Using OAI-PMH for *Resource* Exchange

OAI Metadata Harvesting Workshop, JCDL 03

Michael L. Nelson, Terry L. Harrison

Old Dominion University

Norfolk VA

{mln,tharriso}@cs.odu.edu

OAI-PRH?

- using OAI-PMH for resource extraction / exchange
  - yes, OAI-PMH is for metadata not resources, but its going to happen anyway…
    - mirroring
    - preservation (archive “zipping”)
    - convergence with OAIS
  - assumptions
    - a digital resource
    - rsync et al. neither appropriate nor possible
    - defer metadata vs. data discussion

Possible Approaches

1. Exploit knowledge outside the scope of the OAI-PMH to extract the resource
2. Base64 encode the resource and transmit via OAI-PMH as a separate “metadata” prefix?
3. Separate metadata prefix with instructions on how to extract / scrape the resource
4. Separate metadata format with XML encoded metadata, along with XSLT to decode it

Out of Band Knowledge

Back Forward Stop Refresh Home AutoFill Print Mail

Address: http://naca.larc.nasa.gov/oai2.0/?verb=GetRecord&metadataPrefix=oai_dc&identifier=oai:naca.larc.nasa.gov:1958:naca-tn-4410


```
<?xml version="1.0" encoding="UTF-8" ?>
- <OAI-PMH xmlns="http://www.openarchives.org/OAI/2.0/" xmlns:xsi="http://www.w3.org/2001/
  XMLSchema-instance" xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/ http://
  www.openarchives.org/OAI/2.0/OAI-PMH.xsd">
  <responseDate>2003-05-31T14:07:57+00:00</responseDate>
  <request identifier="oai:naca.larc.nasa.gov:1958:naca-tn-4410" metadataPrefix="oai_dc"
    verb="GetRecord">http%3A%2F%2Fnaca.larc.nasa.gov%2Foai2.0%2Findex.cgi</request>
  - <GetRecord>
    - <record>
      - <header>
          <identifier>oai:naca.larc.nasa.gov:1958:naca-tn-4410</identifier>
          <datestamp>2001-07-27</datestamp>
        </header>
      - <metadata>
        - <oai_dc:dc xmlns:oai_dc="http://www.openarchives.org/OAI/2.0/oai_dc/"
            xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xmlns:dc="http://
            purl.org/dc/elements/1.1/" xsi:schemaLocation="http://www.openarchives.org/
            OAI/2.0/oai_dc/ http://www.openarchives.org/OAI/2.0/oai_dc.xsd">
            <dc:title>Flight measurements of the vibratory stresses on a propeller designed for
              an advance ratio of 4.0 and a Mach number of 0.82</dc:title>
            <dc:creator>Thomas C. O'Bryan</dc:creator>
            <dc:date>SEP 1958</dc:date>
            <dc:identifier>http://naca.larc.nasa.gov/reports/1958/naca-tn-4410/<
              /dc:identifier>
            <dc:type>NACA TN 4410</dc:type>
            <dc:contributor>NACA Langley Aeronautical Laboratory</dc:contributor>
            <dc:description>Results are presented of vibratory-stress measurements obtained
              in flight on a propeller designed for an advance ratio of 4.0 and a forward Mach
              number of 0.82.</dc:description>
          </oai_dc:dc>
        </metadata>
```


1. take url in dc:identifier
2. parse report number
3. append “reportnumber.pdf” to url

Back Forward Stop Refresh Home AutoFill Print Mail

Address: http://ntrs.nasa.gov/?verb=GetRecord&metadataPrefix=oai_dc&identifier=oai:magicnrc:arcrm3731


```
<?xml version="1.0" encoding="UTF-8" ?>
- <OAI-PMH xmlns="http://www.openarchives.org/OAI/2.0/" xmlns:xsi="http://www.w3.org/2001/
  XMLSchema-instance" xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/ http://
  www.openarchives.org/OAI/2.0/OAI-PMH.xsd">
  <responseDate>2003-05-31T14:11:50+00:00</responseDate>
  <request metadataPrefix="oai_dc" verb="GetRecord" identifier="oai:magicnrc:arcrm3731">http://
    ntrs.nasa.gov/index.cgi</request>
  - <GetRecord>
    - <record>
      - <header>
          <identifier>oai:magicnrc:arcrm3731</identifier>
          <datestamp>2002-10-31</datestamp>
        </header>
      - <metadata>
        - <oai_dc:dc xmlns:oai_dc="http://www.openarchives.org/OAI/2.0/oai_dc/"
            xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xmlns:dc="http://
            purl.org/dc/elements/1.1/" xsi:schemaLocation="http://www.openarchives.org/
            OAI/2.0/oai_dc/ http://www.openarchives.org/OAI/2.0/oai_dc.xsd">
            <dc:title>On the flow in an isentropic light piston tunnel</dc:title>
            <dc:type>ARC/R&M-3731</dc:type>
            <dc:creator>T. V. Jones, D. L. Schultz and A. D. Hendley</dc:creator>
            <dc:contributor>Aeronautical Research Council, Great Britain</dc:contributor>
            <dc:date>1973</dc:date>
            <dc:identifier>http://naca.central.cranfield.ac.uk/reports/arc/rm/
              3731.pdf</dc:identifier>
            <dc:description>No Abstract Available</dc:description>
          </oai_dc:dc>
        </metadata>
      - <about>
        - <provenance xmlns="http://www.openarchives.org/OAI/2.0/provenance"
            xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xsi:schemaLocation=
```


