OAIResource Software

Michael Nelson
Computer Science Department
Old Dominion University

Herbert Van de Sompel
Digital Library Research & Prototyping Team
Research Library, Los Alamos National Laboratory

This work supported in part by the Library of Congress

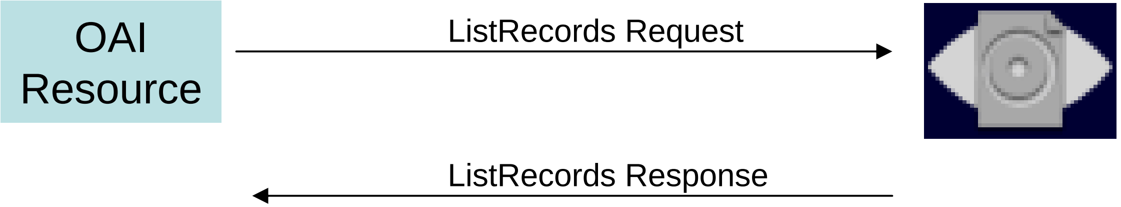
OAIResource

- “Resource-Aware” OAI-PMH harvester
  - combines metadata and resource harvesting in a single application
    - http://dx.doi.org/10.1045/june2005-bekaert
  - developed at LANL Research Library
    - written in Java
    - plug-in architecture for extensibility
  - based on (among others):
    - OAICat (OCLC)
    - Heretrix (Internet Archive)
    - aDORe tools (LANL RL)

Two Primary Use Cases

- Target repository:
  - – supports resource harvesting
    - via a complex object format (e.g., MPEG-21 DIDL, METS, etc.)
  - – does not support resource harvesting
    - uses only Dublin Core or other metadata formats

Repository Does Not Support COs



```
<oai_dc:dc>
   <dc:title>A Simple Parallel-Plate Resonator Technique
      for Microwave.
       Characterization of Thin Resistive Films</dc:title>
   <dc:creator>Vorobiev, A.</dc:creator>
   <dc:subject>ING-INF/01 Elettronica</dc:subject>
   <dc:description>A parallel-plate resonator method is
      proposed for
       non-destructive characterisation of resistive films
      used in
       microwave integrated circuits. A slot made in one
      ... </dc:description>
   <dc:publisher>Microwave engineering
      Europe</dc:publisher>
   <dc:date>2002</dc:date>
   <dc:type>Documento relativo ad una Conferenza o altro
      Evento</dc:type>
   <dc:type>PeerReviewed</dc:type>

      <dc:identifier>http://amsacta.cib.unibo.it/archive/0
      0000014/</dc:identifier>
   <dc:format>pdf

      http://amsacta.cib.unibo.it/archive/00000014/01/GaAs
      _1_Vorobiev.pdf
   </dc:format>
</oai_dc:dc>
```

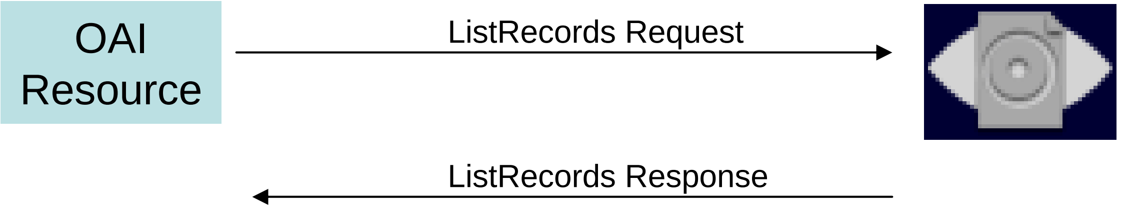
```
<oai_dc:dc>
   <dc:title>A Simple Parallel-Plate Resonator Technique
      for Microwave.
       Characterization of Thin Resistive Films</dc:title>
   <dc:creator>Vorobiev, A.</dc:creator>
   <dc:subject>ING-INF/01 Elettronica</dc:subject>
   <dc:description>A parallel-plate resonator method is
      proposed for
       non-destructive characterisation of resistive films
      used in
       microwave integrated circuits. A slot made in one
      ... </dc:description>
   <dc:publisher>Microwave engineering
      Europe</dc:publisher>
   <dc:date>2002</dc:date>
   <dc:type>Documento relativo ad una Conferenza o altro
      Evento</dc:type>
   <dc:type>PeerReviewed</dc:type>

      <dc:identifier>http://amsacta.cib.unibo.it/archive/0
      0000014/</dc:identifier>
   <dc:format>pdf

      http://amsacta.cib.unibo.it/archive/00000014/01/GaAs
      _1_Vorobiev.pdf
   </dc:format>
</oai_dc:dc>
```

