Tools for a Preservation-Ready Web

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

NDIIPP Digital Preservation Partners Meeting
July 9, 2008

What is Preservation?

- We will define preservation of a web site *W* to be:
  - refreshing
    - copying the bits from place to place
    - R(W) = W_r
  - migrating
    - converting the bits from format f_1 to format f_2
    - M(W) = W_m
  - emulation
    - simulating the original context for the bits
    - E(W) = W_e
  - putting it all together:
    - E(M(R(W))) = W_{rme}

Preservation Function *P*

- We define a preservation function *P*
  - P(W) = W_p

- Intuition is that P makes other functions easier:

  M(W_p) is easier to implement than M(W)

  E(W_p) is easier to implement than E(W)

  R(W_p) is probably easier to implement than R(W)

Web Site Preservation: 2 Problems

The counting problem

What are the members of W?

The representation problem

How do we define P(W)?

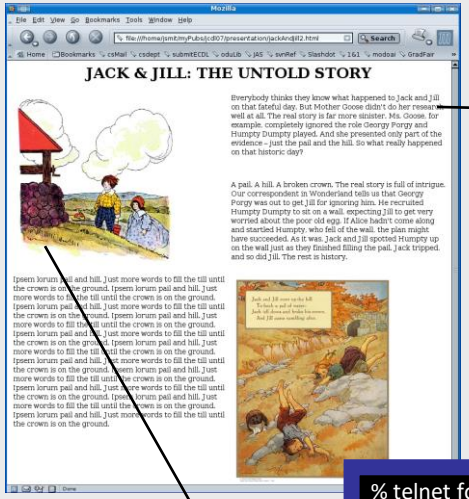
Preservation & the Counting Problem

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

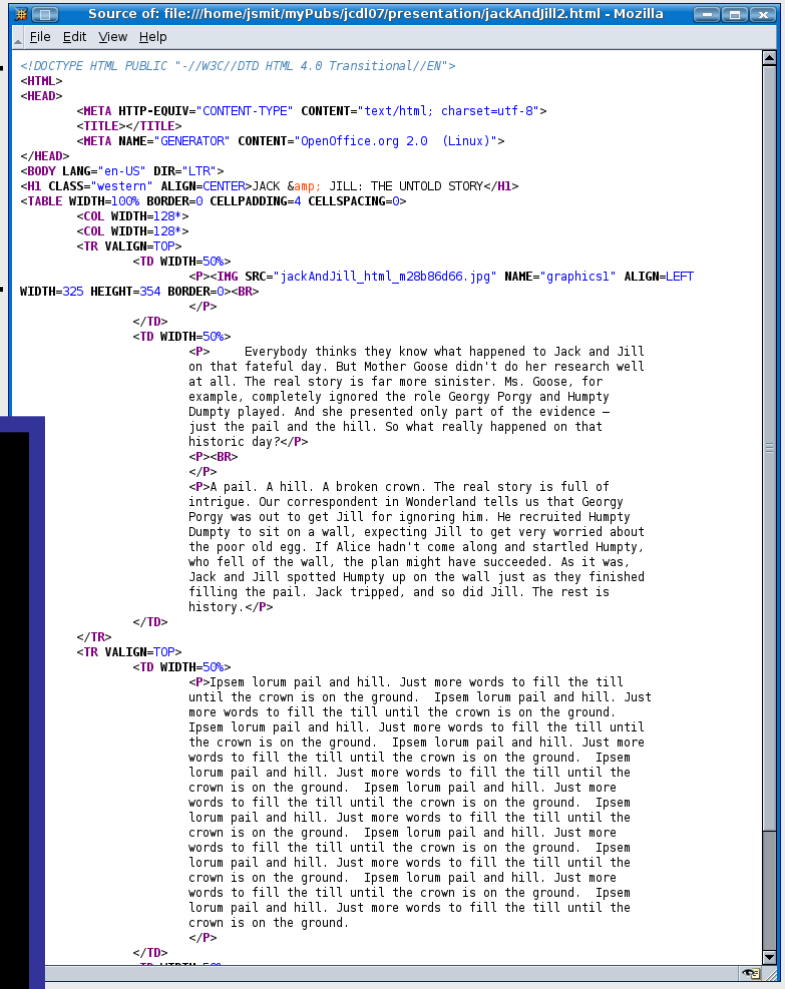
$$W = \{w_1, w_2, w_3, w_4 \ldots w_n\}$$

