Archive Ingest and Handling Test: ODU's Perspective

Michael L. Nelson

Department of Computer Science

Old Dominion University

http://www.cs.odu.edu/~mln/

DLF Fall Forum 2005, Charlottesville VA, November 7 2005

Fortress Model

Five Easy Steps for Preservation:

1. Get a lot of $
2. Buy a lot of disks, machines, tapes, etc.
3. Hire an army of staff
4. Load a small amount of data
5. “Look upon my archive ye Mighty, and despair!”

image from: http://www.itunisie.com/tourisme/excursion/tabarka/images/fort.jpg

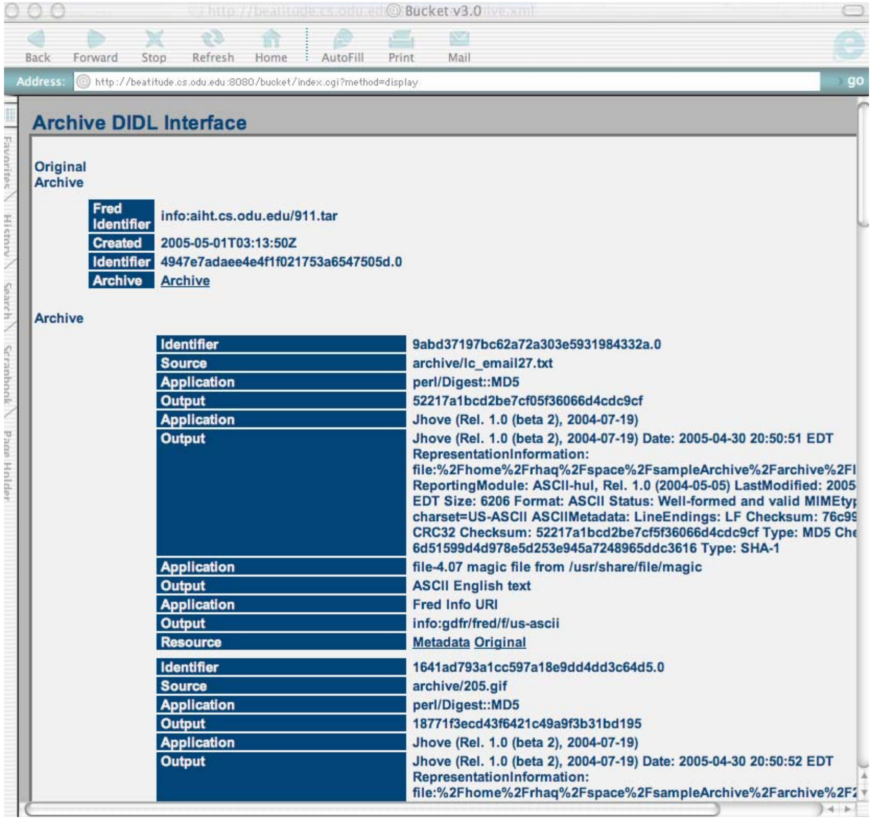
ODU’s Research Goals

- We’re in the CS department, not the library
  - Less infrastructure (bad)
  - More freedom (good)
- Interested in repository/object interaction
  - Long-range vision: repositories fade away; objects are responsible for their own preservation
  - Could we accomplish this with our “bucket” technology?
    - Significant questions about archive granularity
    - Transition to MPEG-21 Digital Item Declaration Language (DIDL) based buckets
- New models for digital preservation?

Buckets

- Buckets: self-contained, web-accessible objects
  - Grew out of research for serving NASA documents, esp. NACA Reports
    - http://naca.larc.nasa.gov/
    - http://doi.acm.org/10.1145/374308.374342
  - implicit assumptions:
    - 1 bucket = 1 logical item (N physical items)
    - Display is for human use
    - Bucket contents are DOM-parsable

Which Interface?