1. Write OAI-PMH harvest to XMLtape
2. Per record in XMLtape: Examine URL in DC.Identifier or DC.Format
   - GET if pdf, jpg, tif, mp3
   - if html, GET and scrape for pdf, jpg, tif, mp3
3. write datastream to Arc file
4. Write connection between XMLtape and ARCfile in ok.csv
5. Write failure in bad.csv

Repository Does Support COs



```
<didl:DIDL>
<didl:Item>
   <didl:Descriptor><didl:Statement mimeType="text/xml;
     charset=UTF-8">
      <dii:Identifier>
        http://amsacta.cib.unibo.it/archive/00000014/
      </dii:Identifier>
   </didl:Statement></didl:Descriptor>
   <didl:Descriptor><didl:Statement mimeType="text/xml;
     charset=UTF-8">
      <oai_dc:dc>
        <dc:title>A Simple Parallel-Plate Resonator
     Technique for
           Microwave. Characterization of Thin Resistive
     Films
        </dc:title>
        <dc:creator>Vorobiev, A.</dc:creator>
        <dc:identifier>

     http://amsacta.cib.unibo.it/archive/00000014/</dc:iden
     tifier>
        <dc:format>application/pdf</dc:format>
        É
     </oai_dc:dc>
   </didl:Statement></didl:Descriptor>
  <didl:Component>
    <didl:Resource mimeType="application/pdf"
     ref="http://amsacta.cib.unibo.it/archive/00000014/01/G
     aAs_1_Vorobiev.pdf"/>
  </didl:Component>
</didl:Item>
</didl:DIDL>
```

```
<didl:DIDL>
<didl:Item>
   <didl:Descriptor><didl:Statement mimeType="text/xml;
     charset=UTF-8">
      <dii:Identifier>
        http://amsacta.cib.unibo.it/archive/00000014/
      </dii:Identifier>
   </didl:Statement></didl:Descriptor>
   <didl:Descriptor><didl:Statement mimeType="text/xml;
     charset=UTF-8">
      <oai_dc:dc>
        <dc:title>A Simple Parallel-Plate Resonator
     Technique for
           Microwave. Characterization of Thin Resistive
     Films
        </dc:title>
        <dc:creator>Vorobiev, A.</dc:creator>
        <dc:identifier>

     http://amsacta.cib.unibo.it/archive/00000014/</dc:iden
     tifier>
        <dc:format>application/pdf</dc:format>
        É
     </oai_dc:dc>
   </didl:Statement></didl:Descriptor>
  <didl:Component>
    <didl:Resource mimeType="application/pdf"
     ref="http://amsacta.cib.unibo.it/archive/00000014/01/G
     aAs_1_Vorobiev.pdf"/>
  </didl:Component>
</didl:Item>
</didl:DIDL>
```

1. Write OAI-PMH harvest to XMLtape
2. Per record in XMLtape: unpack DIDL.
   Per Resource in DIDL:
   - extract base64 if datastream by-val
   - http GET if datastream by-ref (@ ref attribute)
3. write datastream to Arc file
4. Write connection between XMLtape and ARC file to ok.cvs
5. Write failure to bad.csv

Plug-Ins provided in OAIResource Package

- DCFormatProcessor
  - look for pdf, jpg, tif, mp3 URLs in DC.Format (as per eprints.org repositories)
- DCIdentifierProcessor
  - look for html files in DC.Identifier; grab those and extract pdf, jpg, tif, mp3 files one level down
- DidlSigProcessor
  - process DIDLs with XML signatures
    - see: http://dx.doi.org/10.1045/june2005-bekaert
- DidlNoSigProcessor
  - process DIDLs without XML signatures