- For non trivial web sites:
  - The membership of W depends on who is asking
  - W is unknown (unknowable?)
  - W can only be approximated
- *There is no HTTP mechanism to define* W
- Sitemaps are a method to convey locally-held knowledge about W to web crawlers

Preservation & the Representation Problem



HTML metadata



HTTP metadata

```
% telnet foo.edu 80
Trying 82.165.199.160...
Connected to foo.edu.
Escape character is '^]'.

GET /jackJill.jpg HTTP/1.1
Host: foo.edu

HTTP/1.1 200 OK
Date: Mon, 11 Jun 2007 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

ÿØÿà
"#2s¡35Rq‘±³ÁÂ$%Ccruƒ“ĉÃÒÿÄ
```

P(W) Involves the Output of Forensic Metadata Utilities

Standard HTTP Headers --

Last-Modified: Mon, 29 Aug 2005 12:01:40 GMT
ETag: "5800535-3e72-4312f924"
Content-Length: 15986
Content-Type: image/jpeg

EXIF:

File Name	103_0315.JPG
Camera Model Name	Canon EOS DIGITAL REBEL
Date/Time Original	2003:09:30 13:37:51
Shooting Mode	Sports
Shutter Speed	1/2000
Aperture	7.1
Metering Mode	Evaluative
Exposure Compensation	0
ISO	400
Lens	75.0 - 300.0mm
Focal Length	300.0mm
Image Size	3072x2048
Quality	Normal
Flash	Off
White Balance	Auto
Focus Mode	AI Servo AF
Contrast	+1
Sharpness	+1
Saturation	+1
Color Tone	Normal
File Size	1606 kB
File Number	103-0315

MD5 Hash:
58a54e8638db432f4515eedf89f44505

File/Magic:
JPEG image data
JFIF standard 1.00
resolution (DPI)
"LEAD Technologies Inc. V1.01“
33 x 26

JHOVE:
Date: 2007-06-18 14:35:50 EDT RepresentationInformation: /home/crate/apache/htdocs/jackJill.jpg ReportingModule: JPEG-hul, Rel. 1.2 (2005-08-22) LastModified: 2007-01-16 23:09:07 EST Size: 27750 Format: JPEG Version: 1.00 Status: Well-Formed and valid SignatureMatches: JPEG-hul MIMEtype: image/jpeg Profile: JFIF JPEGMetadata: CompressionType: Huffman coding, Baseline DCT Images: Number: 1 Image: NisoImageMetadata: MIMEType: image/jpeg ByteOrder: big-endian CompressionScheme: JPEG ColorSpace: YCbCr SamplingFrequencyUnit: inch XSamplingFrequency: 33 YSamplingFrequency: 26 ImageWidth: 172 ImageLength: 146 BitsPerSample: 8, 8, 8 SamplesPerPixel: 3 Scans: 1 QuantizationTables: QuantizationTable: Precision: 8-bit DestinationIdentifier: 0 Comments: LEAD Technologies Inc. V1.01 ApplicationSegments: APP0

