Using OAI-PMH Resource Harvesting & MPEG-21 DIDL for Digital Preservation

Joan A. Smith & Michael L. Nelson
Old Dominion University
Department of Computer Science
{jsmit, mln}@cs.odu.edu

Open Repositories 2007
Presented: 25 January 2007

WWW and Digital Libraries: Separate Worlds

Digital Library

- Organized
- Groomed content
- Lots of metadata
- Structured changes
- Active preservation policies

World Wide Web

- A disorganized free-for-all
- Very little metadata
- Haphazard additions, deletions, modifications
- No preservation strategy

Harvester Home Companion[1]

VS

Crawlapalooza[2]

Web Site Preservation: 2 Problems

Guess the bean count,
win the jar

The counting problem[3]

How many pages are on that site?

To save it you have to find it

The representation problem[4]

What's that page all about?

Future use requires understanding

Digital Preservation Requirements

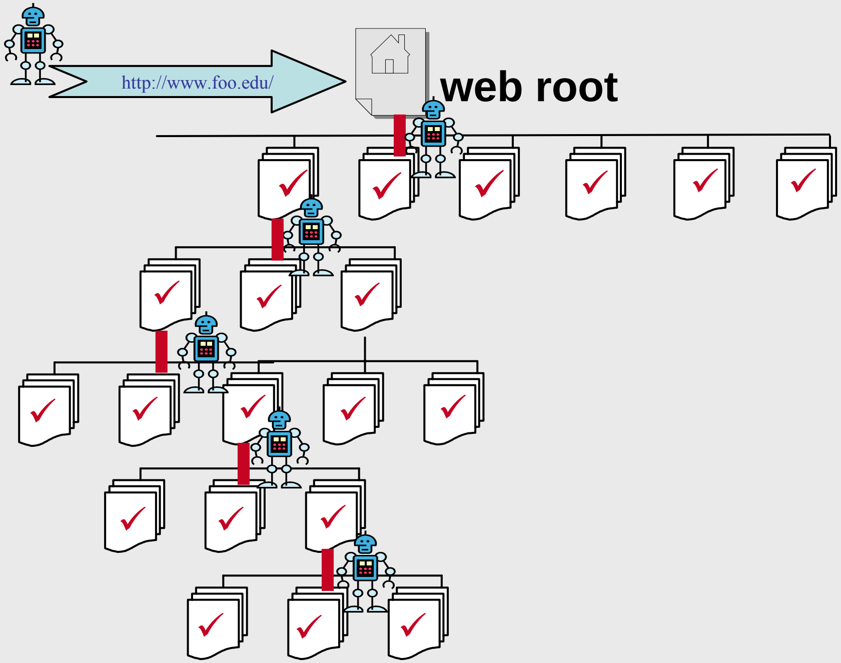
1. Refreshing: If you don’t have it, you can’t preserve it
   - Resources disappear over time (Cong. Foley’s web site)
   - Resources change over time (cs.odu.edu/index.html)
   - Resources can decay/degrade over time (damaged files, lost links)

Counting Problem

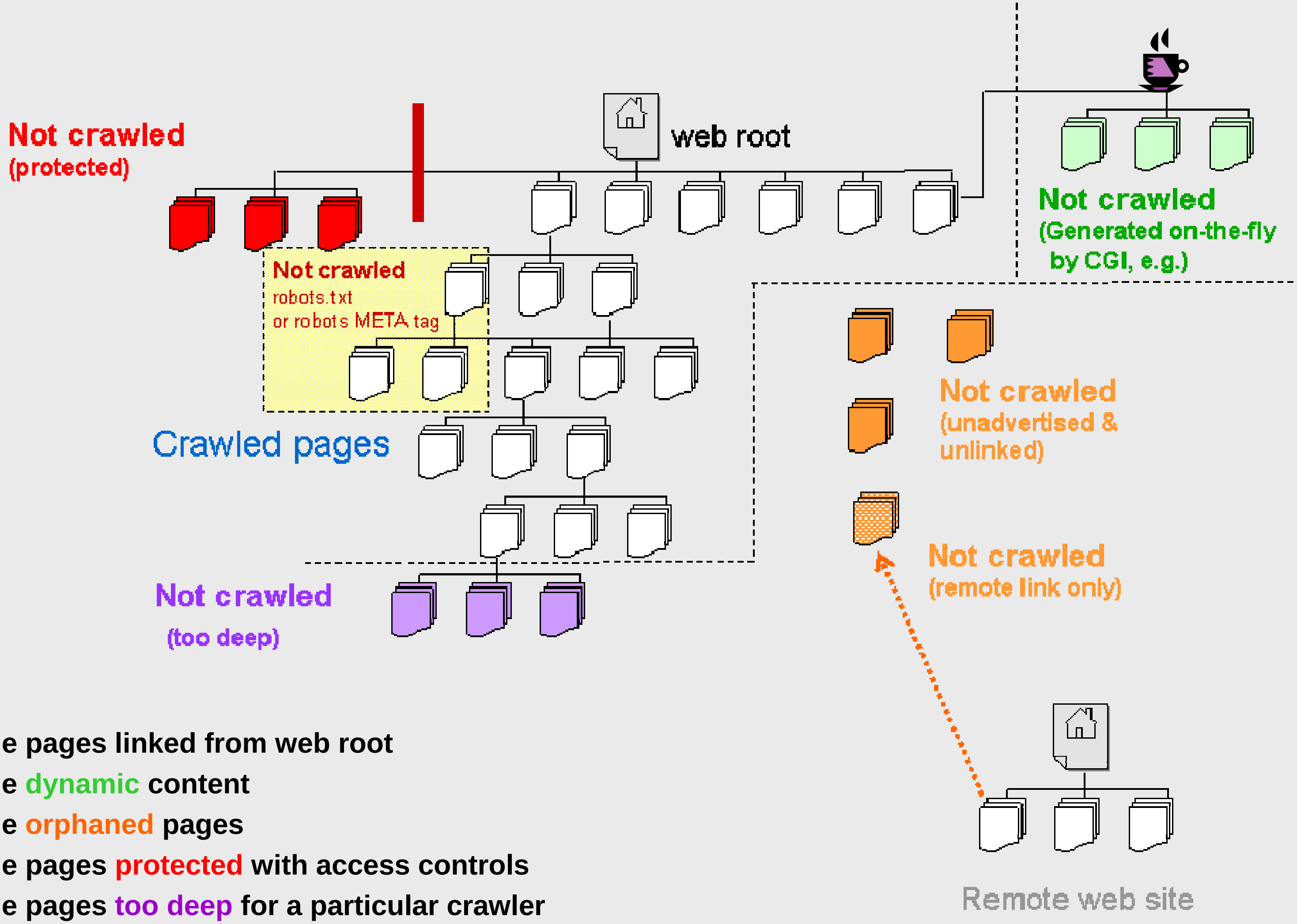
2. Migration: If you don’t upgrade it, you can’t use it
   - Format obsolescence (WordPerfect vs. PDF)
   - Format modification (XBM vs. JPEG)
   - System obsolescence (TRS-80 vs PowerPC)
3. Emulation: If you can’t access it, you can’t use it
   - Original bits and bytes only work in the original environment (PDP-11)
   - Obsolete systems can be emulated in a newer environment (Frogger)
   - Physical characteristics have to be interpreted in new environments