OAIResource Demo

```
[AIHT:~] mln% ssh demo@128.82.5.248
Password:
Welcome to Darwin!
[dhcp-248:~] demo% cd oai-resource/
dhcp-248.cs.odu.edu:/Users/demo/oai-resource % cd bin
dhcp-248.cs.odu.edu:/Users/demo/oai-resource/bin % sh ./OAIResource.sh ../etc/examples/env.properties ../etc/examples/libeprints.open.ac.uk/libeprints.open.ac.uk.properties
java version "1.5.0_02"
Java(TM) 2 Runtime Environment, Standard Edition (build 1.5.0_02-56)
Java HotSpot(TM) Client VM (build 1.5.0_02-36, mixed mode, sharing)
…
INFO  gov.lanl.harvester.ListRecords2Tape - http://libeprints.open.ac.uk/perl/oai2?verb=ListRecords&from=&until=2005-10-15&set=&metadataPrefix=oai_dc,size=47
INFO  gov.lanl.harvester.ListRecords2Tape - http://libeprints.open.ac.uk/perl/oai2?verb=ListRecords&resumptionToken=0/8941500/oai_dc, size=2
INFO  gov.lanl.harvester.ListRecords2Tape - Done -- harvested  49 records
INFO  gov.lanl.ingest.oaitape.DirIngester - project:EPRINT
INFO  gov.lanl.ingest.oaitape.DirIngester - plugin:gov.lanl.ingest.oaitape.simple.DCFormatProcessor
INFO  gov.lanl.ingest.oaitape.DirIngester - lastingest:19000101010101
INFO  gov.lanl.ingest.oaitape.DirIngester - xmltapesdir:/Users/demo/libeprints.open.ac.uk/tape/
INFO  gov.lanl.ingest.oaitape.DirIngester - start ingest tape:tape_8cfa2100-8a27-4e4c-a8b5-e6ff0ca169bc.xml
INFO  gov.lanl.ingest.oaitape.simple.DCFormatProcessor -  http://libeprints.open.ac.uk/archive/00000002/01/LIBARTVICEprints.pdf
Oct 15, 2005 5:22:14 PM org.archive.io.arc.ARCWriter close
INFO: Closed /Users/demo/libeprints.open.ac.uk/data/20051015172200/tape_8cfa2100-8a27-4e4c-a8b5-e6ff0ca169bc/tmp68653e30-37d2-4be6-8913-a3ad05959f5c.arc, size 117533
INFO  gov.lanl.ingest.oaitape.simple.DCFormatProcessor -  http://libeprints.open.ac.uk/archive/00000023/01/Sandrine_EcologicalEconomicsPublicat.pdf
Oct 15, 2005 5:22:16 PM org.archive.io.arc.ARCWriter close
INFO: Closed /Users/demo/libeprints.open.ac.uk/data/20051015172200/tape_8cfa2100-8a27-4e4c-a8b5-e6ff0ca169bc/tmp02ca96eb-5a34-4a23-bb02-9f10c59fbccb.arc, size 241509
INFO  gov.lanl.ingest.oaitape.simple.DCFormatProcessor -  http://libeprints.open.ac.uk/archive/00000025/01/lib_subject_guides.pdf
Oct 15, 2005 5:22:17 PM org.archive.io.arc.ARCWriter close
INFO: Closed /Users/demo/libeprints.open.ac.uk/data/20051015172200/tape_8cfa2100-8a27-4e4c-a8b5-e6ff0ca169bc/tmpd8f66bac-d543-4c16-afd0-950fb37b5e43.arc, size 174875
INFO  gov.lanl.ingest.oaitape.simple.DCFormatProcessor -  http://libeprints.open.ac.uk/archive/00000027/01/RIA_SPAR.pdf
Oct 15, 2005 5:22:17 PM org.archive.io.arc.ARCWriter close
INFO: Closed /Users/demo/libeprints.open.ac.uk/data/20051015172200/tape_8cfa2100-8a27-4e4c-a8b5-e6ff0ca169bc/tmpb963273a-1c64-4f5a-82ea-a39c48efd93b.arc, size 184171
…
INFO: Closed /Users/demo/libeprints.open.ac.uk/data/20051015172200/tape_8cfa2100-8a27-4e4c-a8b5-e6ff0ca169bc/tmpb1a53f32-b6a3-4f55-9171-a7c2155c7d21.arc, size 400013
WARN  gov.lanl.ingest.oaitape.simple.DCFormatProcessor - DCFormatProcessor:null
Oct 15, 2005 5:24:04 PM org.archive.io.arc.ARCWriter close
INFO: Closed /Users/demo/libeprints.open.ac.uk/data/20051015172200/tape_8cfa2100-8a27-4e4c-a8b5-e6ff0ca169bc/tmpf4b3796e-842c-4ff3-a3b5-bcbc13d46999.arc, size 169
WARN  gov.lanl.ingest.oaitape.simple.DCFormatProcessor - DCFormatProcessor:null
Oct 15, 2005 5:24:04 PM org.archive.io.arc.ARCWriter close
INFO: Closed /Users/demo/libeprints.open.ac.uk/data/20051015172200/tape_8cfa2100-8a27-4e4c-a8b5-e6ff0ca169bc/tmp1a24a284-d384-4da3-83b5-a33b019ac61a.arc, size 169
Oct 15, 2005 5:24:04 PM org.archive.io.arc.ARCWriter close
INFO: Closed /Users/demo/libeprints.open.ac.uk/data/20051015172200/tape_8cfa2100-8a27-4e4c-a8b5-e6ff0ca169bc/EPRINT_abe7aae7-701d-44af-8be5-dd20edb20ff6.arc, size 26756730
INFO  gov.lanl.ingest.oaitape.DirIngester - finish tape:tape_8cfa2100-8a27-4e4c-a8b5-e6ff0ca169bc.xml
INFO  gov.lanl.ingest.oaitape.DirIngester - finish directory:20051015172200
```