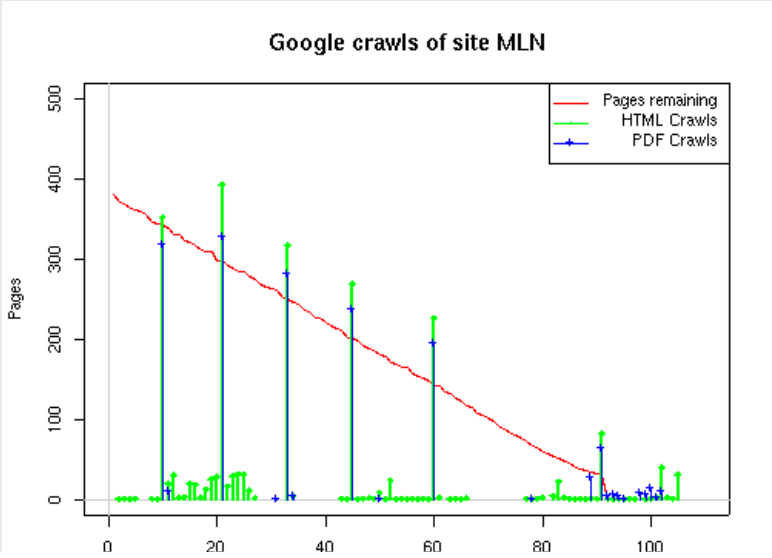
Experiments & Evaluation

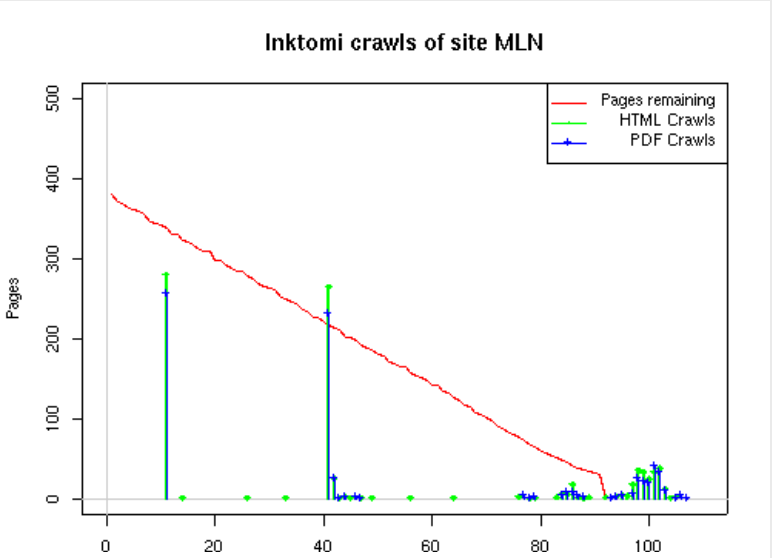
- Counting problem
  - Web crawler behavior on decaying web sites (D-Lib 2006)
  - Web crawler behavior on deep and wide web sites (D-Lib 2008)
  - Defining W on a departmental web site (unpublished)
- Representation problem
  - Dissemination time preservation metadata (JCDL 2007, IWAW 2007, D-Lib 2008)
  - Performance evaluation of metadata utilities (ECDL 2008)

- Reference implementation: mod_oai, an Apache module
  - uses Sitemaps, OAI-PMH resource harvesting for counting problem
  - uses “CRATE” -- base64’d resource + metadata output as the OAI-PMH metadataPrefix for representation problem