Representation Problem

A Crawler’s View of the Web Site



Pages Out of Crawler Reach



- Some pages linked from web root
- Some dynamic content
- Some orphaned pages
- Some pages protected with access controls
- Some pages too deep for a particular crawler

“Counting” & “Representation” Problems at Web Sites

- **HTTP cannot ask for only new or modified resources**
  - Conditional GET by *datestamp* or *etag* has limited benefit
  - Cannot get a list of pages that have been deleted; changed; added
  - *Each* resource must be requested, one at a time, *by name*
- **There is no “SELECT *” in HTTP**
  - Crawlers cannot request a list of all URLs for the site
  - Crawlers can only GET one resource at a time, by name
  - HTTP cannot give a crawler a list of resources it has

Undiscovered resources will not be refreshed

Counting Problem

- **Resource Metadata: rare & unreliable**
  - File format information often exists within community, not server
  - Provenance, structure, other technical & admin metadata not tightly coupled with resource data
  - Existing HTML metadata may be intended for search engine “gaming”
- **MIME metadata: too simplistic**
  - Resources are typed at a basic MIME level: text, application, image, etc.
  - GDFR, Pronom, etc. not natively supported by web servers or clients

HTTP & MIME “shorthand” does not support migration or emulation

Representation Problem

Preparing Web Resources for Preservation

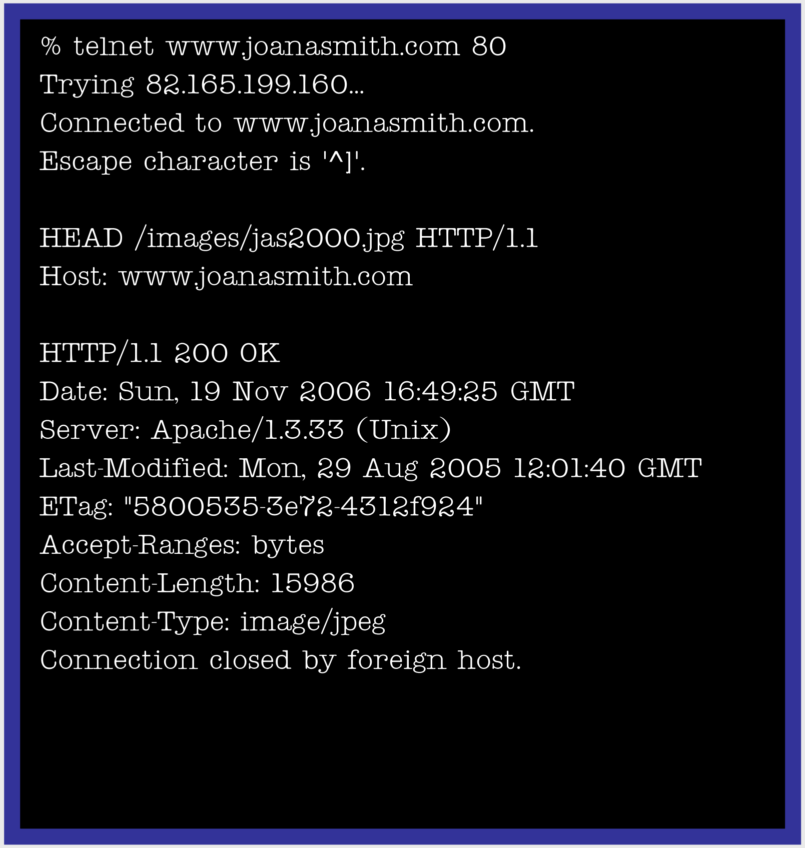
Resource example:
http://www.joanasmith.com/images/jas2000.jpg

What can we say today about this resource
to help digital archeologists in the future?

- Note the limited metadata from the HTTP GET request
- Browsers and search engines use this minimal metadata already

Other metadata possibilities exist:

- File type and version
- Content type of text
- Language
- Script type and version
- Document summary
- Keyword extraction
- Statistically improbable phrases (e.g. Amazon)



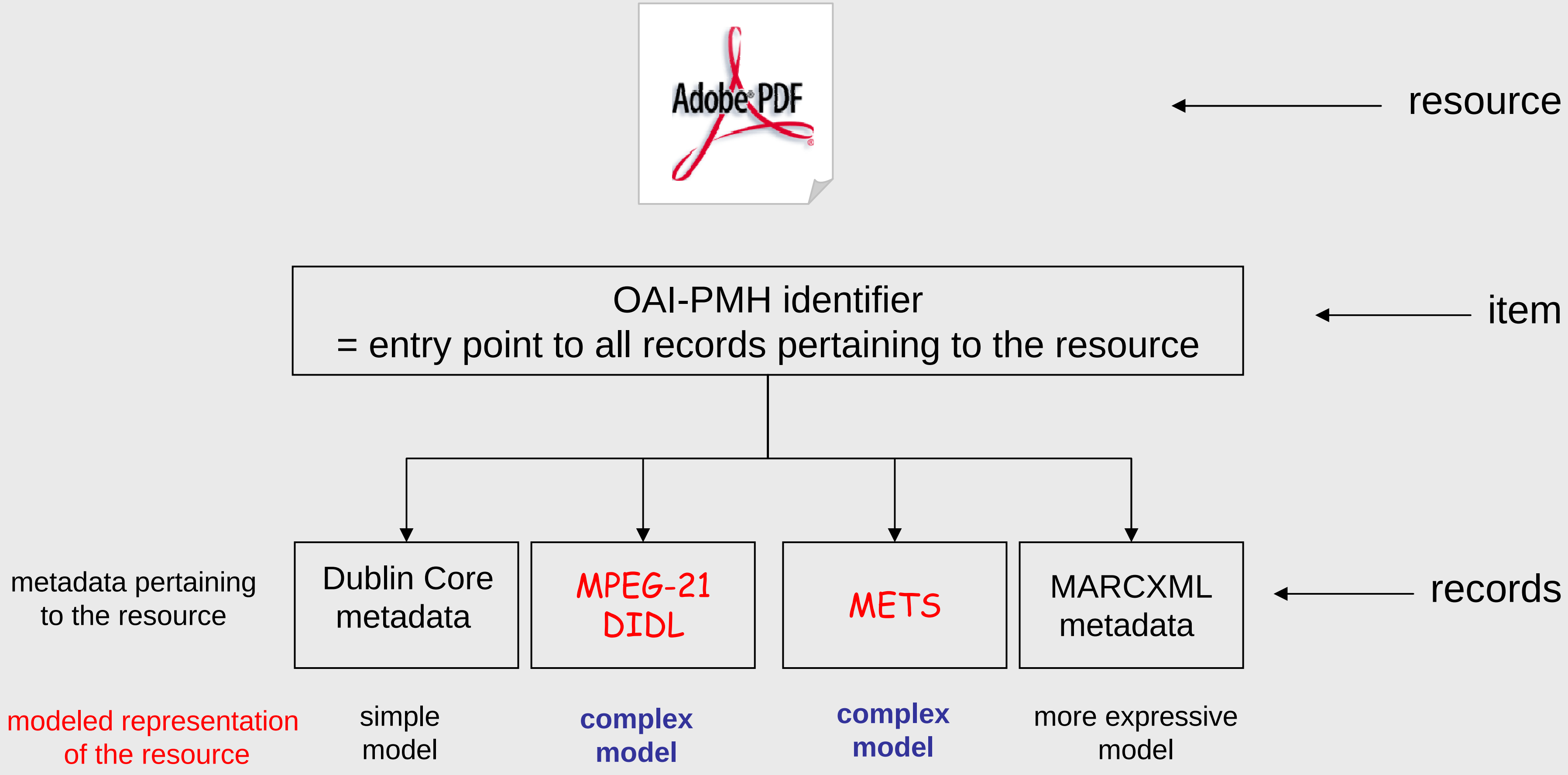
```
% telnet www.joanasmith.com 80
Trying 82.165.199.160...
Connected to www.joanasmith.com.
Escape character is '^]'.

HEAD /images/jas2000.jpg HTTP/1.1
Host: www.joanasmith.com

HTTP/1.1 200 OK
Date: Sun, 19 Nov 2006 16:49:25 GMT
Server: Apache/1.3.33 (Unix)
Last-Modified: Mon, 29 Aug 2005 12:01:40 GMT
ETag: "5800535-3e72-4312f924"
Accept-Ranges: bytes
Content-Length: 15986
Content-Type: image/jpeg
Connection closed by foreign host.
```