Address: http://beatitude.cs.odu.edu:8080/bucket/index.cgi?method=display

Archive DIDL Interface

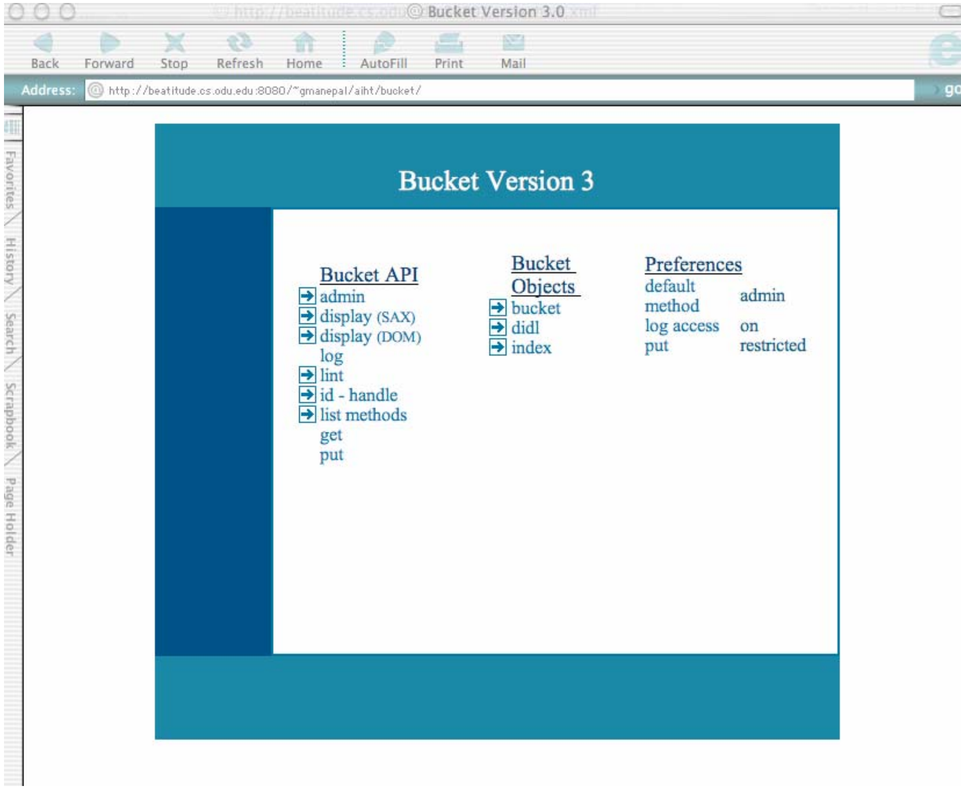
Original Archive

Fred Identifier	info:aiht.cs.odu.edu/911.tar
Created	2005-05-01T03:13:50Z
Identifier	4947e7adaee4e4f1f021753a6547505d.0
Archive	Archive

Archive

Identifier	9abd37197bc62a72a303e5931984332a.0
Source	archive/lc_email27.txt
Application	perl/Digest::MD5
Output	52217a1bcd2be7cf05f36066d4cdc9cf
Application	Jhove (Rel. 1.0 (beta 2), 2004-07-19)
Output	Jhove (Rel. 1.0 (beta 2), 2004-07-19) Date: 2005-04-30 20:50:51 EDT RepresentationInformation: file:%2Fhome%2Frhaq%2Fspace%2FsampleArchive%2Farchive%2Fl ReportingModule: ASCII-hul, Rel. 1.0 (2004-05-05) LastModified: 2005 EDT Size: 6206 Format: ASCII Status: Well-formed and valid MIMEtyp charset=US-ASCII ASCIIMetadata: LineEndings: LF Checksum: 76c99 CRC32 Checksum: 52217a1bcd2be7cf5f36066d4cdc9cf Type: MD5 Che 6d51599d4d978e5d253e945a7248965ddc3616 Type: SHA-1
Application	file-4.07 magic file from /usr/share/file/magic
Output	ASCII English text
Application	Fred Info URI
Output	info:gdfr/fred/f/us-ascii
Resource	Metadata Original

Identifier	1641ad793a1cc597a18e9dd4dd3c64d5.0
Source	archive/205.gif
Application	perl/Digest::MD5
Output	18771f3ecd43f6421c49a9f3b31bd195
Application	Jhove (Rel. 1.0 (beta 2), 2004-07-19)
Output	Jhove (Rel. 1.0 (beta 2), 2004-07-19) Date: 2005-04-30 20:50:52 EDT RepresentationInformation: file:%2Fhome%2Frhaq%2Fspace%2FsampleArchive%2Farchive%2F2



Address: http://beatitude.cs.odu.edu:8080/~gmanepal/aiht/bucket/

Bucket Version 3

Bucket API

- ➔ admin
- ➔ display (SAX)
- ➔ display (DOM)
- log
- ➔ lint
- ➔ id - handle
- ➔ list methods
- get
- put

