7-9-2008

Tools for a Preservation-Ready Web

Joan A. Smith
Old Dominion University

Michael L. Nelson
Old Dominion University, mnelson@odu.edu

Follow this and additional works at: https://digitalcommons.odu.edu/computerscience_presentations

Part of the Archival Science Commons, and the Computer Sciences Commons

Recommended Citation
https://digitalcommons.odu.edu/computerscience_presentations/25

This Book is brought to you for free and open access by the Computer Science at ODU Digital Commons. It has been accepted for inclusion in Computer Science Presentations by an authorized administrator of ODU Digital Commons. For more information, please contact digitalcommons@odu.edu.
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
July 9, 2008
NDIIPP Partners Meeting

Slide # 6

Preservation & the Representation Problem

% 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
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

Everybody thinks they know what happened to Jack and Jill on that fateful day. But Mother Goose didn't do her research well at all! The real story is far more sinister. Mr. Goose, for example, completely ignored the role George Porogy and Humpty Dumpty played. And she presented only part of the evidence -- just the tail and the hill. So what really happened on that historic day?

A hill. A broken crown. The real story is full of intrigue. Our correspondent in Wonderland tells us that George Porogy was out to get Jill for ignoring him. He recruited Humpty Dumpty to sit on a wall, expecting Jill to get very worried about the poor old egg. If Alice hadn't come along and startled Humpty, who fell off the wall, the plan might have succeeded. As it was, Jack and Jill spotted Humpty up on the wall just as they finished filling the pot. Jack tripped, and so did Jill. The rest is history.

Preservation & the Representation Problem
P(W) Involves the Output of Forensic Metadata Utilities

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:
- 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

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 JPEG
- Metadata: 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

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

<table>
<thead>
<tr>
<th>Source</th>
<th>Files</th>
<th>URLs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Self-Crawl</td>
<td>406</td>
<td>538</td>
</tr>
<tr>
<td>External Crawl</td>
<td>406</td>
<td>761</td>
</tr>
<tr>
<td>File System</td>
<td>2,052</td>
<td>2,052*</td>
</tr>
</tbody>
</table>
mod_oai implementation

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)


From site foo, Using OAI-PMH

Give me all resources And their preservation metadata
dating from 9/21/2007 through today

that are MIME type video-MPEG

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 request headers>

<http response headers>
Content-Length: 101702
Content-Type: image/jpeg
...
</http response headers>

<metadata>
  <name>File/Magic:</name>
  <version>JP EG image data</version>
  <output>JFIF standard 1.00 resolution (DPI)</output>
  <name>”LEAD Technologies Inc. V1.01 Ò</name>
  <version>33 x 26</version>
</metadata>

<metadata>
  <name>File/Magic:</name>
  <version>JPEG image data</version>
  <output>JFIF standard 1.00 resolution (DPI)</output>
  <name>”LEAD Technologies Inc. V1.01 Ò</name>
  <version>33 x 26</version>
</metadata>

<metadata>
  <name>File/Magic:</name>
  <version>JPEG image data</version>
  <output>JFIF standard 1.00 resolution (DPI)</output>
  <name>”LEAD Technologies Inc. V1.01 Ò</name>
  <version>33 x 26</version>
</metadata>

<metadata>
  <name>JHOVE TOOL:</name>
  <version>Date: 2007-06-18 14:35:50 EDT RepresentationInformation:/home/crate/apache/htdocs/jackJill.jpg</version>
  <output>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: JFIFJPEGMetadata CompressionType Huffman coding, Baseline DCT Images: Number: 1 Image: NisoiImageMetadata CompressionType Huffman coding, Baseline DCT XSamplingFrequencyUnit: inch XSamplingFrequency: 33 YSamplingFrequencyUnit: inch YSamplingFrequency: 24 BitsPerSample: 8, 8, 8 SamplesPerPixel: 3 Encoding: 8-bit DestinationIdentifier: 0 ApplicationSegments: APP0
</metadata>

<base64’d resource>KNiAwIG9iago8PC9MZW5ndGgg….</base64’d resource>

• If you like to look at lots of XML, go to: http://www.cratemodel.org/modoai/?verb=List
  Records&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

Figure 2

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

OAI-PMH MPEG-21 DIDL Metadata Format

MIME / GDFR Type Copyright Originator Description
CRATE: Apache Configuration File

```xml
<Location /modoai>
  SetHandler modoai-handler
</Location>
```

- `<Location /modoai>`
- `SetHandler modoai-handler`

```xml
modoai_plugin
  "jhove"
  "/opt/jhove/jhove -m jpeg-hul %s"
  "/opt/jhove/jhove --v"
  "image/jpeg"
```

- `modoai_plugin`
  `- plugin element: one utility per element`
  `- each has a label, used as a metadata “ID tag”`
  `- the command-line or script to call the utility`
  `- include the version number of the installed utility`
  `- which MIME types should be analyzed (any jpeg)`

```xml
modoai_plugin
  "ots"
  "/usr/local/bin ots --summary %s"
  "/usr/local/bin ots -v"
  "text/*"
```

- `modoai_plugin`
  `- Use on all text (plain, HTML, XML, etc.) resources`

```xml
modoai_plugin
  "jhove"
  "/opt/jhove/jhove -m pdf-hul %s"
  "/opt/jhove/jhove --v"
  "application/pdf"
```