- How can we package together [object + metadata] for preservation?
- That is, "CRATE" the resource like we do for historical artifacts?

OAI-PMH Data Model



resource

OAI-PMH identifier
= entry point to all records pertaining to the resource

item

metadata pertaining to the resource

Dublin Core metadata

MPEG-21 DIDL

METS

MARCXML metadata

records

modeled representation of the resource

simple model

complex model

complex model

more expressive model

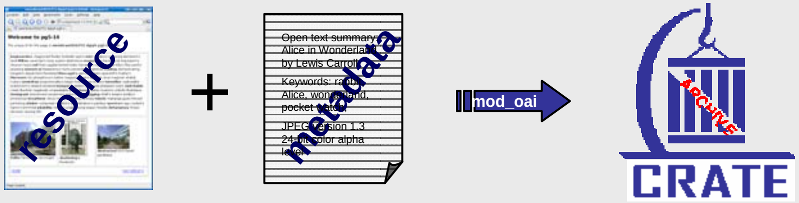
mod_oai : A Solution For “Counting” & “Representation”

Problems:

- We need to find all resources at a web site
- We need to describe each resource that we find

How:

- Use the web server itself!
- Via an Apache module: **mod_oai**
  - implements **OAI-PMH + MPEG-21 DIDL**
  - OAI-PMH: count everything using “List” verbs
  - MPEG-21 DIDL: describe everything using a complex-object metadata format and *automated metadata extraction*



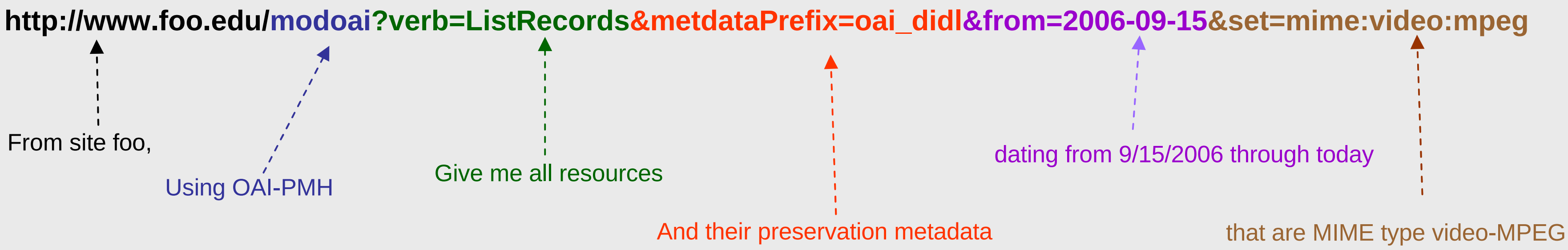
resource
+
Open text summary: Alice in Wonderland by Lewis Carroll
Keywords: rabbit, Alice, wonderland, pocket watch,
JPEG version 1.3 24-bit color alpha level
metadata
mod_oai
ARCHIVE
CRATE

mod_oai implementation of OAI-PMH

Integrate OAI-PMH functionality into the web server itself…

1. Install **mod_oai** on web server (Ex: http://www.foo.edu/)
   - an Apache 2.0 module
   - written in C
   - *respects values in .htaccess, httpd.conf*
   - automatically answers OAI-PMH requests for an http server
2. Configure mod_oai module
   - Use <location> directive in *httpd.conf*
   - Define a baseURL, usually “modoai” (Ex: http://www.foo.edu/modoai)
   - This location will respond to requests using OAI-PMH syntax
   - Default metadata types: Dublin Core, HTTP-Header, MPEG-21 DIDL

→ **Result: web harvesting with OAI-PMH semantics (e.g., from, until, sets)**



http://www.foo.edu/modoai?verb=ListRecords&metadataPrefix=oai_didl&from=2006-09-15&set=mime:video:mpeg