Bucket Objects

- ➔ bucket
- ➔ didl
- ➔ index

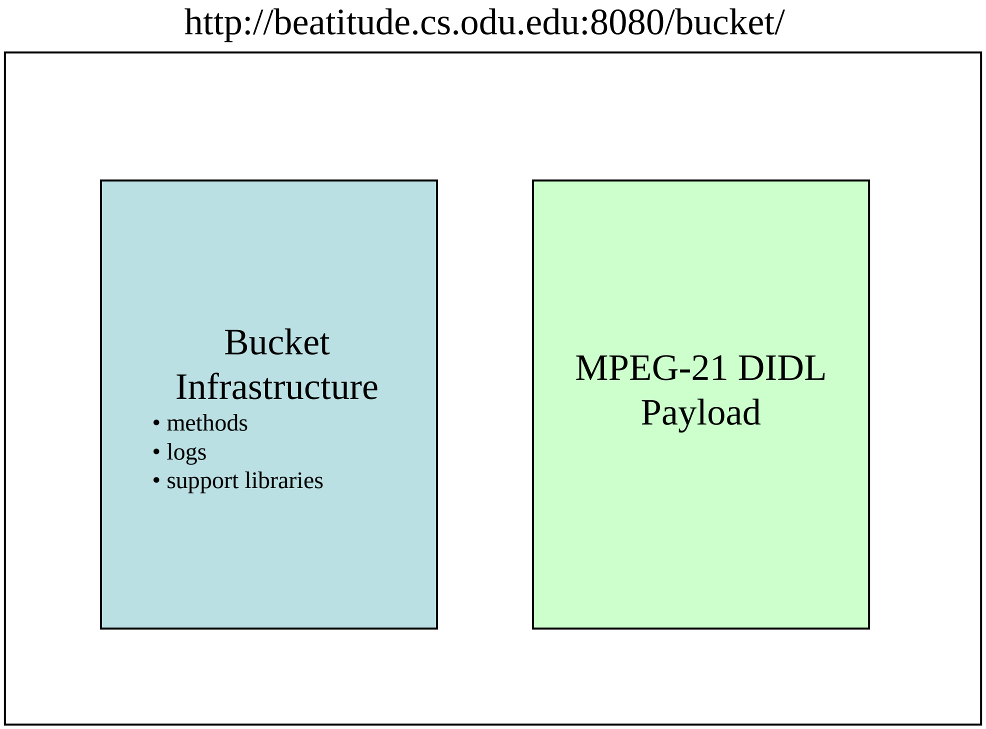
Preferences

default method	admin
log access	on
put	restricted

Display based on web use

Display based on archival use

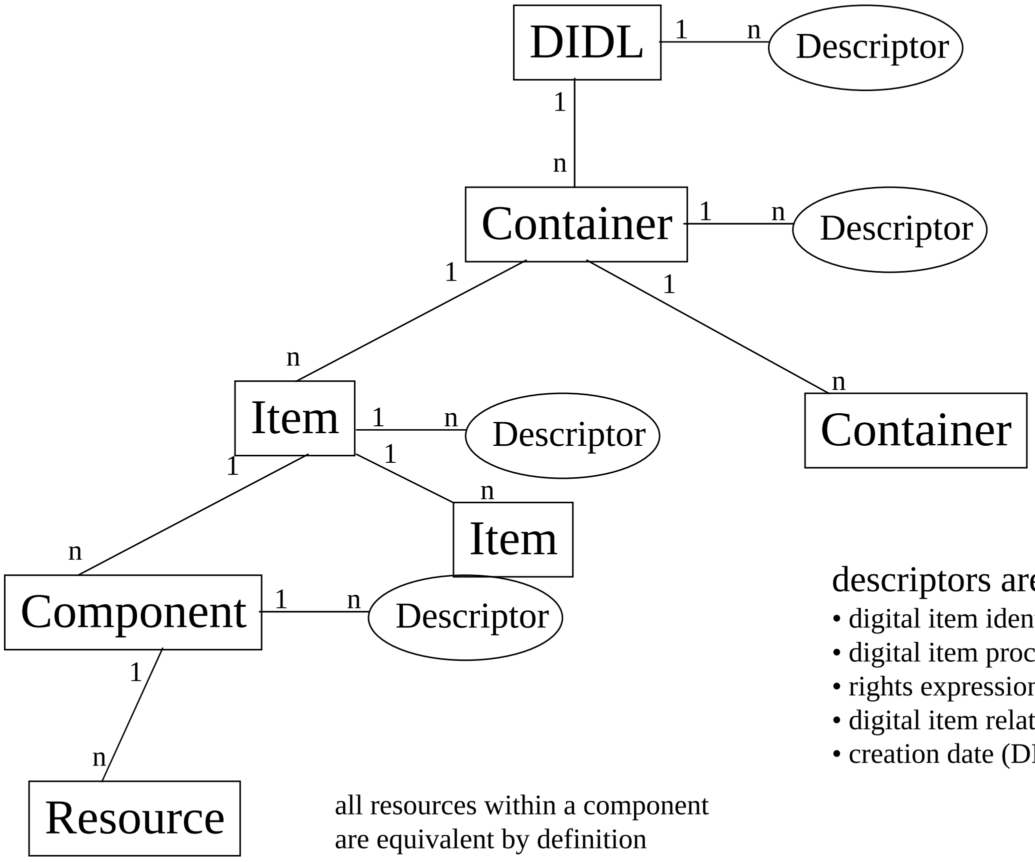
Bucket / MPEG-21 Model



MPEG-21 DIDL

- A generic, powerful complex object metadata format
  - – Based on an abstract data model
  - – Semantics separated from syntax
    - i.e. the tags don't mean anything -- a little disconcerting at first glance
  - – Digital library use championed by LANL
    - http://www.dlib.org/dlib/november03/bekaert/11bekaert.html
    - http://www.dlib.org/dlib/february04/bekaert/02bekaert.html
    - http://arxiv.org/abs/cs.DL/0502028