OAIResource Output

```
dhcp-248.cs.odu.edu:/Users/demo/libeprints.open.ac.uk % ls -R
data/                lastingest.txt        tape/
lastharvest.txt        log/

./data:
20051015172200/

./data/20051015172200:
tape_8cfa2100-8a27-4e4c-a8b5-e6ff0ca169bc/

./data/20051015172200/tape_8cfa2100-8a27-4e4c-a8b5-e6ff0ca169bc:
EPRINT_abe7aae7-701d-44af-8be5-dd20edb20ff6.arc
bad.csv
ok.csv

./log:
20051015172200/

./log/20051015172200:
tape_8cfa2100-8a27-4e4c-a8b5-e6ff0ca169bc/

./log/20051015172200/tape_8cfa2100-8a27-4e4c-a8b5-e6ff0ca169bc:
log4joutput.log

./tape:
20051015172200/

./tape/20051015172200:
tape_8cfa2100-8a27-4e4c-a8b5-e6ff0ca169bc.xml
dhcp-248.cs.odu.edu:/Users/demo/libeprints.open.ac.uk %
```

resources

mapping

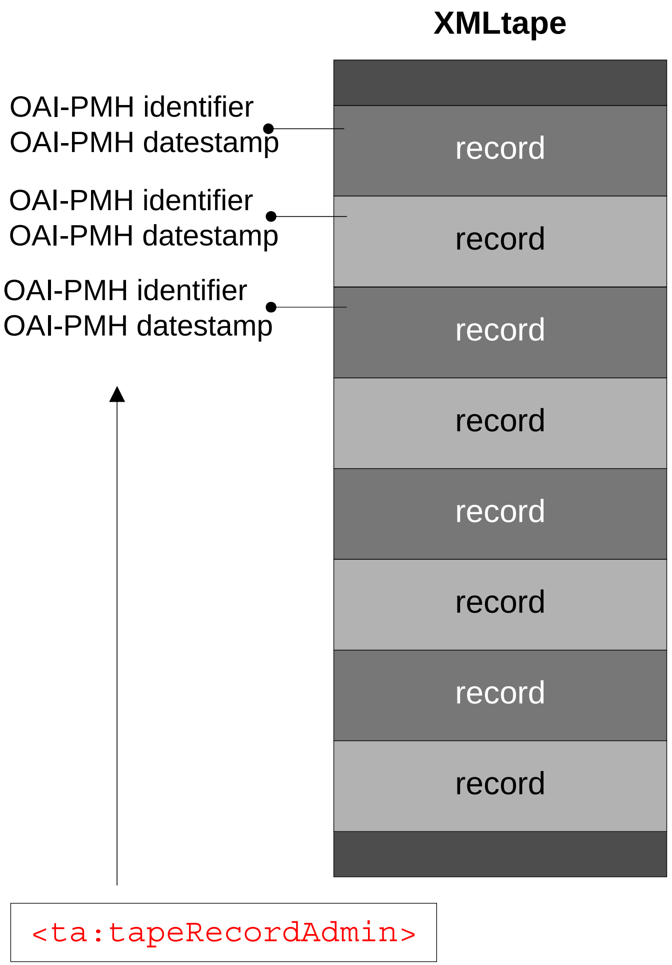
run time mesgs

metadata

aDORe XMLtape

- An XML file that concatenates the XML-based representations of multiple DOs
- Structure is defined by an XML Schema
  - http://purl.lanl.gov/aDORe/schemas/2005-08/XMLtape.xsd
  - tape-level administrative section:
    - Open-ended content
    - Plug-in for processing-related information, indication of related ARCfiles:
      - http://purl.lanl.gov/aDORe/schemas/2005-08/XMLtapeBasics.xsd
  - concatenation of records, each of which consists of:
    - record-level administrative section
      - identifier and datestamp of the contained record
      - other record-level administrative information
    - a record (can be from any XML Namespace). DIDL in case of aDORe:
