referenced", or "redirected"
      - Though optional according to Fedora-METS schema
    - Determine based on filename or file format
      - Archivematica add "checksum" into filename for files generated during the preservation workflow

	Origin	Strategy
III XML	Output from III XML Server	Not Stored. Proprietary format. Transformed into MarcXML
ProQuest XML	Received from ProQuest	Not Stored. Contains confidential/sensitive information. Transformed into MarcXML
MarcXML	Generated from III XML & ProQuest XML	Stored as MarcXML datastream. Widely used, highly descriptive standard
MODS	Derived from MarcXML	Stored as MODS datastream. Preferred standard for MSU-L collections
DC	Derived from MODS	Stored as required DC datastream. Used by Fedora Commons and OAI-PMH
NDLTD	Derived from DC	Stored as NDLTD datastream. Collection specific XML
METS 1.8	Output from Archivematica	Stored as METS datastream. Describes technical, administrative, preservation and provenance
FOXML	Generated by Fedora Commons	Stored on filesystem. Describes component parts of digital objects

Questions?

- Lucas Mak (makw@msu.edu)
- Aaron Collie (collie@msu.edu)
- Devin Higgins (higgi135@msu.edu)