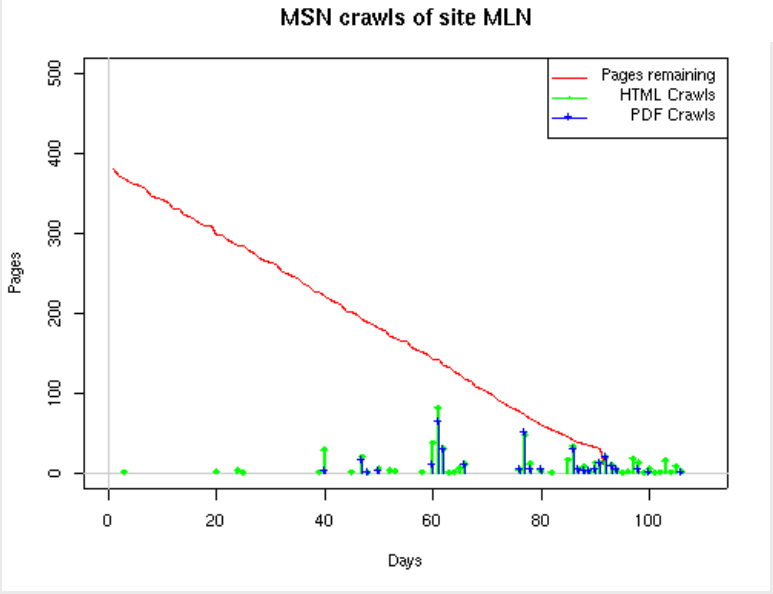
Decaying Web Sites

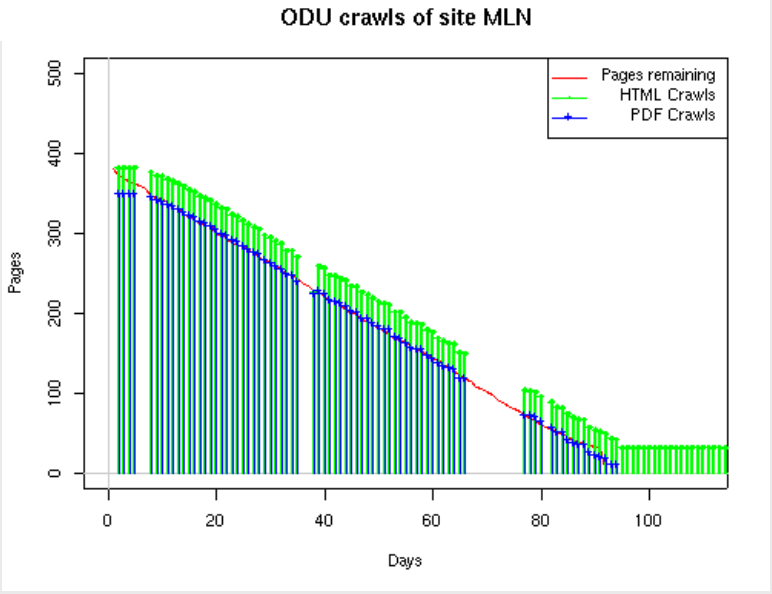
(D-Lib 2006)





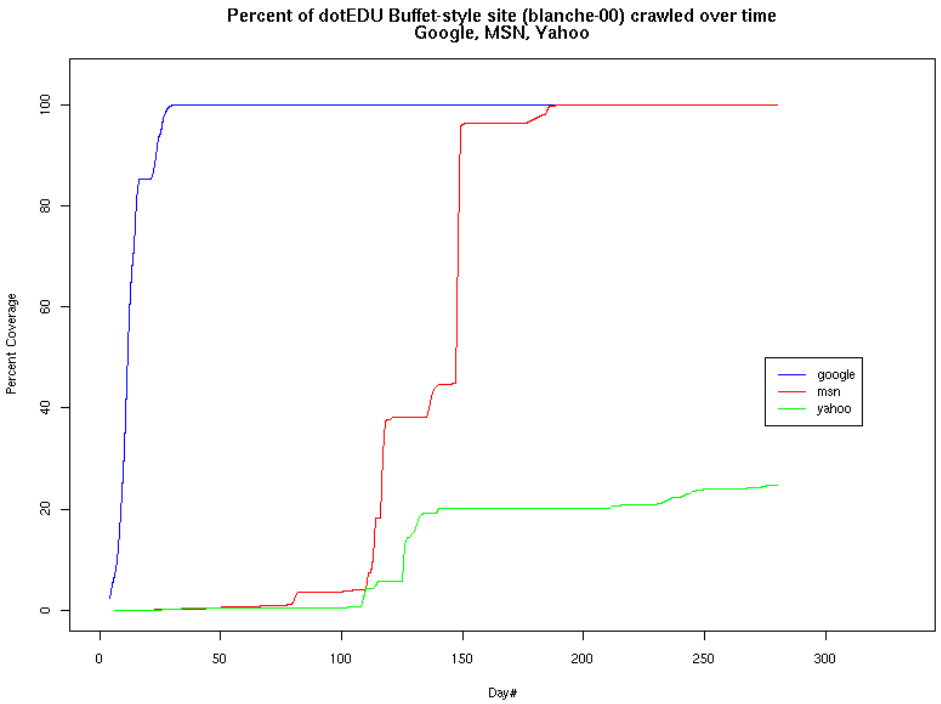
- Lots of pages die in between crawler visits
- IA never came in 3+ months