      - http://purl.lanl.gov/aDORe/schemas/2005-08/DIDL.xsd
- An XMLtape is a valid and well-formed XML file
- Independent from chosen XML-based Compound Object Format

aDORe XMLtape



aDORe XMLtape

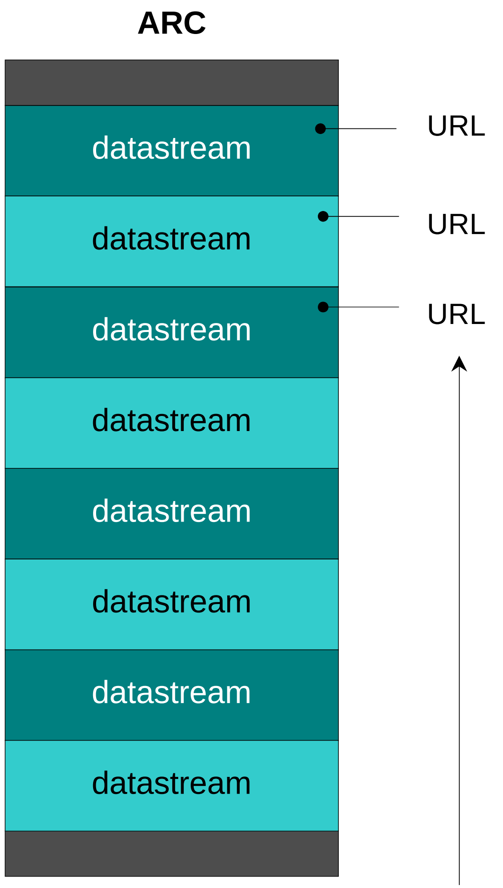
```
<?xml version="1.0" encoding="UTF-8"?>
<ta:tape xmlns:ta="http://library.lanl.gov/2005-08/aDORe/XMLtape/"
   <ta:tapeAdmin>
      ...
   </ta:tapeAdmin>
   <ta:tapeRecord>
      <ta:tapeRecordAdmin>
         <ta:identifier>oai:open.ac.uk.OAI2:2</ta:identifier>
         <ta:date>2005-03-29Z</ta:date>
         <ta:recordAdmin>
            ...
         </ta:recordAdmin>
      </ta:tapeRecordAdmin>
      <ta:record>
         <oai_dc:dc>...</oai_dc:dc>
      </ta:record>
   </ta:tapeRecord>
</ta:tape>
```

aDORe ta:tape

Internet Archive ARCfile

- Concatenation of binary files
- Designed and used by the Internet Archive (Wayback machine)
  - > 400 TB web data
- Under revision by the International Internet Preservation Consortium (IIPC): WARC file format
  - Input from LANL to facilitate non-Web-crawling use case
- The ARC file format is structured as follows:
  - file header that provides administrative information about the ARC file itself
  - a sequence of document records, consisting of:
    - a header line containing some, mainly crawl-related, metadata.
      - URI of the crawled document
      - timestamp of acquisition of the data
      - size of the data block
    - a response to a protocol request such as an HTTP GET