MPEG-21 DIDL Data Model

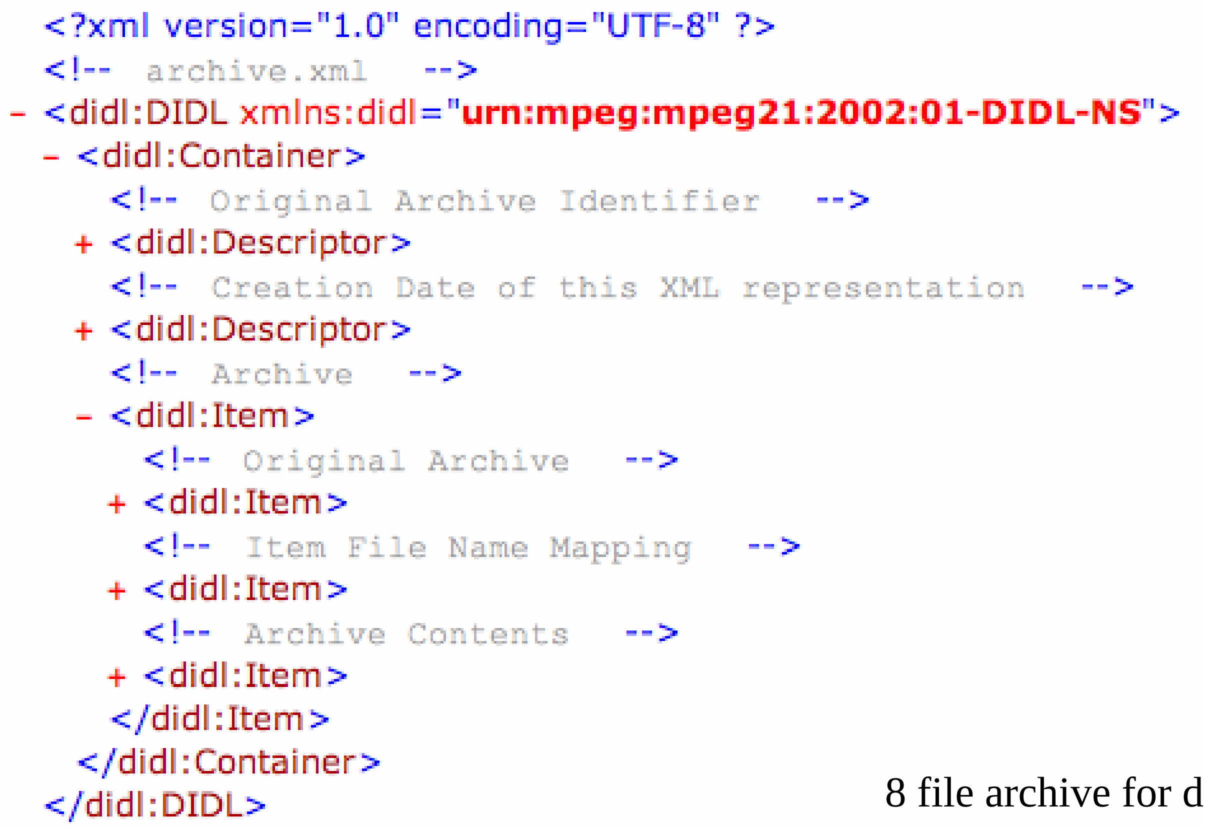


all resources within a component
are equivalent by definition

How to encode Archive?
- 1 file = 1 DID
- 1 archive = 1 container
- 1 archive = 1 component
- 1 file = 1 component

descriptors are used to convey:
- digital item identification (DII)
- digital item processing (DIP)
- rights expression language (REL)
- digital item relations (DIR)
- creation date (DIDT)

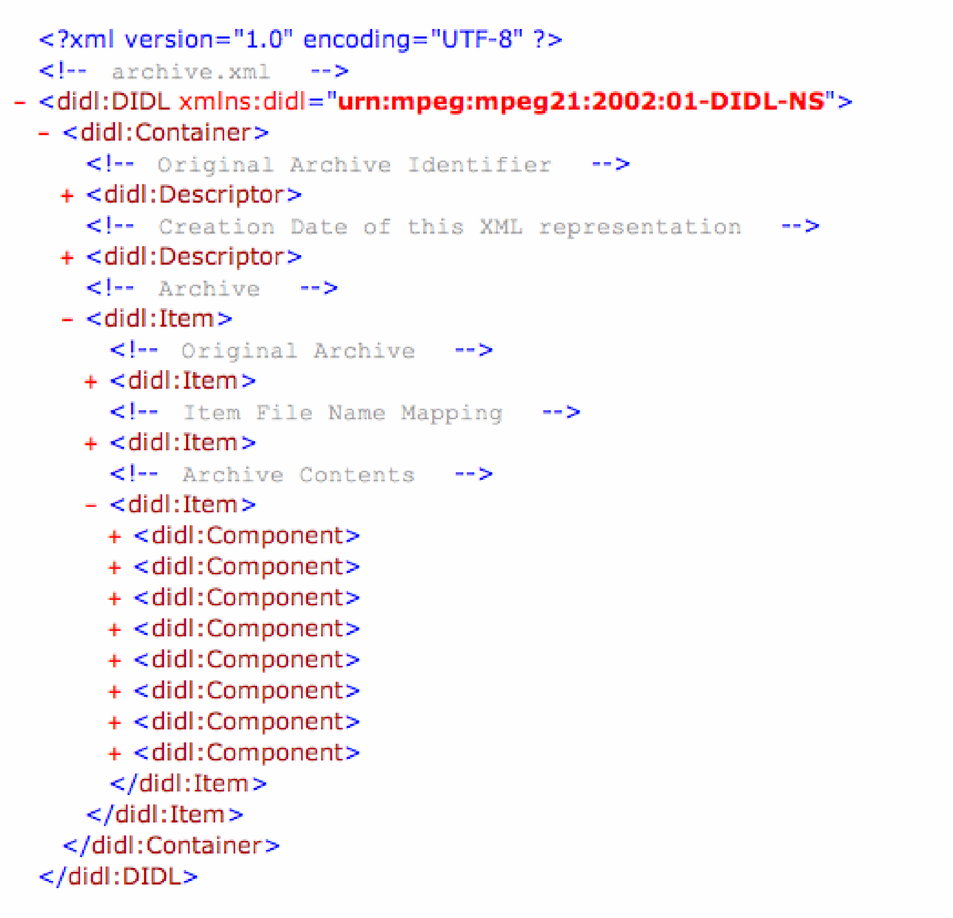
1 File = 1 Component



```
  <?xml version="1.0" encoding="UTF-8" ?>
  <!--  archive.xml   -->
- <didl:DIDL xmlns:didl="urn:mpeg:mpeg21:2002:01-DIDL-NS">
  - <didl:Container>
      <!--  Original Archive Identifier   -->
    + <didl:Descriptor>
      <!--  Creation Date of this XML representation   -->
    + <didl:Descriptor>
      <!--  Archive   -->
    - <didl:Item>
        <!--  Original Archive   -->
      + <didl:Item>
        <!--  Item File Name Mapping   -->
      + <didl:Item>
        <!--  Archive Contents   -->
      + <didl:Item>
      </didl:Item>
    </didl:Container>
  </didl:DIDL>
```