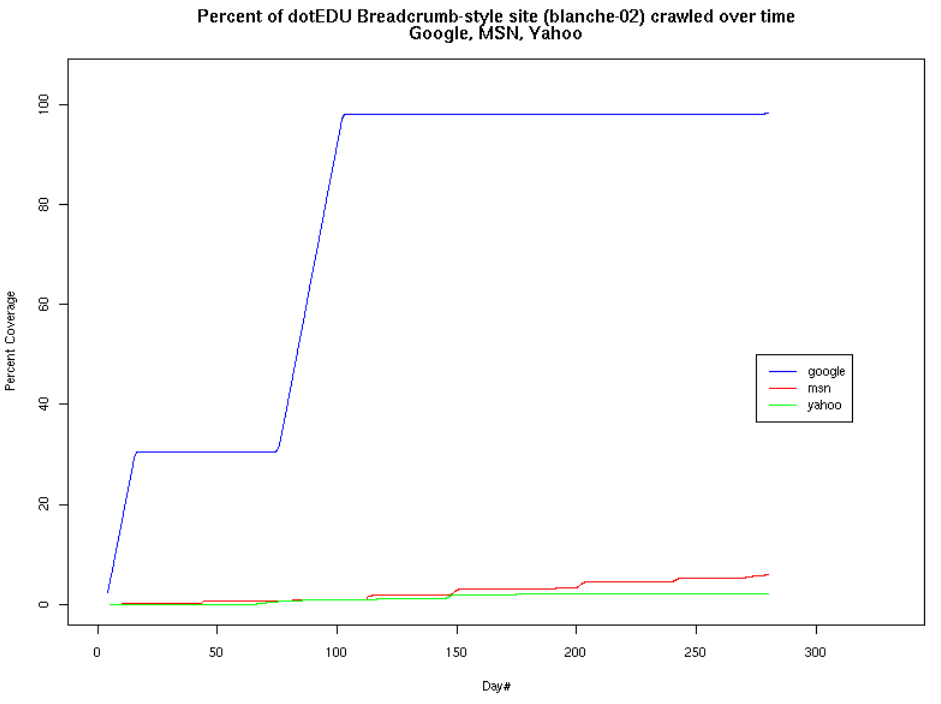




Deep & Wide Web Sites

(D-Lib 2008)

