

Bringing Many Tools Together

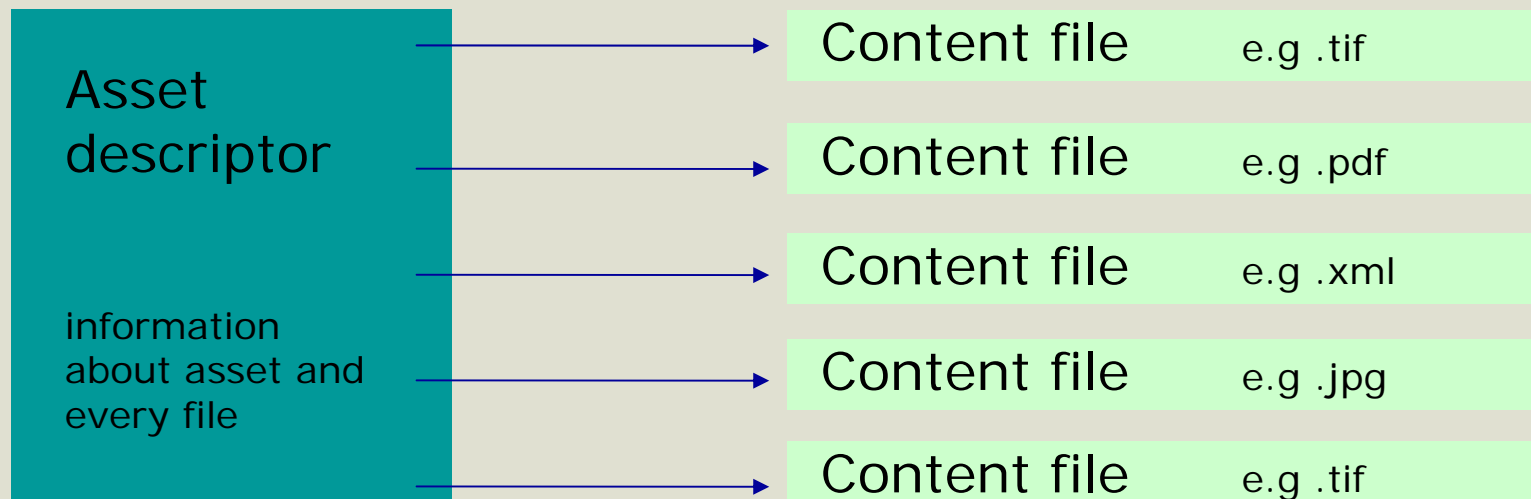
Building a system of co-operating OAIS's
in the MathArc project