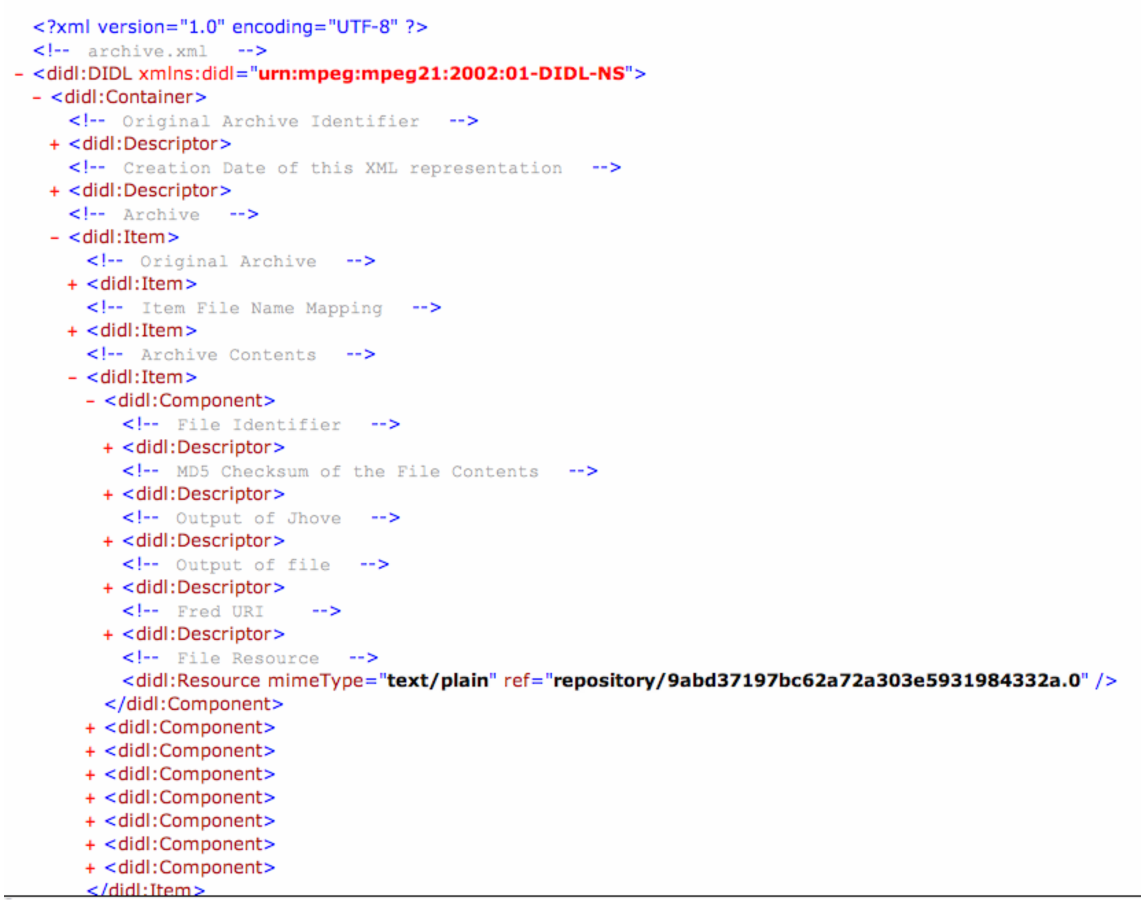
8 file archive for demo purposes…
http://www.cs.odu.edu/~mln/aiht/

Looking Inside the Archive



```
  <?xml version="1.0" encoding="UTF-8" ?>
  <!--  archive.xml  -->
- <didl:DIDL xmlns:didl="urn:mpeg:mpeg21:2002:01-DIDL-NS">
  - <didl:Container>
      <!--  Original Archive Identifier  -->
    + <didl:Descriptor>
      <!--  Creation Date of this XML representation  -->
    + <didl:Descriptor>
      <!--  Archive  -->
    - <didl:Item>
        <!--  Original Archive  -->
      + <didl:Item>
        <!--  Item File Name Mapping  -->
      + <didl:Item>
        <!--  Archive Contents  -->
      - <didl:Item>
        + <didl:Component>
        + <didl:Component>
        + <didl:Component>
        + <didl:Component>
        + <didl:Component>
        + <didl:Component>
        + <didl:Component>
        + <didl:Component>
        </didl:Item>
      </didl:Item>
    </didl:Container>
  </didl:DIDL>
```

Looking at a Single File…



```
 <?xml version="1.0" encoding="UTF-8" ?>
 <!--  archive.xml   -->
- <didl:DIDL xmlns:didl="urn:mpeg:mpeg21:2002:01-DIDL-NS">
 - <didl:Container>
     <!--  Original Archive Identifier   -->
   + <didl:Descriptor>
     <!--  Creation Date of this XML representation   -->
   + <didl:Descriptor>
     <!--  Archive   -->
   - <didl:Item>
       <!--  Original Archive   -->
     + <didl:Item>
       <!--  Item File Name Mapping   -->
     + <didl:Item>
       <!--  Archive Contents   -->
     - <didl:Item>
       - <didl:Component>
           <!--  File Identifier   -->
         + <didl:Descriptor>
           <!--  MD5 Checksum of the File Contents   -->
         + <didl:Descriptor>
           <!--  Output of Jhove   -->
         + <didl:Descriptor>
           <!--  Output of file   -->
         + <didl:Descriptor>
           <!--  Fred URI    -->
         + <didl:Descriptor>
           <!--  File Resource   -->
           <didl:Resource mimeType="text/plain" ref="repository/9abd37197bc62a72a303e5931984332a.0" />
         </didl:Component>
       + <didl:Component>
       + <didl:Component>
       + <didl:Component>
       + <didl:Component>
       + <didl:Component>
       + <didl:Component>
       + <didl:Component>
       </didl:Item>
```