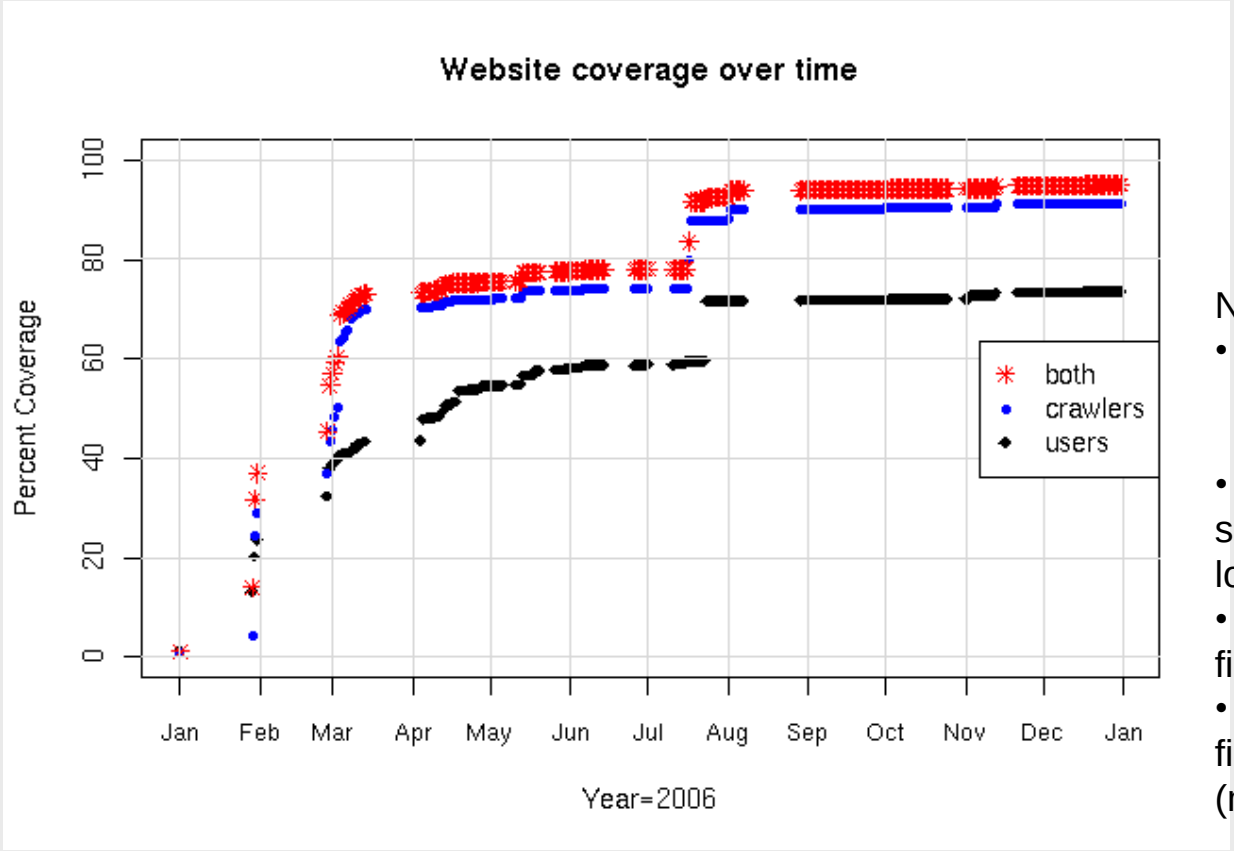




"Buffet" = Level 1 links to levels 2, 3, 4 … n

"Bread Crumb" = Level 1 links to level 2, level 2 links to level 3, etc.

Coverage of www.cs.odu.edu



Notes:
- Departmental snapshot (no ~user URLs; CGI files removed; spotty http logs)
- Google Python Sitemap script crashed on ill-formed log data
- 100% defined in terms of file system count
- Results written in a Sitemap file for mod_oai processing (more later)

Source	Files	URLs
Self-Crawl	406	538
External Crawl	406	761
File System	*2,052*	*2,052**

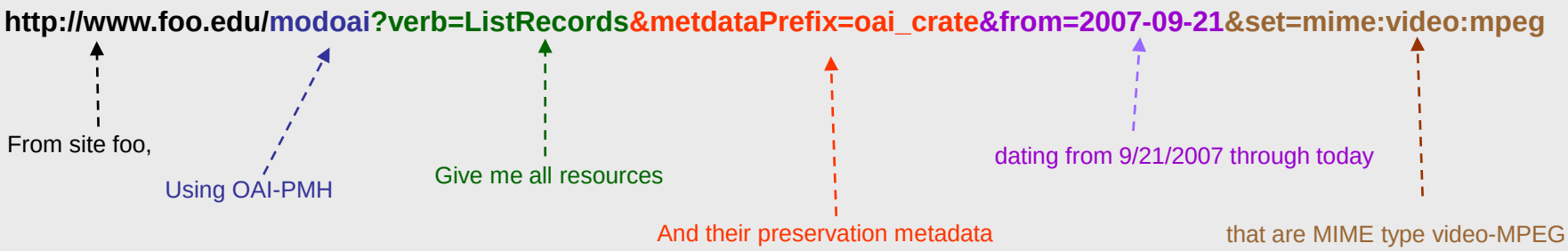
mod_oai implementation

(JCDL 2007, IWAW 2007, D-Lib 2008)

Integrate OAI-PMH functionality into the web server itself…

1. Use **mod_oai**
   - an Apache 2.0 module
   - automatically answers OAI-PMH requests for an http server
   - written in C
   - *respects values in .htaccess, httpd.conf*
2. Install mod_oai on http://www.foo.edu/
3. Define baseURL: http://www.foo.edu/modoai

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



Uses a public (or private) Sitemap for the definition of W (used to be dynamic file system inspection); create the Sitemap as the union of multiple tools

CRATE

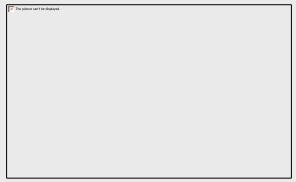
<crate>
  <http request headers> Accept: text/xml,application/xml,application/xhtml+xml,text/html;q=0.9,text/plain;q=0.8,image/png,*/*;q=0.5
Accept-Language: en-us,en;q=0.5
…
  <http response headers> Content-Length: 101702
Content-Type: image/jpeg
…
  <metadata>
    <name>
    <version>
    <output>
  </metadata>

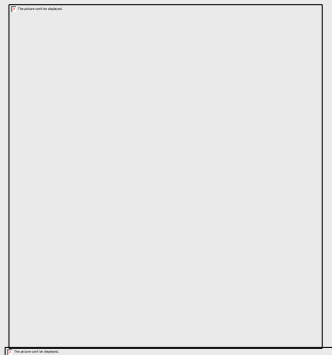
  <metadata>
    <name>
    <version>
    <output>
  </metadata>

