Using OAI-PMH for *Resource* Exchange

OAI Metadata Harvesting Workshop, JCDL 03

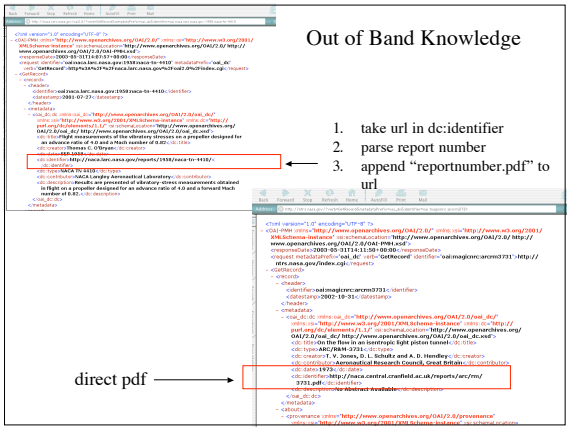
Michael L. Nelson, Terry L. Harrison
Old Dominion University
Norfolk VA
{mln,tharriso}@cs.odu.edu

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

- 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

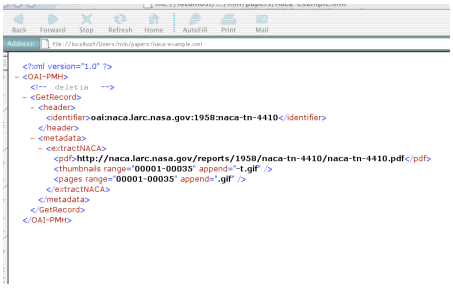
Metadata as Instructions

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



```
<?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