Design Decisions: File Storage

- Store each file as a <Component>
  - Big: each file is base64'd into the DIDL
  - Small: each file is ref'd from the DIDL to a directory
    - Filename = MD5 hash of the original file name (not contents!) + a version number
    - Example:

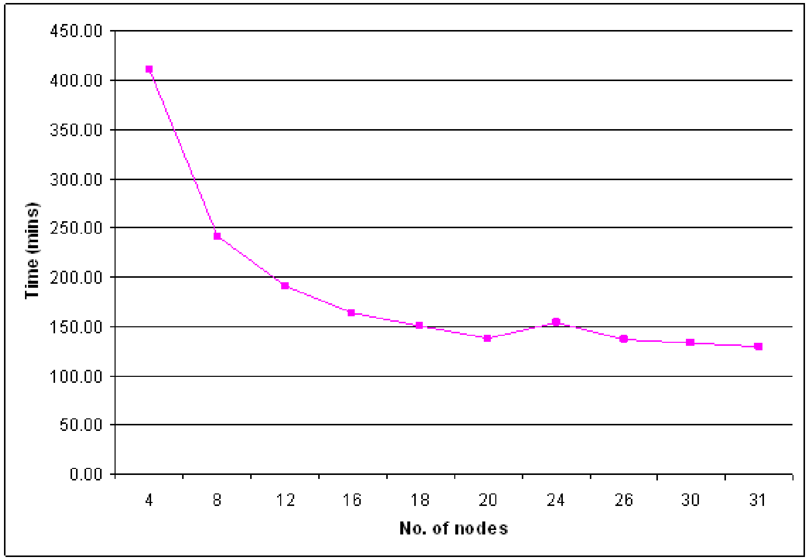
```
<didl:Resource mimeType="image/gif"ref="repository/1641ad793a1cc597a18e9dd4dd3c64d5.0" />
```

Archive Sizes

Name	XML File Size (bytes)	Notes
DIDL.xml	15382712841 (15 GB)	By-Value. This first upload did not contain the files "outside" of the original tar file.
DIDL2.xml	35633037513 (35 GB)	By-Value. This upload contained all files (tar file + database files)
DIDL3.xml	322653621 (322 MB)	By-Reference. All files. File size does not include tar file (26 GB).
DIDL4.xml	407093487 (407 MB)	By-Reference. Harvard Import. File size does not include tar file (24 GB).

Design Decisions: Ingestion

- For every program/process to apply to a file, create a corresponding <Descriptor>
  - Jhove
  - Unix “file”
  - Fred URI
  - MD5 of file contents
- Expandable, scriptable list of metadata extraction / analysis programs
- Ingestion is parallelized over a workstation cluster

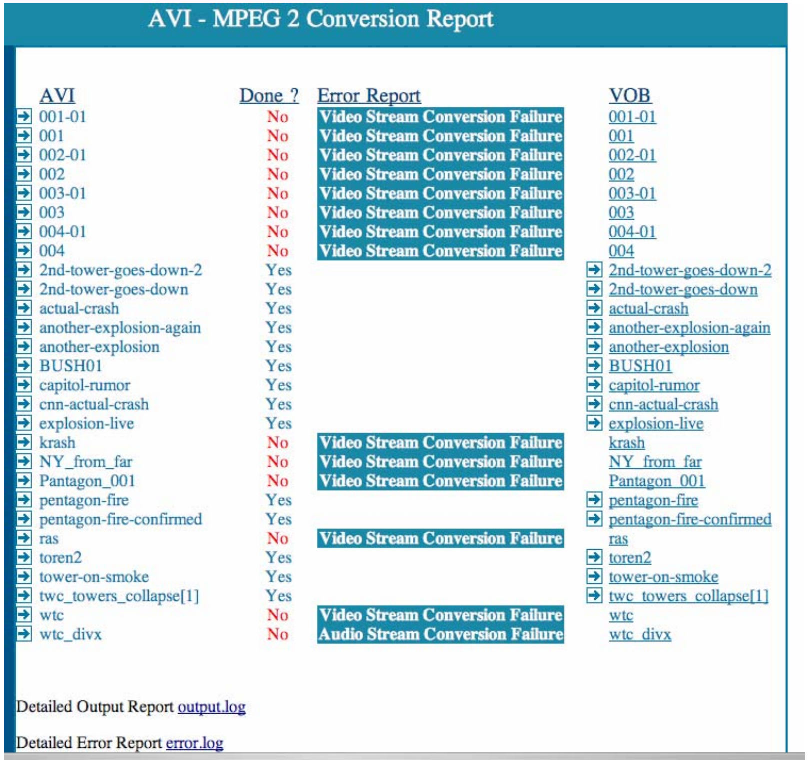


Example Output: MD5

```
<didl:Descriptor>
<didl:Statement mimeType="text/xml; charset=UTF-8">
  <dc:creator xmlns:dc="http://purl.org/dc/elements/1.1/"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xsi:schemaLocation="http://purl.org/dc/elements/1.1/
http://dublincore.org/schemas/xmls/simpledc20021212.xsd">perl/Digest::M
D5</dc:creator>
  <dc:description xmlns:dc="http://purl.org/dc/elements/1.1/"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xsi:schemaLocation="http://purl.org/dc/elements/1.1/
http://dublincore.org/schemas/xmls/simpledc20021212.xsd">52217a1bcd2b
e7cf05f36066d4cdc9cf</dc:description>
</didl:Statement>
</didl:Descriptor>
```

Conversion: AVI -> VOB

- Investigated PDF -> SVG, but tools were not mature
- Selected “transcode” for AVI -> VOB conversion
  - http://www.transcoding.org/
- Also implemented ImageMagick based rules for standard graphics conversion