Markus Enders
William Kehoe
Adam Smith

iPRES 2006
October 10, 2006

MathArc document model

Asset - used as transmission packet



Cornell University Library's OAIS

Data management

- MySQL

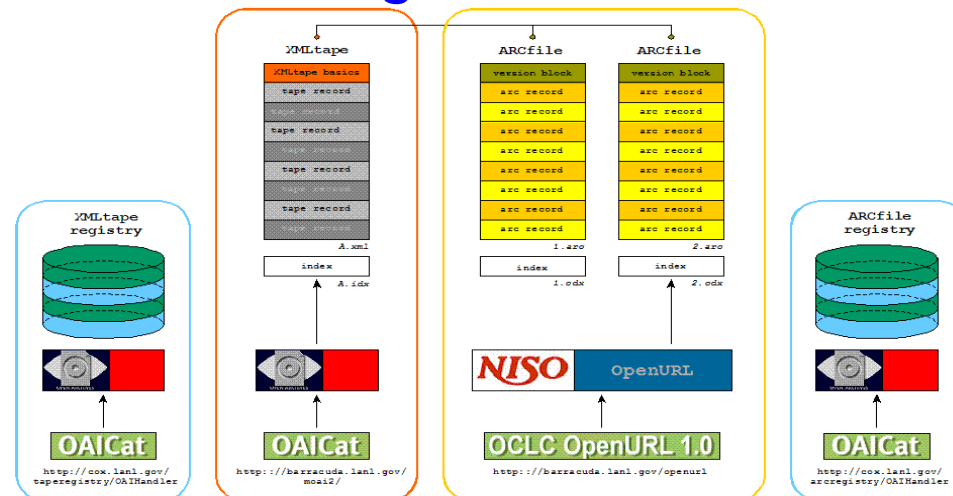
Access

- Django

Ingest

- Java
- METS
- PREMIS
- OAI_DC
- OAI-PMH
- JHOVE
- XMLBeans
- iBatis

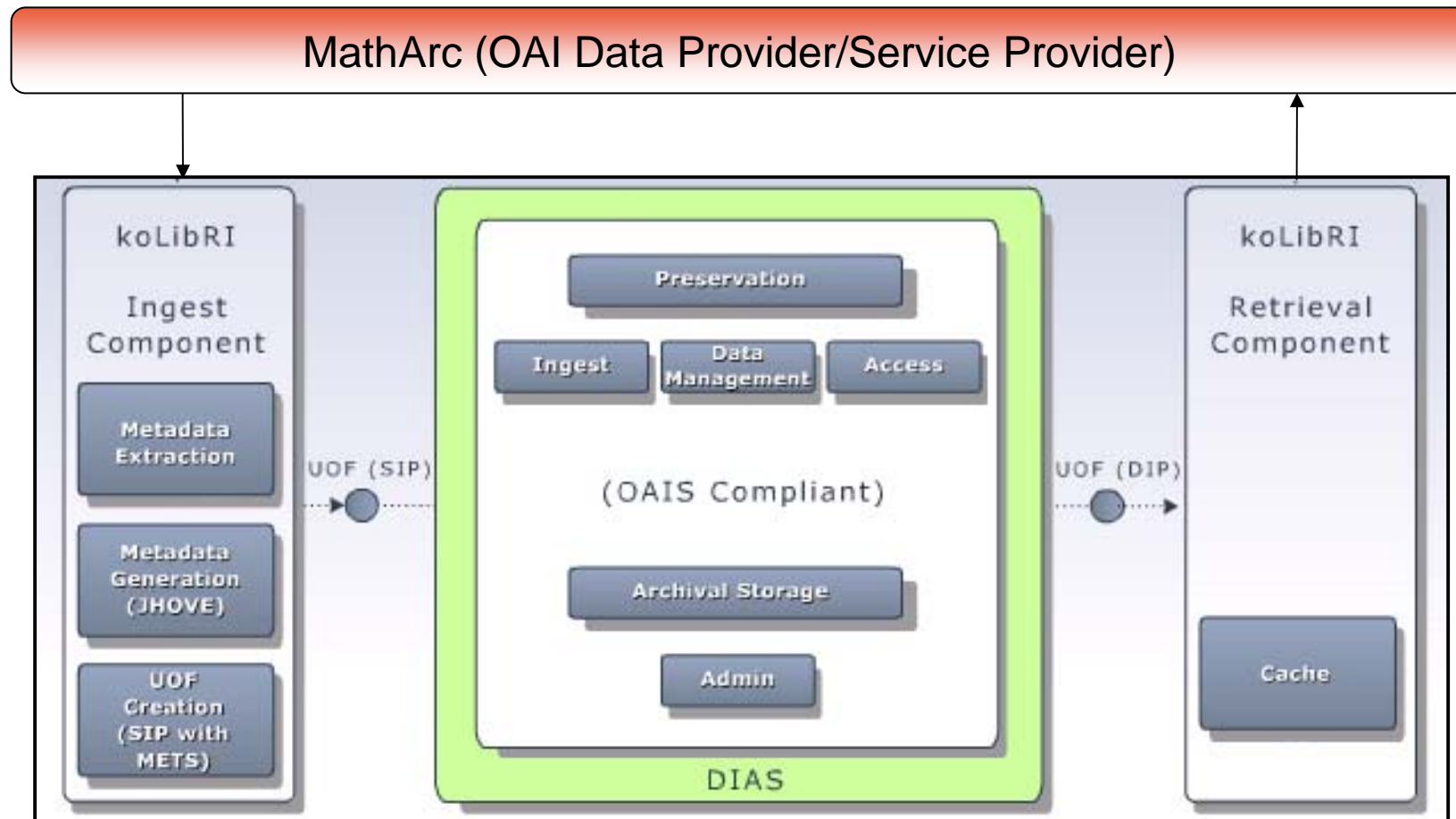
Archival Storage: aDORe



Handle System used instead of OpenURLs

- exposed as a Web Service using Apache Axis in Tomcat
- Aspects for logging, authentication

Göttingen State and University Library's OAIS



MathArc

METS based metadata

METS:

one structMap with a single <div> for
the asset

```
<structMap TYPE="ASSET">  
  <div ORDER="1" LABEL="File list" DMDID="DMD0 DMD1" TYPE="ASSET">  
    </div>  
</structMap>
```

MathArc

METS based metadata

METS:

single FileSet with a single fileGrp

this fileGrp must contain all files for
this asset

Technical metadata which are available
for all files are stored in METS<file>-
attributes

