Integrating Preservation Functions into the Web Server

Joan A. Smith
Advisor: Michael L. Nelson
12 June 2008

ODU

Curriculum Vitae

- *EDUCATION*
  - Ph.D. Computer Science, Old Dominion University, 2008
  - M.A. Computer Education, Hampton University, 1988
  - B.A. Natural Science, University of the State of New York, 1986
- *PROFESSIONAL EXPERIENCE*
  - 2004–Present Research Assistant, Old Dominion University
  - 2000–2004 Business Owner and Consultant
  - 1998–2000 Northrop Grumman, Inc.
  - 1989–1998 Inter-National Research Institute
  - 1987–1989 Electronic Institute of Technology
- *PUBLICATIONS & PRESENTATIONS*
  - http://www.joanasmith.com/pubs.html
  - http://www.joanasmith.com/ppt.html

Published Dissertation Research

1. **A Quantitative Evaluation of Dissemination-Time Preservation Metadata**. J.A. Smith and M.L. Nelson. *Proceedings of the 12th European Conference on Research and Advanced Technology for Digital Libraries*. (ECDL Sept. 2008).
2. **Creating Preservation-Ready Web Resources**. J.A. Smith and M.L. Nelson. *D-Lib Magazine*. January/February 2008.
3. **CRATE: A Simple Model for Self-Describing Web Resources**. J.A. Smith and M.L. Nelson. *Proceedings of the 7th International Web Archiving Workshop IWAW'07*. June 2007.
4. **Generating Best Effort Preservation Metadata for Web Resources at Time of Dissemination**. J.A. Smith and M.L. Nelson. *Proceedings of JCDL 2007*. June 2007.
5. **Efficient, Automatic Web Harvesting**. M.L. Nelson, J.A. Smith, I. Garcia del Campo, H. Van de Sompel and X. Liu. *Proceedings of ACM WIDM 2006.* November 2006.
6. **Site Design Impact on Robots: An Examination of Search Engine Crawler Behavior at Deep and Wide Websites**. J.A. Smith and M.L. Nelson. *D-Lib Magazine*. March/April 2008.
7. **Reconstructing Websites for the Lazy Webmaster**. F. McCown, J.A. Smith and M.L. Nelson. In *Dynamics of Search Engines: An Introduction.* Icfai University Press, 2007.
8. **Using The Web Infrastructure To Preserve Web Pages**. M.L. Nelson, F. McCown, J.A. Smith, and M. Klein. *International Journal on Digital Libraries.* July 2007.
9. **Lazy Preservation: Reconstructing Websites for the Lazy Webmaster**. F. McCown, J.A. Smith, M.L. Nelson, and J. Bollen. *Proceedings of ACM WIDM 2006*. November 2006.
10. **Reconstructing Websites for the Lazy Webmaster**. F. McCown, J.A. Smith, M.L. Nelson, and J. Bollen. *Technical Report*, Old Dominion University. December 2005.
11. **Observed Web Robot Behavior On Decaying Web Subsites**. J.A. Smith, F. McCown, and M.L. Nelson. *D-Lib Magazine*. February 2006.
12. **How Much Preservation Do I Get If I Do Absolutely Nothing**? M. Klein, F. McCown, J.A. Smith, and M.L. Nelson. In *Content Engineering: Konzepte, Technologien und Anwendungen in der Medienproduktion*. Gito-Verlag, Berlin, 2007.
13. **Repository Replication Using NNTP and SMTP**. J.A. Smith, M. Klein, and M.L. Nelson. *Proceedings of European Conference on Digital Libraries*. September 2006.

Research Questions

Can a web server actively support and contribute to web preservation?

- Can it address the counting problem?
  - Enumerate site resources efficiently and accurately
- Can it address the representation problem?
  - Package resources together with relevant metadata
- Is it feasible to use the web server?
  - Impact on performance
  - Long-term viability of response object

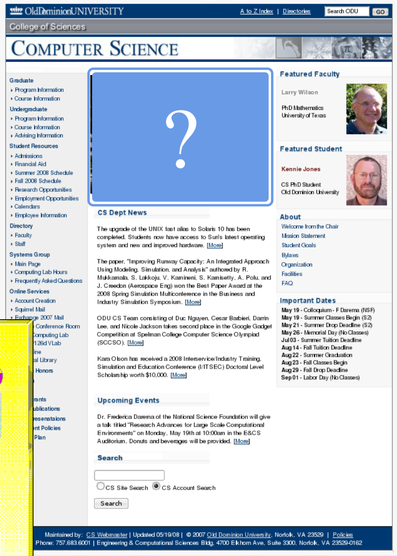
Outline (1)