AVI - MPEG 2 Conversion Report

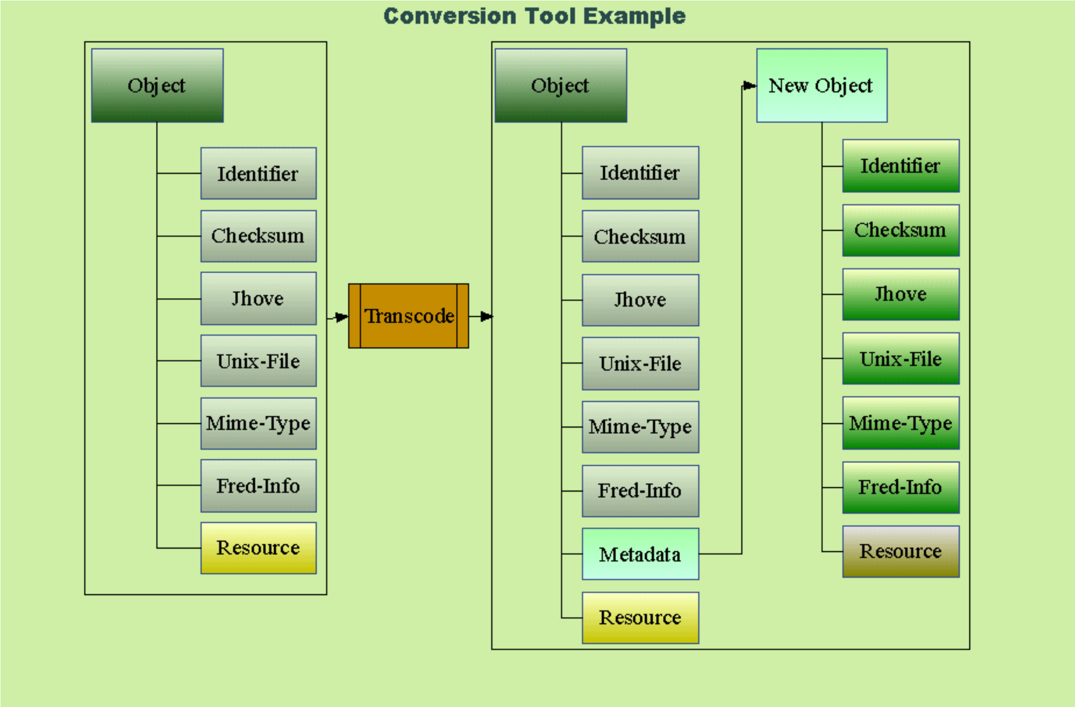
AVI	Done ?	Error Report	VOB
001-01	No	Video Stream Conversion Failure	001-01
001	No	Video Stream Conversion Failure	001
002-01	No	Video Stream Conversion Failure	002-01
002	No	Video Stream Conversion Failure	002
003-01	No	Video Stream Conversion Failure	003-01
003	No	Video Stream Conversion Failure	003
004-01	No	Video Stream Conversion Failure	004-01
004	No	Video Stream Conversion Failure	004
2nd-tower-goes-down-2	Yes		2nd-tower-goes-down-2
2nd-tower-goes-down	Yes		2nd-tower-goes-down
actual-crash	Yes		actual-crash
another-explosion-again	Yes		another-explosion-again
another-explosion	Yes		another-explosion
BUSH01	Yes		BUSH01
capitol-rumor	Yes		capitol-rumor
cnn-actual-crash	Yes		cnn-actual-crash
explosion-live	Yes		explosion-live
krash	No	Video Stream Conversion Failure	krash
NY_from_far	No	Video Stream Conversion Failure	NY_from_far
Pantagon_001	No	Video Stream Conversion Failure	Pantagon_001
pentagon-fire	Yes		pentagon-fire
pentagon-fire-confirmed	Yes		pentagon-fire-confirmed
ras	No	Video Stream Conversion Failure	ras
toren2	Yes		toren2
tower-on-smoke	Yes		tower-on-smoke
twc_towers_collapse[1]	Yes		twc_towers_collapse[1]
wtc	No	Video Stream Conversion Failure	wtc
wtc_divx	No	Audio Stream Conversion Failure	wtc_divx

Detailed Output Report output.log

Detailed Error Report error.log

http://beatitude.cs.odu.edu:8080/~gmanepal/Transcode.html

Conversion: Linking Old to New



If the previous version of the Resource was specified as:
<didl:Resource mimeType="**image/jpeg**"
ref="**repository/9abd37197bc62a72a303e5931984332a.0**" />
then the new version of the resource is specified as:
<didl:Resource mimeType="**image/png**"
ref="**repository/9abd37197bc62a72a303e5931984332a.1**" />