Internet Archive ARC file



```
URL IP-address Archive-date Content-type Archive-length
```

Internet Archive ARC file

```
filedesc://IA-001102.arc 0 19960923142103 text/plain 76
1 0 Alexa Internet
URL IP-address Archive-date Content-type Archive-length


http://www.dryswamp.edu:80/index.html 127.10.100.2 19961104142103 text/html
202
HTTP/1.0 200 Document follows
Date: Mon, 04 Nov 1996 14:21:06 GMT
Server: NCSA/1.4.1
Content-type: text/html Last-modified: Sat,10 Aug 1996 22:33:11 GMT
Content-length: 30
<HTML>
Hello World!!!
</HTML>
```

sample ARC file

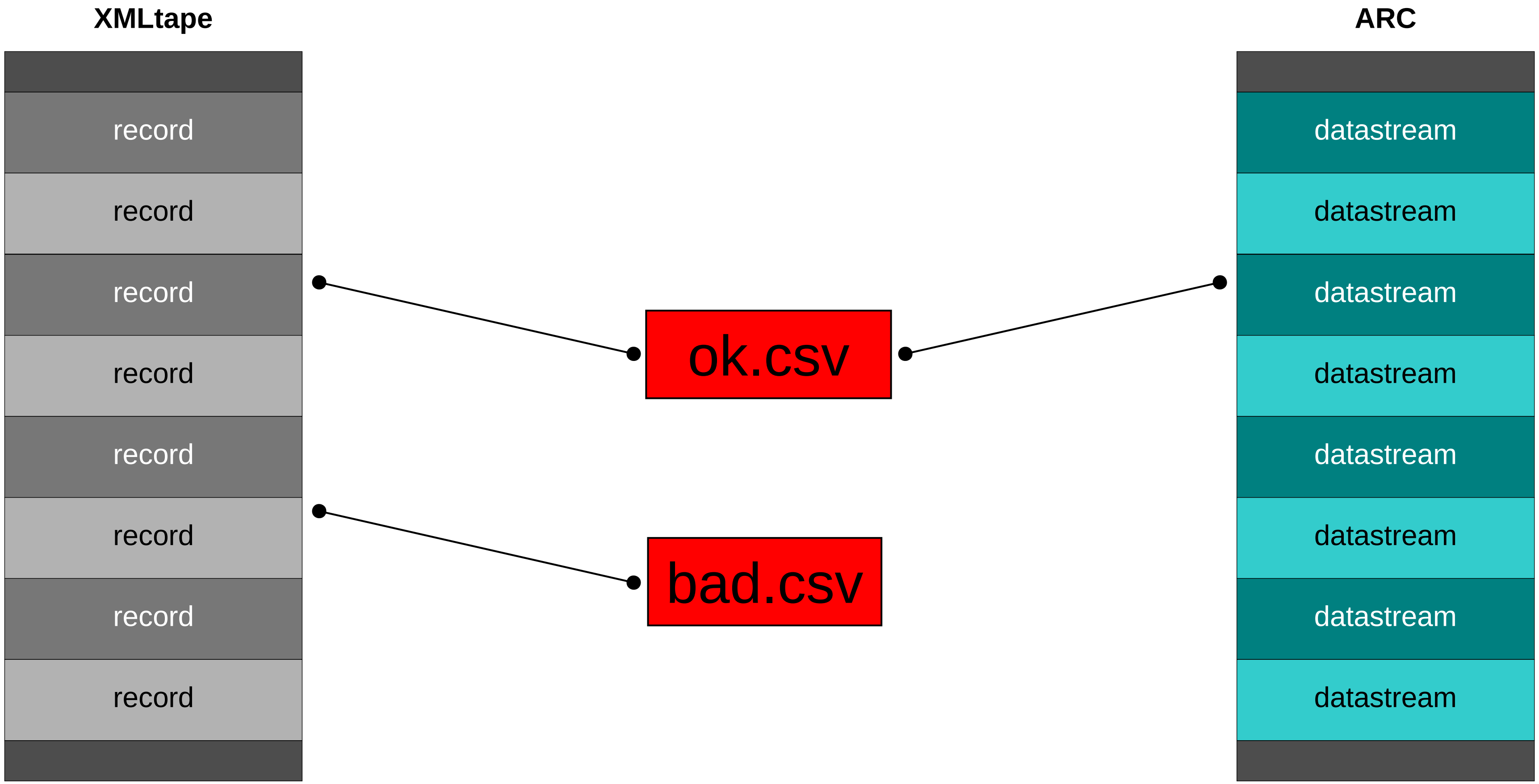
Internet Archive ARC file in OAIResource

```
filedesc://singletape.arc 0.0.0.0 20050922142103 text/plain 76 1 0
Internet Archive
URL IP-address Archive-date Content-type Archive-length

info:lanl-repo/ds/39c2fa93-fa22-4c19-90af-b5f58b9b989a 0.0.0.0 20050907221344
application/pdf 415025
  %PDF-1.3
  %âãÏÓ
  290
  0 obj
  <<
  /Linearized 1
  /O 295
  /H [ 3642 1057 ]
  /L 415025
  …
```