① Background: The Challenge of Digital Preservation
② Research Focus: Website Preservation
③ The Counting Problem
④ The Representation Problem
⑤ The CRATE Reference Model
⑥ MODOAI
⑦ Future Work
⑧ Contributions
⑨ Questions & Comments

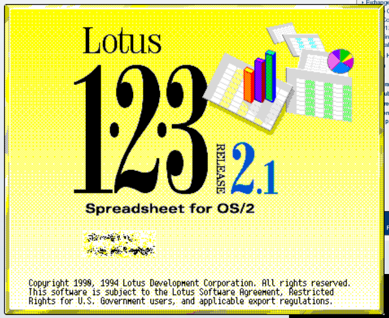
Digital Preservation Issues

- **Refreshing:**
  - If you don't have it, you can't preserve it
  - Resources disappear, change, degrade
  - A "Counting" Problem
- **Migration:**
  - If you don't upgrade it, it won't work
  - Formats & systems change, die, evolve
  - A "Representation" Problem
- **Emulation:**
  - If you can't imitate it, you can't understand it
  - Old systems are emulated in new environments
  - A "Representation" Problem

Refresh…



Migrate…



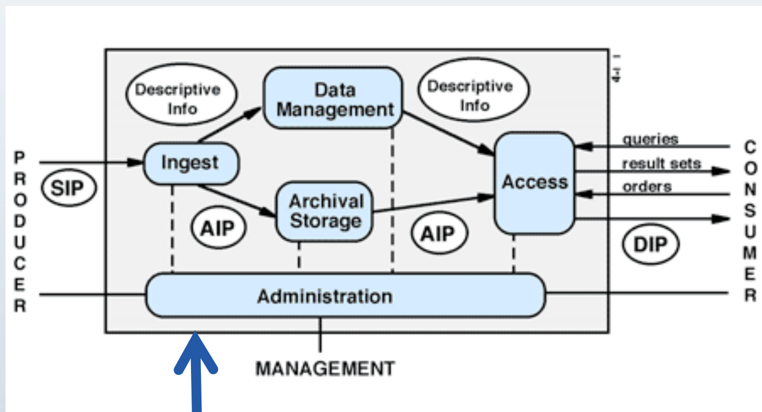
Emulate…

Model for an Open Archival Information System (OAIS)

- A General **Model of Preservation** (physical or digital)
  - **SIP** = Submission Information Package
  - **AIP** = Archival Information Package
  - **DIP** = Dissemination Information Package