Harvard Ingest

- Harvard’s model was the most similar to our MPEG-21 model
- Ingesting from another archive is (roughly) the same as initial ingest
  - Save any metadata that was delivered in the original METS file as a <Descriptor>
    - We don’t trust it, but it might be useful for future forensics
  - Re-ingest in the normal way
- Our export is part of the bucket API:
  - http://beatitude.cs.odu.edu:8080/bucket/?method=get&id=didl

```
<didl:Descriptor>
<didl:Statement mimeType="text/xml; charset=UTF-8">
  <dc:creator xmlns:dc="http://purl.org/dc/elements/1.1/"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xsi:schemaLocation="http://purl.org/dc/elements/1.1/
http://dublincore.org/schemas/xmls/simpledc20021212.xsd">External Metadata</dc:creator>
  <dc:description xmlns:dc="http://purl.org/dc/elements/1.1/"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xmlns:aes="http://www.aes.org/audioObject"
xmlns:app="http://hul.harvard.edu/ois/xml/ns/drs/app" xmlns:mix="http://www.loc.gov/mix/"
xmlns:tcf="http://www.aes.org/tcf" xmlns:txt="http://www.loc.gov/METS/text/"
xmlns:xlink="http://www.w3.org/TR/xlink" xsi:schemaLocation="http://purl.org/dc/elements/1.1/
http://dublincore.org/schemas/xmls/simpledc20021212.xsd">
<file ID="F1" MIMETYPE="image/jpeg" SEQ="1" SIZE="194914" ADMID="T1"
CHECKSUM="a7969810684c468525313b8282501405" CHECKSUMTYPE="MD5"
OWNERID="aiht/websites/chnm/september11/REPOSITORY/CONTRIBUTORS/1199_photos/
wtc_web/wetc5.jpg">
  <FLocat LOCTYPE="URL" xlink:type="simple"
xlink:href="file:///aiht/data/2004/12/17/0/122.jpg" />
  </file>
<mix:mix>
<mix:BasicImageParameters>
<mix:Format>
  <mix:MIMEType>image/jpeg</mix:MIMEType>
  </mix:Format>
<mix:Compression>
  <mix:CompressionType>6</mix:CompressionType>
  </mix:Compression>
<mix:PhotometricInterpretation>
  <mix:ColorSpace>6</mix:ColorSpace>
  </mix:PhotometricInterpretation>
<mix:File>
  <mix:Orientation>1</mix:Orientation>
  </mix:File>
  </mix:BasicImageParameters>
<mix:ImageCreation>
<mix:DigitalCameraCapture>
  <mix:DigitalCameraModel>Canon Canon EOS D30</mix:DigitalCameraModel>
  </mix:DigitalCameraCapture>
  </mix:ImageCreation>
<mix:ImagingPerformanceAssessment>
<mix:SpatialMetrics>
  <mix:SamplingFrequencyUnit>2</mix:SamplingFrequencyUnit>
  <mix:ImageWidth>540</mix:ImageWidth>
  <mix:ImageLength>360</mix:ImageLength>
  </mix:SpatialMetrics>
<mix:Energetics>
  <mix:BitsPerSample>8 8 8</mix:BitsPerSample>
  </mix:Energetics>
  </mix:ImagingPerformanceAssessment>
  </mix:mix>
  </dc:description>
</didl:Statement>
</didl:Descriptor>
```

“In Vivo” Preservation

- As part of the ingest process, we looked for copies of the ingested web page in the “living web”
  - – Idea: find all replicated / similar pages and maintain pointers to them
  - – Problem: We could find related documents, but finding copies was difficult
    - Term Frequency (TF) – easy to compute
    - Inverse Document Frequency (IDF) – difficult to compute
    - Solution: lexical signatures, Phelps & Wilensky:
      - – http://www.dlib.org/dlib/july00/wilensky/07wilensky.html
  - – Spinoff research:
    - Terry Harrison’s MS thesis
    - Frank McCown’s Ph.D. dissertation
    - Joan Smith’s Ph.D. dissertation
    - NSF proposal on “in vivo” preservation

The DIP is the TMD*

- Using METS or MPEG-21, there is no need for a separate transfer metadata format
- METS & MPEG-21 can be the lumps of XML exchanged between harvesters & repositories
  - http://www.dlib.org/dlib/december04/vandesompel/12vandesompel.html
- Web servers can be made to automatically expose their contents via OAI-PMH
  - http://www.modoai.org/

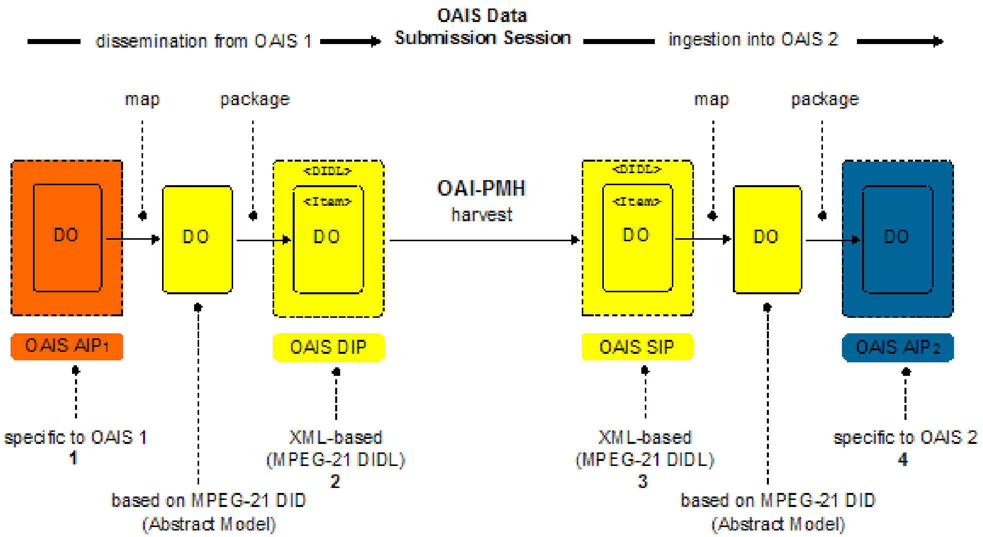


Figure 1, Bekaert & Van de Sompel
http://www.dlib.org/dlib/june05/bekaert/06bekaert.html

* Eat your heart out, Marshal McLuhan