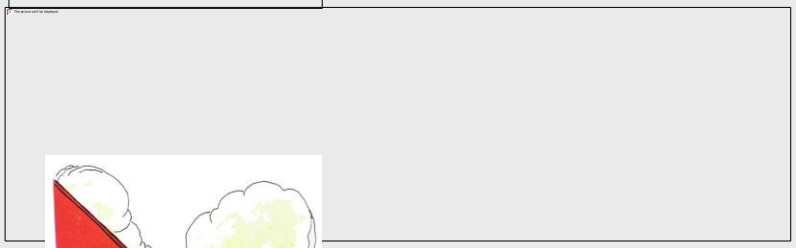
  <metadata>
    <name>
    <version>
    <output>
  </metadata>

  <base64'd resource> KNiAwIG9iago8PC9MZW5ndGgg….

</crate>

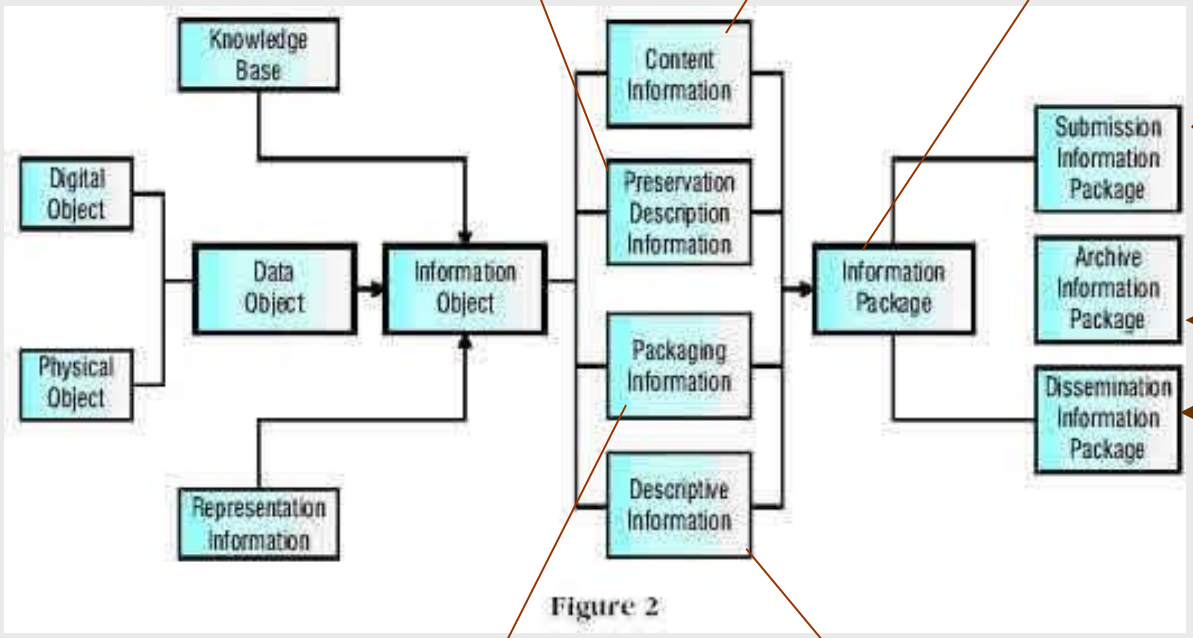






- If you like to look at lots of XML, go to: http://www.cratemodel.org/modoai/?verb=ListRecords&metadataPrefix=oai_crate
- Intuition: now is the best time to describe a file, and the local user has a good idea as to what should be said (does not preclude further analysis by harvester)