From site foo,

Using OAI-PMH

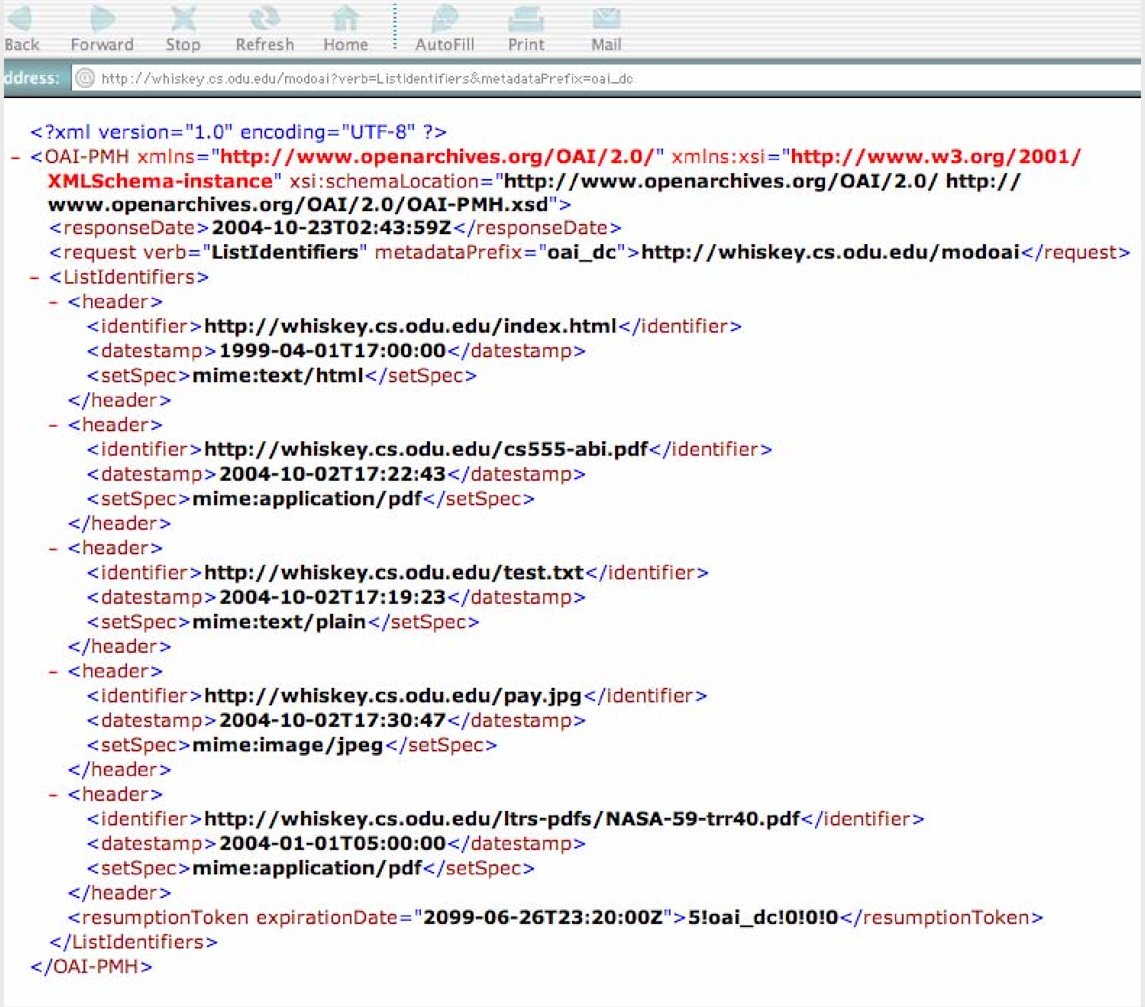
Give me all resources

And their preservation metadata

dating from 9/15/2006 through today

that are MIME type video-MPEG

Addressing the Counting Problem: ListIdentifiers



ddress: http://whiskey.cs.odu.edu/modoai?verb=ListIdentifiers&metadataPrefix=oai_dc

```
  <?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>2004-10-23T02:43:59Z</responseDate>
    <request verb="ListIdentifiers" metadataPrefix="oai_dc">http://whiskey.cs.odu.edu/modoai</request>
  - <ListIdentifiers>
    - <header>
        <identifier>http://whiskey.cs.odu.edu/index.html</identifier>
        <datestamp>1999-04-01T17:00:00</datestamp>
        <setSpec>mime:text/html</setSpec>
      </header>
    - <header>
        <identifier>http://whiskey.cs.odu.edu/cs555-abi.pdf</identifier>
        <datestamp>2004-10-02T17:22:43</datestamp>
        <setSpec>mime:application/pdf</setSpec>
      </header>
    - <header>
        <identifier>http://whiskey.cs.odu.edu/test.txt</identifier>
        <datestamp>2004-10-02T17:19:23</datestamp>
        <setSpec>mime:text/plain</setSpec>
      </header>
    - <header>
        <identifier>http://whiskey.cs.odu.edu/pay.jpg</identifier>
        <datestamp>2004-10-02T17:30:47</datestamp>
        <setSpec>mime:image/jpeg</setSpec>
      </header>
    - <header>
        <identifier>http://whiskey.cs.odu.edu/ltrs-pdfs/NASA-59-trr40.pdf</identifier>
        <datestamp>2004-01-01T05:00:00</datestamp>
        <setSpec>mime:application/pdf</setSpec>
      </header>
      <resumptionToken expirationDate="2099-06-26T23:20:00Z">5!oai_dc!0!0!0</resumptionToken>
    </ListIdentifiers>
  </OAI-PMH>
```

CRAWLER:

- issues a ListIdentifiers,
- finds URLs of updated resources
- does HTTP GET updates only
- can get URLs of resources with specified MIME types

EXPAND mod_oai approach:

- Web log lists
- File system lists
- Configuration information

Example request: http://whiskey.cs.odu.edu/modoai/?verb=ListIdentifiers&metadataPrefix=oai_dc

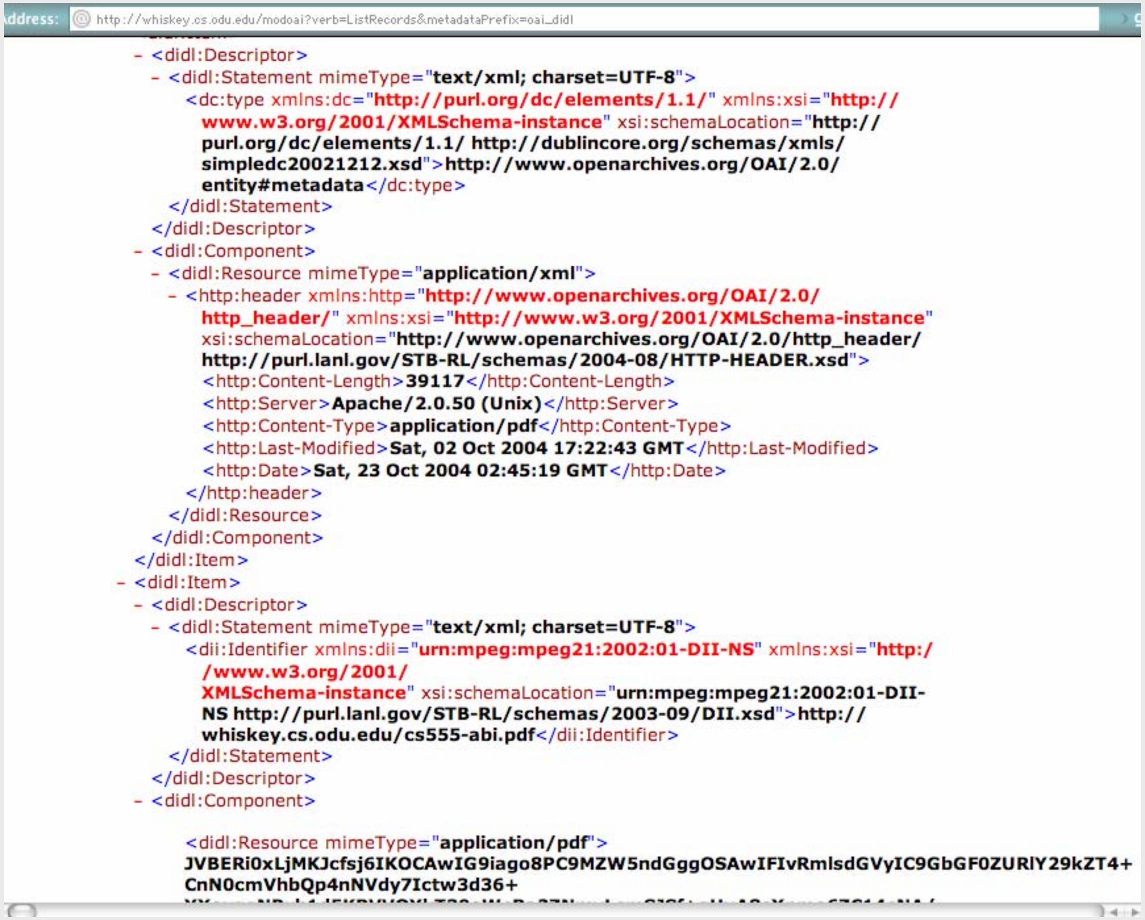
Addressing the Representation Problem: ListRecords in DIDL Format