MathArc

METS based metadata

Size

Creation date

Checksum / checksum type

Mime type

```
<fileGrp ID="ASSET" ADMID="TechMD-LMER-Object">  
  <file ID="FILE0001" MIMETYPE="application/pdf"  
    SIZE="984058" CREATED="2003-01-12T11:55:26"  
    CHECKSUM="bf28a5b1abfa00f83f82ce68eab0123299408f29"  
    CHECKSUMTYPE="SHA-1" ADMID="TECHMD1">  
    <FLocat LOCTYPE="URL" xlink:type="simple"  
      xlink:href="file:///./diss.pdf"/>  
  </file>  
</fileGrp>
```

MathArc

METS based metadata

File type specific technical metadata is stored
in <techMd> section

e.g Mix for still images / JHOVE

Digital provenance metadata is stored
in <digiProv> section

PREMIS (mandatory)

Can contain additional metadata schema; not
used actively

MathArc

METS based metadata

A single `<amdMD>` section for
the asset
for each file

This section contains:

`<digiProvMD>`
and `<techMD>` sections

MathArc

METS based metadata

Further structural data (<div> beside the asset) aren't stored in the asset descriptor

Additional information must be stored in separate METS file which can be included in asset as content file

MathArc

PREMIS based metadata

on Asset level: stored in premis:object

objectIdentifier

purl / urn of the asset

preservationLevel always set to "0"

objectCategory always set to "representation"

creatingApplication

information about the software which created the
original SIP (imported to DIAS)

MathArc

PREMIS based metadata

on Asset level: stored in premis:object

Environment

information about the environment needed to render/interact the complex object (asset) – information about the content delivery platform

dependency: templates used in the DMS, configuration file of the DMS

software: DMS software, incl. swDependency ,

hardware: hardware the DMS is running on

MathArc

PREMIS based metadata

on file level: stored in premis:object

objectIdentifier

unique file id within the archive

preservationLevel always "0"

objectCategory always "file"

MathArc

PREMIS based metadata

objectCharacteristics

contains technical properties about the file:

<fixity>, <size> are not supported; use METS file-attributes instead

<format> is optional; contains format information from registries - empty element

PREMIS - migration

on asset level:

relationship

used to store a link to the old asset version;
link from a relationship to a related event

event

result of a concrete action, executed on a certain asset

eventIdentifier
eventType
eventDateTime
eventDetail

PREMIS - migration

on file level:

event

result of a concrete action, executed on a certain file

eventIdentifier

eventType

eventDateTime

eventDetail

agent

Linked to the event describing the software used for migration

no relationship used on file level

PREMIS - deletion

Asset descriptor (METS)
will contain no content files

eventIdentifier
eventType = deletion
eventDateTime
eventDetail

A "tombstone" is stored in the preservation systems and transmitted via OAI-PMH.

PREMIS – metadata update

In case only metadata has changed but content file(s) hasn't

eventIdentifier
eventType = updateAssetMetadata
eventDateTime
eventDetail

Premis file is similar to Migration event

PREMIS – Inconsistency

In case the data isn't consistent

- during transmission
checksums are checked after transmission
- in archive
regular checks of contentfiles and
checksums required

eventId
eventType = inconsistencyDiscovered
eventDateTime
eventDetail

Only a tombstone is transmitted (as in case of deletion event)

Harvesting Göttingen (SUB) assets from Cornell (CUL)

private LOCKSS Network

```
<partner>
  <identifierPrefix>SUB</identifierPrefix>
  <name>Niedersächsische Staats- und Universitätsbibliothek</name>
  <service_provider_IP>
    134.76.176.172
  </service_provider_IP>
  <data_provider_URL>
    http://gdz-srv3.sub.uni-goettingen.de:8081/matharc/
  </data_provider_URL>
</partner>
```

1. partners.xml request

2. OAI ListSets request

CUL

5. make OAI-PMH calls for assets

6. check local configuration (cul.xml)

7. process asset based on action item

4. available sets response

SUB

3. check local configuration (sub.xml)

```
<harvestedSets>
  <harvestedSet status="active" id="SUB:Mathematica">
    <latestHarvestDate partner_id="SUB">
      2007-08-01T12:00:00+01:00
    </latestHarvestDate>
    <action_item eventType="migration">
      ingestAsset
    </action_item>
    <action_item eventType="updateAssetMetadata">
      ingestAsset
    </action_item>
    <action_item eventType="replacement">
      ingestAsset
    </action_item>
    <action_item eventType="inconsistencyDiscovered">
      verifyAsset
    </action_item>
  </harvestedSet>
</harvestedSets>
```

```
<accessinformation partner="SUB">
  <ownedSets>
    <ownedSet id="SUB:Mathematica">
      <allowed_partners>
        <partner_identifierPrefix status="active" shareable="yes">
          CUL
        </partner_identifierPrefix>
      </allowed_partners>
    </ownedSet>
  </ownedSets>
</accessinformation>
```

Thank you

MathArc:

<http://www.library.cornell.edu/dlit/MathArc>