CRATE and the OAIS Information Model



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

Base64-encoded resource

CRATE

OAI-PMH MPEG-21 DIDL Metadata Format

MIME / GDFR Type Copyright Originator Description

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: Apache Configuration File

```
        <Location /modoai>                          ← Apply these rules to http://foo.edu/modoai
          SetHandler modoai-handler                 ← Use modoai to process these requests

on a    modoai_plugin                               ← plugin element: one utility per element
single      "jhove"                                 ← each has a label, used as a metadata "ID tag"
text        "/opt/jhove/jhove -m jpeg-hul %s"       ← the command-line or script to call the utility
line        "/opt/jhove/jhove --v"                  ← include the version number of the installed utility
            "image/jpeg"                            ← which MIME types should be analyzed (any jpeg)


                                                    ← EOL here
         modoai_plugin
            "ots"                                   Open Text Summarizer
            "/usr/local/bin ots –summary %s"        ← "%s" means substitute resource name here
            "/usr/local/bin ots -v"
            "text/*"                                ← Use on all text (plain, HTML, XML, etc.) resources

        modoai_plugin
            "jhove"                                 ← Another invocation of the JHOVE utility
            "/opt/jhove/jhove -m pdf-hul %s"        ← Note the different hul used here
            "/opt/jhove/jhove --v"                  ← report the version
            "application/pdf"                       ← Use on all PDF resources (only)

        modoai_plugin
            "pronom"                                ← the PRONOM DROID tool
            "java -jar DROID.jar -L%s –SsigFile.xml"
            "java -jar DROID.jar -v"                ← report the version
            "*/*"                                   ← Use this utility on every resource

        </Location /modoai>
```

Tested CRATE Plug-Ins for mod_oai

Name	Description
Exif	Image/video metadata extractor
Jhove	Image analysis
DC	dcTag html extactor
Droid	Pronom registry info
MetaX	Meta-extractor
OTS	Open Text Summarizer
wc	unix word count utility
file	unix file utility (magic cookie)
md5, sha	unix md5sum, shasum utilities

Time Required to CRATE Web Site

(ECDL 2008)

Server response time to other web requests: < 2% throughput delta

Request Parameters	Active Utilities	Response Time in Min:Sec By Server Load			Response Size (Bytes)
		0 %	50 %	100%	
wget (full crawl)	None	00:27.16s	00:28.55s	00:28.89s	77,982,064
ListIdentifiers:oai_dc	None	00:00.14s	00:00.46s	00:00.20s	130,357
ListRecords:oai_dc	None	00:00.34s	00:00.37s	00:00.37s	756,555
ListRecords:oai_crate	None	00:02.47s	00:08.34s	00:03.38s	106,148,676
ListRecords:oai_crate	File	00:09.56s	00:09.72s	00:09.50s	106,429,668
ListRecords:oai_crate	MD5sum	00:04.55s	00:04.52s	00:04.40s	106,278,907
ListRecords:oai_crate	SHA	00:19.36s	00:19.70s	00:19.96s	106,190,722
ListRecords:oai_crate	SHA-1	00:04.57s	00:04.49s	00:05.37s	106,316,236
ListRecords:oai_crate	WC	00:06.14s	00:06.11s	00:05.92s	106,419,750
ListRecords:oai_crate	Exif	00:04.60s	00:04.79s	00:04.51s	106,163,645
ListRecords:oai_crate	DC	00:31.13s	00:29.47s	00:28.66s	106,612,082
ListRecords:oai_crate	OTS	00:35.81s	00:36.43s	00:35.83s	106,285,422
ListRecords:oai_crate	MetaX	01:13.71s	01:15.99s	01:13.96s	106,257,162
ListRecords:oai_crate	Jhove	00:54.74s	00:54.99s	00:54.84s	106,297,738
ListRecords:oai_crate	Droid	44:14.01s	45:29.76s	47:23.29s	106,649,382
ListRecords:oai_crate	All *but Droid*	03:34.58s	03:38.84s	03:42.60s	107,906,032
ListRecords:oai_crate	All	47:42.45s	48:53.97s	50:09.76s	108,407,266

Future Work

- OAI-ORE support
  - CRATEs as Resource Maps
- Defining CRATEs as an http encoding format
  - like gzip, zip, etc.
  - can return a CRATE in response to a regular http request with appropriate q values (not just OAI-PMH harvest request)
- Third party metadata
  - how can my web server use your installation of Jhove?
- Tighter http log / Sitemap integration:
  - “Sitemap strict” -- don’t serve a file unless it appears in a Sitemap
  - “Sitemap synch” -- in real-time, add/delete entries in Sitemap based on 200 / 404 responses

For more information

- More info, code:

  http://www.modoai.org/

  http://code.google.com/p/modoai/

- A joint research project between:
  - Old Dominion University and
  - LANL Digital Library Research & Prototyping Team
- Research supported by the Andrew Mellon Foundation & the Library of Congress