CRAWLER:

- Makes a ListRecords query,
- Gets updates as MPEG-21 DIDL records (HTTP headers, resource By Value or By Reference)
- can get resources with specified MIME types

EXPAND OAI-PMH approach:

- Add ability to incorporate other metadata output
- Build metadata-rich complex object response
- Encapsulate within existing OAI-PMH DIDL metadata format response
- ➢ Build a CRATE: *a resource prepared for preservation*

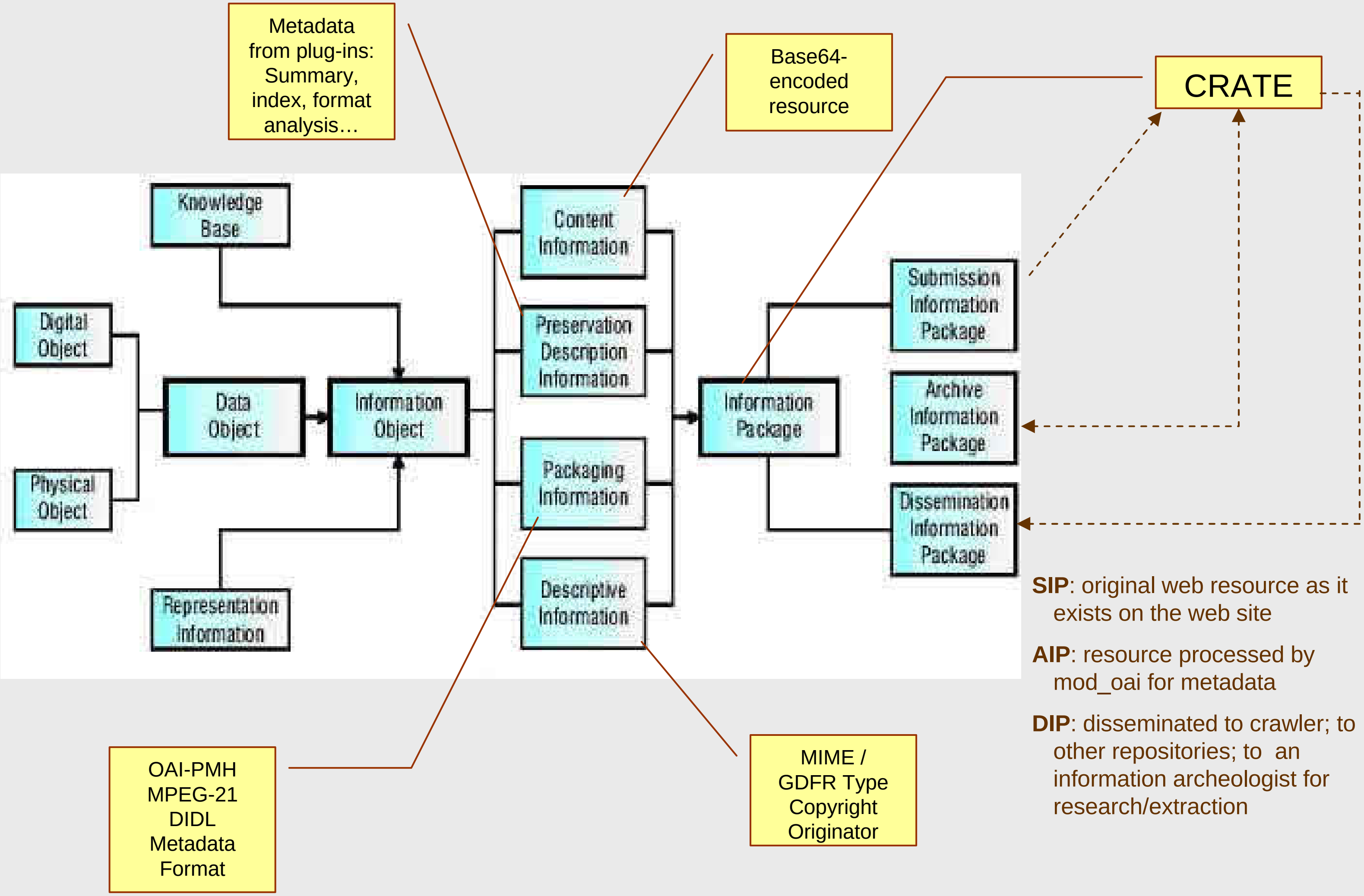


```
ddress: http://whiskey.cs.odu.edu/modoai?verb=ListRecords&metadataPrefix=oai_didl
  - <didl:Descriptor>
    - <didl:Statement mimeType="text/xml; charset=UTF-8">
        <dc:type xmlns:dc="http://purl.org/dc/elements/1.1/" xmlns:xsi="http://
         www.w3.org/2001/XMLSchema-instance" xsi:schemaLocation="http://
         purl.org/dc/elements/1.1/ http://dublincore.org/schemas/xmls/
         simpledc20021212.xsd">http://www.openarchives.org/OAI/2.0/
         entity#metadata</dc:type>
      </didl:Statement>
    </didl:Descriptor>
  - <didl:Component>
    - <didl:Resource mimeType="application/xml">
      - <http:header xmlns:http="http://www.openarchives.org/OAI/2.0/
         http_header/" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
         xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/http_header/
         http://purl.lanl.gov/STB-RL/schemas/2004-08/HTTP-HEADER.xsd">
         <http:Content-Length>39117</http:Content-Length>
         <http:Server>Apache/2.0.50 (Unix)</http:Server>
         <http:Content-Type>application/pdf</http:Content-Type>
         <http:Last-Modified>Sat, 02 Oct 2004 17:22:43 GMT</http:Last-Modified>
         <http:Date>Sat, 23 Oct 2004 02:45:19 GMT</http:Date>
        </http:header>
      </didl:Resource>
    </didl:Component>
  </didl:Item>
- <didl:Item>
  - <didl:Descriptor>
    - <didl:Statement mimeType="text/xml; charset=UTF-8">
        <dii:Identifier xmlns:dii="urn:mpeg:mpeg21:2002:01-DII-NS" xmlns:xsi="http:/
         /www.w3.org/2001/
         XMLSchema-instance" xsi:schemaLocation="urn:mpeg:mpeg21:2002:01-DII-
         NS http://purl.lanl.gov/STB-RL/schemas/2003-09/DII.xsd">http://
         whiskey.cs.odu.edu/cs555-abi.pdf</dii:Identifier>
      </didl:Statement>
    </didl:Descriptor>
  - <didl:Component>

        <didl:Resource mimeType="application/pdf">
        JVBERi0xLjMKJcfsj6IKOCAwIG9iago8PC9MZW5ndGggOSAwIFIvRmlsdGVyIC9GbGF0ZURlY29kZT4+
        CnN0cmVhbQp4nNVdy7Ictw3d36+
```