direct pdf

Out of Band Knowledge

- pros: tailored, no “accidental” harvesting
- cons: not scalable wrt # of repositories & harvesters, false negatives

	no metadata change	metadata change
no data change	ok	unnecessary download
data change	missed update!	ok

assumption: change in metadata means a change in data -- not always true!

Base64 Encoding

- define separate metadata formats
  - base64:application/pdf
  - base64:application/powerpoint
- pros: describable with OAI-PMH semantics, accomplished with standard OAI-PMH tools
- cons: heavyweight (could use compression), suitable for simple objects only, accidental harvesting would produce high loads for repositories and harvesters

Metadata as Instructions

Back Forward Stop Refresh Home AutoFill Print Mail

Address: http://www.archive.org/movies/movies-details-db.php?collection=prelinger&collectionid=19069

INTERNET ARCHIVE

Web | Moving Images | Texts | Audio | Software | Patron Info

Prelinger Archives | Computer Chronicles | SIGGRAPH Electronic Theatre | Net Café | World at War

Search: Moving Images GO! Advanced Search Anonymous User (login or join us)

View movie scenes
Run Time: 9:15

Watch movie:
Modem (64kb)
Broadband (256k)
Mpeg4 Streaming (55M)

Download movie via FreeCache *(beta test)*:
MPG (93M)
MPEG (245M)
AVI (30M)

Download movie:
MPG (93M)
MPEG (245M)
AVI (30M)

Download w/ IAFM [help]

This program is on:

Duck and Cover 1951

Famous Civil Defense film for children in which Bert the Turtle shows what to do in case of atomic attack.

Sponsor: U.S. Federal Civil Defense Administration
Producer: Archer Productions, Inc.
Audio/Visual: Sd, B&W
Keywords: Atomic-nuclear; Civil defense; Animation

Average User Rating: ★★★★☆ **Viewed** 60,350 **times**

Reviews

★★★☆☆ April 27, 2003
Reviewer: Rocketeer
Subject: One of the Most Misunderstood Films of All Time
Perhaps no other film in history has generated more urban legends and mythology than *Duck and Cover*.

Randell Smith of Texas illuminates some of these myths in his review so I won't repeat them here. I'll just say that those who think the techniques taught in the film would be useless in an atomic attack literally don't know what they're talking about.

For every type of explosion there is a fringe area where survival is possible if the right steps are

cf. http://genomebiology.com/2003/4/6/R40

Metadata as Instructions

- the resource described in <dc:identifier> could be a complex object
  - may not be appropriate to:
    - “tar” the object into a single file
    - expose all constituent objects through OAI-PMH
  - define a metadata prefix that provides machine readable instructions on how extract the complex object
    - METS?

Metadata as Instructions

Back Forward Stop Refresh Home AutoFill Print Mail

Address: file://localhost/Users/mln/papers/naca-example.xml


```
<?xml version="1.0" ?>
- <OAI-PMH>
  <!-- deletia -->
  - <GetRecord>
    - <header>
        <identifier>oai:naca.larc.nasa.gov:1958:naca-tn-4410</identifier>
      </header>
    - <metadata>
      - <extractNACA>
          <pdf>http://naca.larc.nasa.gov/reports/1958/naca-tn-4410/naca-tn-4410.pdf</pdf>
          <thumbnails range="00001-00035" append="-t.gif" />
          <pages range="00001-00035" append=".gif" />
        </extractNACA>
      </metadata>
    </GetRecord>
  </OAI-PMH>
```

XSLT

- if the resource is already XML encoded, include an XSLT to transform into the desired format
  - use separate metadata formats or even sets for the harvester to express their transformation preferences?
- pros: elegant, limited work for repository
- cons: assumes client-side transformation capability, applicable only for XML-encodable resources