- `modoai_plugin`
  `- Another invocation of the JHOVE utility`
  `- Note the different hul used here`
  `- report the version`
  `- Use on all PDF resources (only)`

```xml
modoai_plugin
  "pronom"
  "java -jar DROID.jar -L%ss -SsigFile.xml"
  "java -jar DROID.jar -v"
  "*/*
```

- `modoai_plugin`
  `- the PRONOM DROID tool`
  `- report the version`
  `- Use this utility on every resource`

EOL here

```
Open Text Summarizer

"%s" means substitute resource name here

Use on all text (plain, HTML, XML, etc.) resources
```

Another invocation of the JHOVE utility

Note the different hul used here

report the version

Use on all PDF resources (only)

```
Tested CRATE Plug-Ins for mod_oai

<table>
<thead>
<tr>
<th>Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exif</td>
<td>Image/video metadata extractor</td>
</tr>
<tr>
<td>Jhove</td>
<td>Image analysis</td>
</tr>
<tr>
<td>DC</td>
<td>dcTag html extractor</td>
</tr>
<tr>
<td>Droid</td>
<td>Pronom registry info</td>
</tr>
<tr>
<td>MetaX</td>
<td>Meta-extractor</td>
</tr>
<tr>
<td>OTS</td>
<td>Open Text Summarizer</td>
</tr>
<tr>
<td>wc</td>
<td>unix word count utility</td>
</tr>
<tr>
<td>file</td>
<td>unix file utility (magic cookie)</td>
</tr>
<tr>
<td>md5, sha</td>
<td>unix md5sum, shasum utilities</td>
</tr>
</tbody>
</table>
Quantitative Evaluation of Using MODOAI to Build a CRATE

- Created “typical” website
  - 1084 resources – PDF, HTML, Applications, Images
    - Complete Sitemap file
- Tested in commercial environment (Kronos, Inc)
- Installed metadata utilities
  - Some Java
  - Some OS-Native
  - Some locally compiled
- Collected CPU performance data using Jmeter
- Compared CRATE with simple crawl
  - Time to complete crawl
  - Size of response
  - Response time by load variation
  - Impact on non-Crate requests
- Compared time for individual utilities
  - Response time by load factor
  - Response size by utility
### Time Required to CRATE Web Site

(ECDL 2008)

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

<table>
<thead>
<tr>
<th>Request Parameters</th>
<th>Active Utilities</th>
<th>Response Time in Min:Sec By Server Load</th>
<th>Response Size (Bytes)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>0%</td>
<td>50%</td>
</tr>
<tr>
<td>wget (full crawl)</td>
<td>None</td>
<td>00:27.16s</td>
<td>00:28.55s</td>
</tr>
<tr>
<td>ListIdentifiers:oai_dc</td>
<td>None</td>
<td>00:00.14s</td>
<td>00:00.46s</td>
</tr>
<tr>
<td>ListRecords:oai_dc</td>
<td>None</td>
<td>00:00.34s</td>
<td>00:00.37s</td>
</tr>
<tr>
<td>ListRecords:oai_crate</td>
<td>None</td>
<td>00:02.47s</td>
<td>00:08.34s</td>
</tr>
<tr>
<td>ListRecords:oai_crate</td>
<td>File</td>
<td>00:09.56s</td>
<td>00:09.72s</td>
</tr>
<tr>
<td>ListRecords:oai_crate</td>
<td>MD5sum</td>
<td>00:04.55s</td>
<td>00:04.52s</td>
</tr>
<tr>
<td>ListRecords:oai_crate</td>
<td>SHA</td>
<td>00:19.36s</td>
<td>00:19.70s</td>
</tr>
<tr>
<td>ListRecords:oai_crate</td>
<td>SHA-1</td>
<td>00:04.57s</td>
<td>00:04.49s</td>
</tr>
<tr>
<td>ListRecords:oai_crate</td>
<td>WC</td>
<td>00:06.14s</td>
<td>00:06.11s</td>
</tr>
<tr>
<td>ListRecords:oai_crate</td>
<td>Exif</td>
<td>00:04.60s</td>
<td>00:04.79s</td>
</tr>
<tr>
<td>ListRecords:oai_crate</td>
<td>DC</td>
<td>00:31.13s</td>
<td>00:29.47s</td>
</tr>
<tr>
<td>ListRecords:oai_crate</td>
<td>OTS</td>
<td>00:35.81s</td>
<td>00:36.43s</td>
</tr>
<tr>
<td>ListRecords:oai_crate</td>
<td>MetaX</td>
<td>01:13.71s</td>
<td>01:15.99s</td>
</tr>
<tr>
<td>ListRecords:oai_crate</td>
<td>Jhove</td>
<td>00:54.74s</td>
<td>00:54.99s</td>
</tr>
<tr>
<td>ListRecords:oai_crate</td>
<td>Droid</td>
<td>44:14.01s</td>
<td>45:29.76s</td>
</tr>
<tr>
<td>ListRecords:oai_crate</td>
<td>All but Droid</td>
<td>03:34.58s</td>
<td>03:38.84s</td>
</tr>
<tr>
<td>ListRecords:oai_crate</td>
<td>All</td>
<td>47:42.45s</td>
<td>48:53.97s</td>
</tr>
</tbody>
</table>
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://www.modoai.org/)

- 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