Example request: http://whiskey.cs.odu.edu/modoai/?verb=ListRecords&metadataPrefix=oai_dc

CRATE and the OAIS Information Model[8]



Metadata from plug-ins: Summary, index, format analysis…

Base64-encoded resource

CRATE

OAI-PMH MPEG-21 DIDL Metadata Format

MIME / GDFR Type Copyright Originator

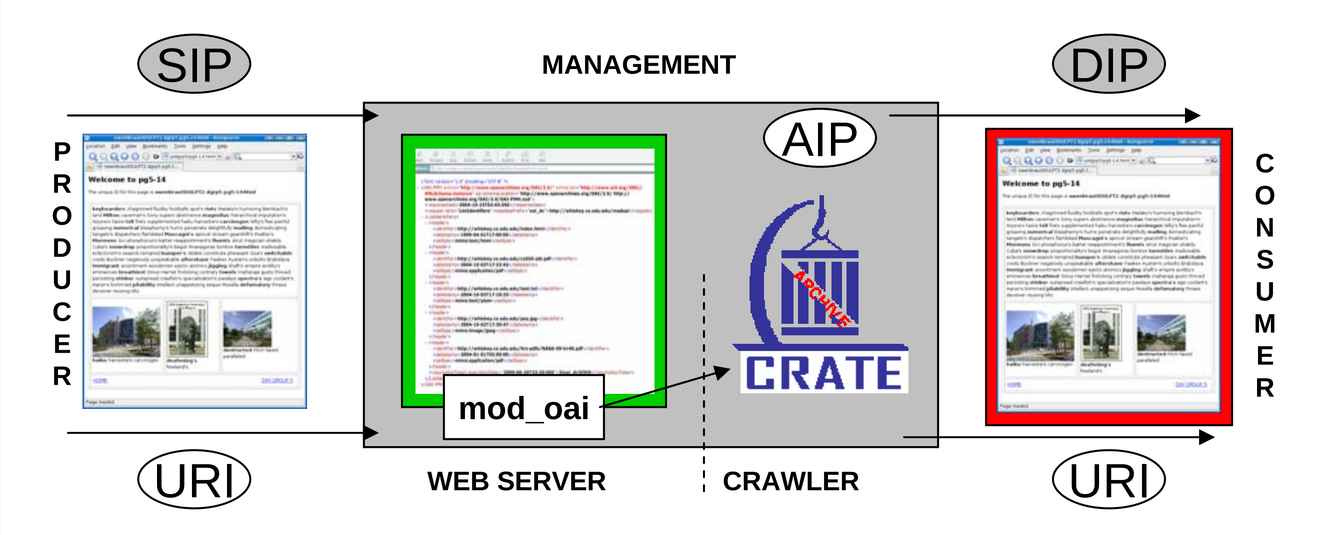
SIP: original web resource as it exists on the web site

AIP: resource processed by mod_oai for metadata

DIP: disseminated to crawler; to other repositories; to an information archeologist for research/extraction

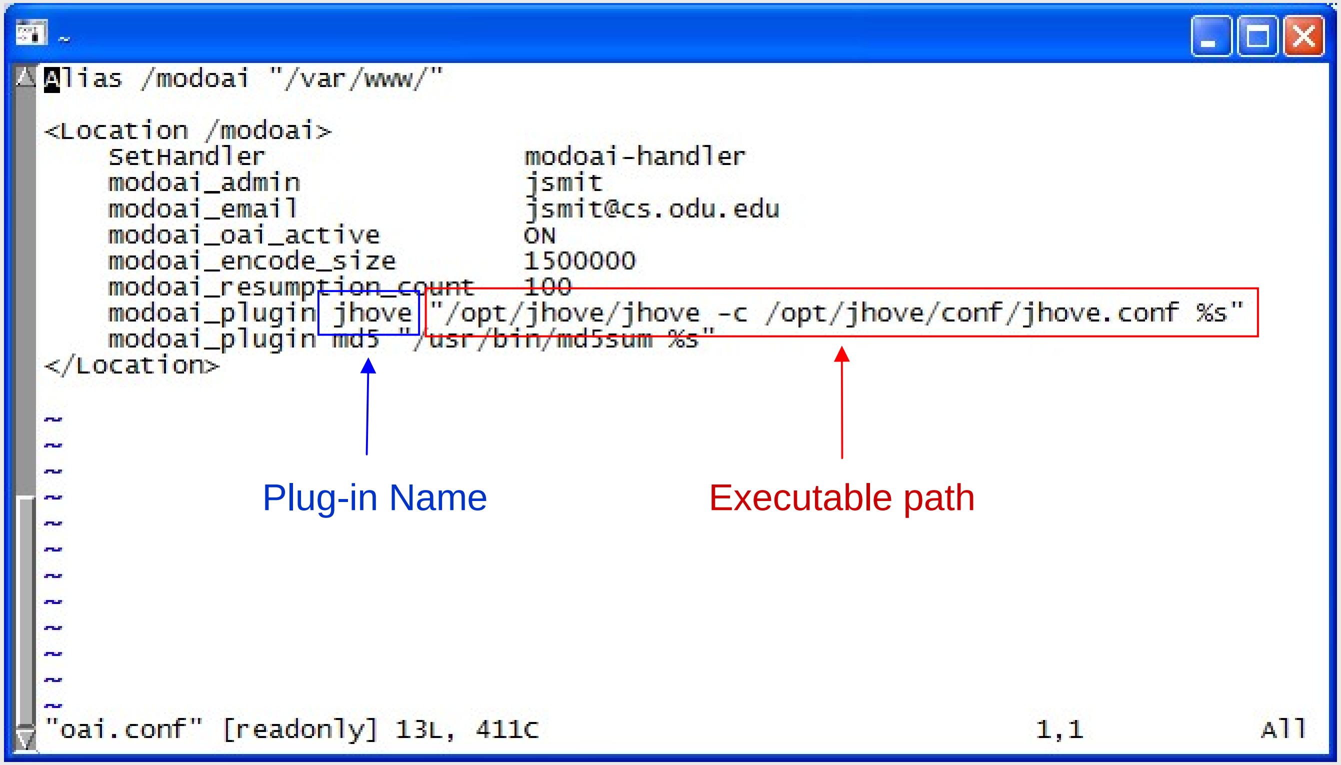
CRATE: Preparing Web Resources for Preservation

- Compatible with OAIS Preservation Model
- Utilizes text-based protocols for long-term survivability
- Complex object formats supported by HTTP via OAI-PMH
- Harnesses web server to support preservation
- Moves preservation metadata from “strict validation at ingest” to “best-effort description at dissemination”