sample OAIResource ARC file

OAIResource XMLtape/ARCfile storage



XMLtape

record

record

record

record

record

record

record

record

ok.csv

bad.csv

ARC

datastream

datastream

datastream

datastream

datastream

datastream

datastream

datastream

needs post-processing for ingest into repository, for creation of an application, etc.

Mapping Between XMLTape & ARCfiles

```
% more bad.csv
# "tape_record_id, arc_date, message"
oai:open.ac.uk.OAI2:53,20051015212314,Unable to locate supported file.
oai:open.ac.uk.OAI2:5,20051015212404,
oai:open.ac.uk.OAI2:70,20051015212404,
```

```
% more ok.csv
# "tape_record_id, arc_id, arc_date, ref, derefXPath, sourceURI, digest, localIdentifier"
oai:open.ac.uk.OAI2:2,
EPRINT_abe7aae7-701d-44af-8be5dd20edb20ff6,20051015212214,
20051015212214,
,
//dc:format/*,
http://libeprints.open.ac.uk/archive/00000002/01/LIBARTVICEprints.pdf,
urn:sha1:KwJ0KYN27AO92aY3JZxsxqaH/CY=,
info:lanl-repo/ds/ffe52b26-58c0-4e6d-b4d5-ebb7a9fe82d1
…
oai:open.ac.uk.OAI2:23,EPRINT_abe7aae7-701d-44af-8be5-
dd20edb20ff6,20051015212215,,//dc:format/*,http://libeprints.open.ac.uk/archive/00000023/01/Sandrine_EcologicalEconomicsPublicat.pdf,urn:s
ha1:YG+ja4cmX7wmTfgrGXPDI7eE8bY=,info:lanl-repo/ds/6e559952-dcef-4c65-a669-ef023b4522fe
oai:open.ac.uk.OAI2:36,EPRINT_abe7aae7-701d-44af-8be5-
dd20edb20ff6,20051015212221,,//dc:format/*,http://libeprints.open.ac.uk/archive/00000036/01/fish_out_of_water_-_ray_and_rose.pdf,urn:sha1:
l5S2qHum4KB+febcUdnNJgTbfXc=,info:lanl-repo/ds/13f3d72b-40d2-4646-8f07-ac99f9137558
oai:open.ac.uk.OAI2:39,EPRINT_abe7aae7-701d-44af-8be5-
dd20edb20ff6,20051015212225,,//dc:format/*,http://libeprints.open.ac.uk/archive/00000039/01/Simon_Sustainability_Gap.pdf,urn:sha1:JfA5TeC
B
fxF8kkwZAqKu8mFvg8I=,info:lanl-repo/ds/8668c560-79d2-4053-b2e4-e00240872000
…
```

Papers

- Jeroen Bekaert and Herbert Van de Sompel. A Standards-based Solution for the Accurate Transfer of Digital Assets. D-Lib Magazine, June 2005. http://dx.doi.org/10.1045/june2005-bekaert
- Xiaoming Liu, Luda Balakireva, Patrick Hochstenbach and Herbert Van de Sompel. File-based storage of Digital Objects and constituent datastreams: XMLtapes and Internet Archive ARC files. ECDL 2005 paper in Lecture Notes in Computer Science at http://dx.doi.org/10.1007/11551362_23 . Preprint at http://arxiv.org/abs/cs.DL/0503016