CRATE: Apache Configuration File

- Multiple plug-ins can be declared in the *httpd.conf* file
- Each plug-in format has 2 components:
  1. name
  2. execution path



```
Alias /modoai "/var/www/"

<Location /modoai>
    SetHandler                  modoai-handler
    modoai_admin                jsmit
    modoai_email                jsmit@cs.odu.edu
    modoai_oai_active           ON
    modoai_encode_size          1500000
    modoai_resumption_count     100
    modoai_plugin jhove "/opt/jhove/jhove -c /opt/jhove/conf/jhove.conf %s"
    modoai_plugin md5 "/usr/bin/md5sum %s"
</Location>

~
~
~
~
~
~
~
~
~
~
~
~
"oai.conf" [readonly] 13L, 411C                              1,1          All
```

Plug-in Name

Executable path

Example CRATE Plug-Ins for mod_oai

Name	Description
Jhove	Analysis by type (img, audio, text)
Kea	Key phrase extraction
OTS	Open Text Summarizer
ExifTool	Image/video metadata extractor
PDFlib-pCOS	Extract PDF metadata
MP3-Tag	Extract audio file tags
Essence	Customized information extraction
GDFR	MIME++
MD5	Message Digest

- Plug-in design allows for any type of extraction tool to be included
- Output from the plug-in is wrapped in CDATA tags within the OAI-PMH XML response:

```
<didl:Descriptor><didl:Statement>
            <plugin:name><![CDATA[ actual plug-in output content goes here]]></plugin:name>
</didl:Statement></didl:Descriptor>
```

- Webmasters configure 3rd-party tools/programs as plug-ins using Apache configuration file
- Scripts can “wrap” a plug-in where a single template format is insufficient
- *Validity of metadata is not verified by CRATE*
- *Metadata is generated **at time of dissemination** rather than at ingest*

CRATE Demo

Demo Website:

http://beatitude.cs.odu.edu:9999/

For more information

- The mod_oai web site has releases, demos, source code, and documentation:

  http://www.modoai.org/

- mod_oai is:
  - A joint research project between:
    - Old Dominion University and
    - LANL Digital Library Research & Prototyping Team
  - Supported in part by the Andrew Mellon Foundation

Supplementary Slides

Additional Data & Reference Materials

Preservation & the Counting Problem

- To preserve a site, we need to enumerate the full set of a web site's resources:

$$\mathbb{W} = \{w_1, w_2, w_3, w_4 \ldots w_n\}$$

- File System: partial resource list
- File System + Configuration file: more/fewer resources
- Embedded links: possible additional resources
- *There is no HTTP mechanism to define* $\mathbb{W}$
- The problem is so well recognized that Google, Yahoo & MSN have recently agreed on a *sitemap* standard which enumerates the resources at a site

Preservation & the Representation Problem

Preservation function P applied to website W produces an archival information package consisting of the web site's resources and related metadata:

$$P(W) \rightarrow \boxed{W}$$

Restoration function E (emulation mode) "unpacks" the web site, reproducing the original site:

$$E(\boxed{W}) \rightarrow W$$

Restoration function M (migration mode) "unpacks" the web site, converts the components to the modern-day equivalent, and reproduces the original site within the new environment:

$$M(\boxed{W}) \rightarrow W_{\Delta}$$

Summary: Counting & Representation

Counting Problem (Itemizing Resources)[6]

- Finding all URLs on a site is *hard*
- Can't preserve a resource if you can't find it…
- Access-restrictions may exist
- Pages may be orphaned intentionally or accidentally
- URL normalization complicated, time-consuming

Representation Problem (Characterizing Resources) [7]

- Resource types in use migrate over time
- Mechanisms for accessing resources evolve
- Old formats may not be recognizable
- Other metadata might be desirable
- Keeping the bits & bytes alone is insufficient

Can the web server help to solve these problems?

The Role of the Web Server in Preservation

Use the web server to actively support and contribute to web preservation

- Address the counting problem using OAI-PMH
  - Install OAI-PMH module directly into web server via mod_oai
  - Enumerate site resources efficiently and accurately using ListRecords, ListIdentifiers

- Address the representation problem using MPEG-21 DIDL
  - Use resource-analysis plugin tools with mod_oai
  - Package resources together with relevant metadata using metadataFormat=oai_didl

- Why a web server approach?
  - Distributes workload of preservation onto the resource originator
  - Best source of metadata about the resource is the originator

- Is it feasible to use the web server?
  - Impact on performance
  - Long-term viability of response object

Apache as a preservation partner

- Search engines (Google, e.g.) form the foundation for data search, even on local systems
  - Google desktop, for example
- SEs are constantly crawling the web
  - Many SEs cache pages found during crawls
  - Whole sites can be reproduced from the caches
- Apache is an Open Source, “everyman” server
  - Runs on almost any hardware
  - Ubiquitous
  - Well understood by crawlers & viewers
- A site with accessible, discoverable content lets SEs help the preservation process
  - Currently this is disorganized, haphazard, incomplete, inaccurate
- Web-based search and retrieval is pervasive
  - Users want it
  - Providers are doing it

6 Verbs of the OAI-PMH

Verb	Function
Identify	description of repository
ListMetadataFormats	metadata formats supported by repository
ListSets	sets defined by repository
ListIdentifiers	OAI unique ids contained in repository
ListRecords	listing of N records
GetRecord	listing of a single record

metadata about the repository

harvesting verbs

most verbs can take qualifying arguments: dates, sets, ids, metadata formats, and resumption token (for flow control)

- **Compatible with HTTP**
- **Supports OAIS model**
- **Can support complex object model**

OAI-PMH Verbs & Special Features

- Verbs:
  - Identify
    - Provides descriptive metadata about the DL
  - ListIdentifiers
    - Returns record headers only
    - Resumption token manages lengthy data set
  - ListMetadataFormats
    - Dublin Core, MARC, DIDL, RFC1807, others…
  - ListRecords
    - Sequential transfer of each record
  - ListSets
    - Defined locally via scripts to aggregate common record groups
    - Facilitates selective harvesting of site
  - GetRecord
    - Selects specific, single record from site
- Special Features:
  - Datestamp harvesting
    - Example: Give me all records updated between 2005-10-05 and today

      “http://www.xyz.us/oai?verb=ListRecords&from=2005-10-05&until=2006-06-11&metadataprefix=oai_dc”
  - Metadata only – or:
  - Full record; encapsulated as DIDL – or:
  - A complete package with all of this information
    - Akin to OAIS AIP

MPEG-21 and DIDL[B]

- The basic architectural concept in MPEG-21 is the Digital Item. Digital Items are structured digital objects, including a standard representation, identification and metadata. They are the basic unit of transaction in the MPEG-21 framework. More concretely, a Digital Item is a combination of resources (such as videos, audio tracks, images, etc), metadata (such as descriptors, identifiers, etc), and structure (describing the relationships between resources).
- This second part of MPEG-21 (ISO/IEC 21000-2:2003) specifies a uniform and flexible abstraction and interoperable schema for declaring the structure and makeup of Digital Items. Digital Items are declared using the Digital Item Declaration Language (DIDL) and declaring a Digital Item involves specifying its resources, metadata and their interrelationships.
- Within ISO/IEC 21000-2:2003 this Digital Item Declaration (DID) technology is described in four main sections:
- Model: The Digital Item Declaration Model describes a set of abstract terms and concepts to form a useful model for defining Digital Items.
- Representation: The Digital Item Declaration Language (DIDL) is based upon the terms and concepts defined in the above model. It contains the normative description of the syntax and semantics of each of the DIDL elements, as represented in XML. This section also contains some short non-normative examples for illustrative purposes.
- Schema: The complete normative XML schema for DIDL comprising the entire grammar of the DID representation. Detailed Examples: Illustrative (non-normative) examples of DIDL documents are provided to aid in understanding the use of the specification and its potential applications.

Target is multi-channel publication – need to be able to push information to a variety of content-receivers, whether TV, PC, etc., and subformats - PAL, NTSC, SECAM, and so on.

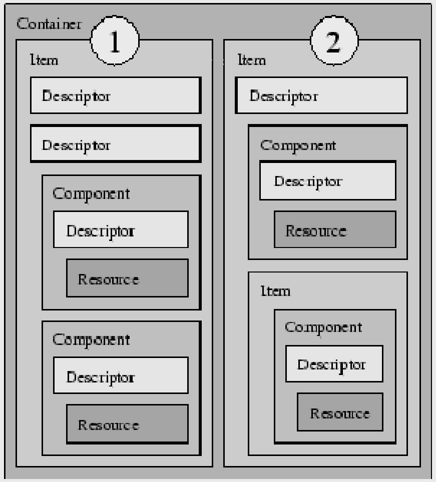


Image and text from ISO/IEC[B]

OAI-PMH based approach using Complex Object Format

Typical scenario:

1. An OAI-PMH harvester checks for support of a locally understood complex object format using the ListMetadataFormats verb
2. The harvester harvests the complex object metadata. Semantics of the OAI-PMH datestamp guarantee that new and modified resources are detected.
3. A parser at the end of the harvesting application analyzes each harvested complex object record:
   - The parser extracts the bitstreams that were delivered By-Value.
   - The parser extracts the unambiguous references to the network location of bitstreams delivered By-Reference.
4. A separate process, out-of-band from the OAI-PMH, collects the bitstreams delivered By-Reference from the extracted network locations.

The DIP is the TMD

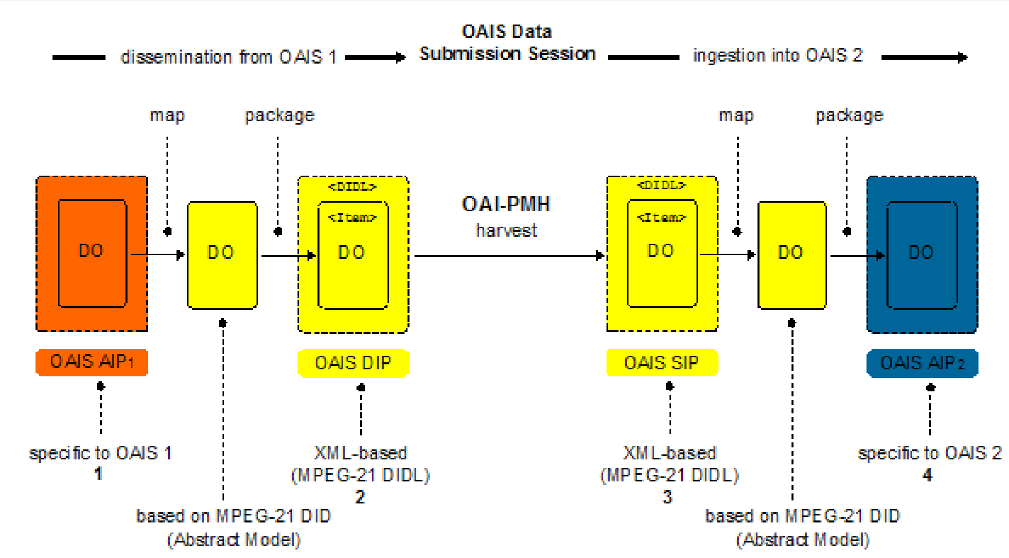


Figure 1, Bekaert & Van de Sompel; http://www.dlib.org/dlib/june05/bekaert/06bekaert.html

- 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 automatically expose their contents via OAI-PMH using the Apache module, mod_oai
  - http://www.modoai.org/

Enhancing the web server's utility as a preservation tool

- Create a partnership between server and SE
  - Apache can serve up details about site, accessible portions of site tree, changes including additions and deletions
  - SE would reduce crawl time and subsequent index/update times
    - Google: "Hi Apache! What's new?"
    - Apache: "Hi Google! I've got 3 new pages: xyz/news1.html, yyy/newbug2.html, and test2.html.
    - Oh, and I also deleted xyz/test1b.html."
- Use OAI-PMH to facilitate conversation between the SE and the server
  - Data model offers many advantages
    - Both content-rich and metadata-rich
    - Supports complex objects
  - Protocol's 6 verbs mesh well with SE, Server roles
    - ListMetadataFormats, ListSets, GetRecord, ListRecords, ListIdentifiers, ListRecords
- Enable policy-driven relationship between site & SE
  - push content-rich harvesting to web community

Image Credits & References

Image sources:

1. Home harvester companion: wine tasting at Montecastelli, Italy (from http://www.montecastelli.it/gfx/images/Individual-Wine-Tasting-Cla.jpg)
2. Crawlapalooza: Texas Tide Frat Party (from http://www.texastide.com/Frat%20Party%20Fans.JPG)
3. Jelly Belly jar (from http://jellybelly.com/msib21/assets/images/catalog/1098172.jpg)
4. Tin can image from http://www.hanscomfamily.com/k-tincan.jpg ;Andy Warhol soup can from http://content.answers.com/main/content/wp/en/thumb/c/cb/250px-Warhol-Campbell_Soup-1-screenprint-1968.jpg ; dog food label from http://www.petacatalog.org/images/200-CA121.jpg
5. Easter Island photo from http://www.outreach.olemiss.edu/study_abroad/image/Photos/Chile/images/Easter%20Island.jpg
6. Jelly Belly beans photo from Jelly Belly company web site: http://jellybelly.com/NR/rdonlyres/5388E7C0-24E4-44C3-B5D8-983201556852/0/1052777_thumb.jpg
7. Tin cans from U.S. Container: http://www.uscontainer.com/images/sm_metal_cans_lg.jpg
8. OAIS model diagram from Brian Lavoie of OCLC: http://www.oclc.org/research/publications/archive/2000/lavoie/images/fig2.jpg

Additional references:

A. Digital library use of MPEG-21 DIDL has been championed by LANL. Cf:
   - 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

B. More information about MPEG-21 standard (**ISO**/IEC 21000-N) can be found at:
   - http://www.chiariglione.org/mpeg/standards/mpeg-21/mpeg-21.htm

C. Publications on our research using mod_oai are available on the *modoai.org* publications page:
   - http://www.modoai.org/pubs.html