from today through all tomorrows

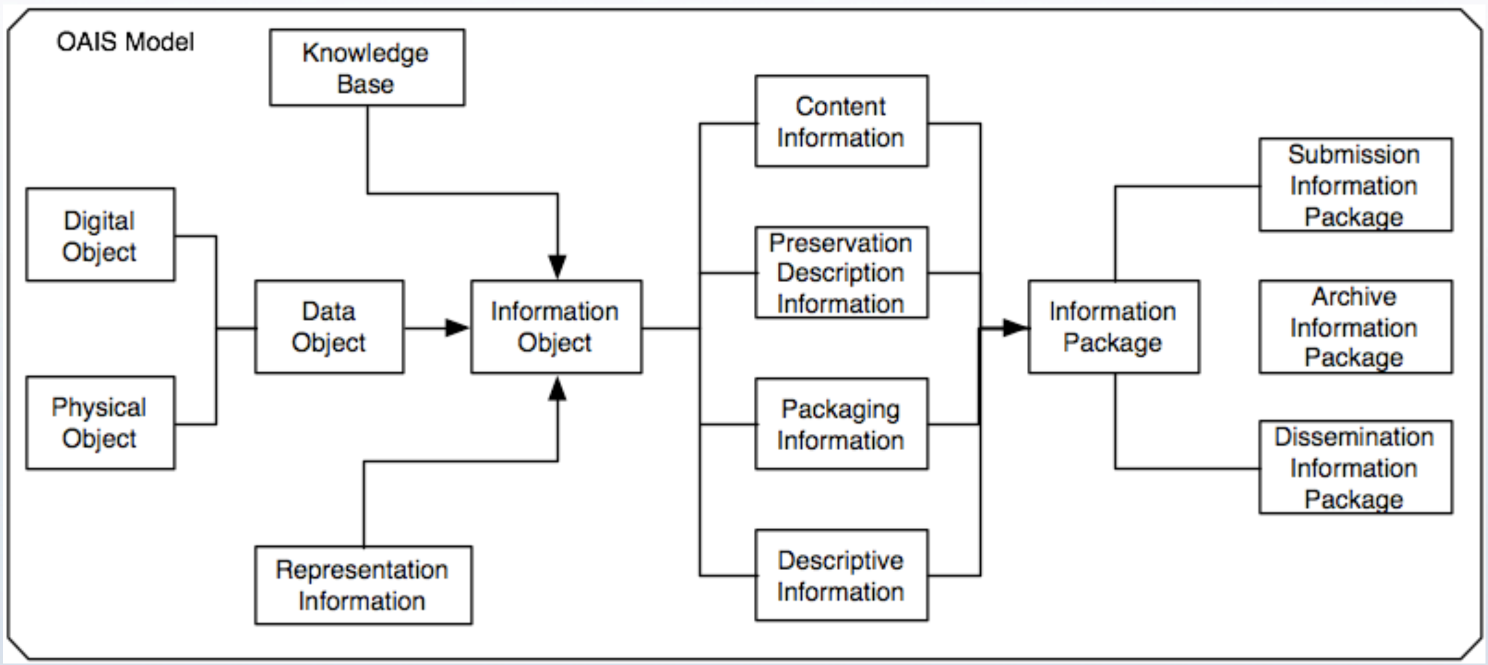
today



future

Note the complicated, active, on-going role of the archivist

OAIS Model

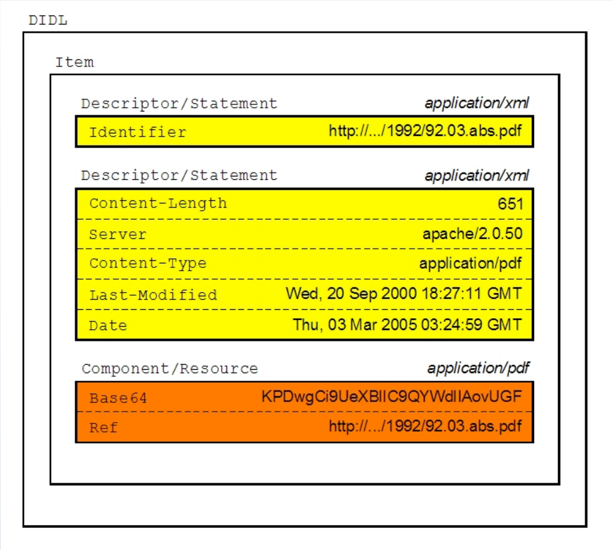


OAIS Model
Knowledge Base
Digital Object
Physical Object
Data Object
Information Object
Representation Information
Content Information
Preservation Description Information
Packaging Information
Descriptive Information
Information Package
Submission Information Package
Archive Information Package
Dissemination Information Package

Complex Objects

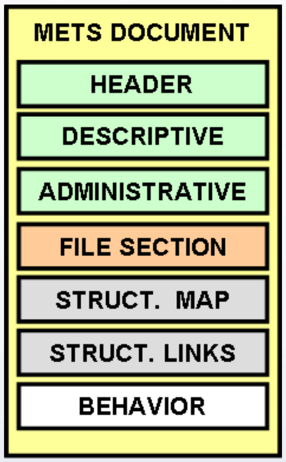
DVD
Audio
+
Video
+
Menus

- Representation of a digital object by means of a wrapper XML document
- Represented resource can be:
  - simple digital object (consisting of a single datastream): foo.txt
  - compound digital object (consisting of multiple datastreams) foo.asp
- Unambiguous approach to convey identifiers of the digital object and its constituent datastreams.
- Include datastream:
  - By-Value: embedding of Base64-encoded datastream
  - By-Reference: embedding network location of the datastream
  - not mutually exclusive; equivalent
- Include a variety of secondary information
  - By-Value
  - By-Reference
  - Descriptive metadata, rights information, technical metadata…

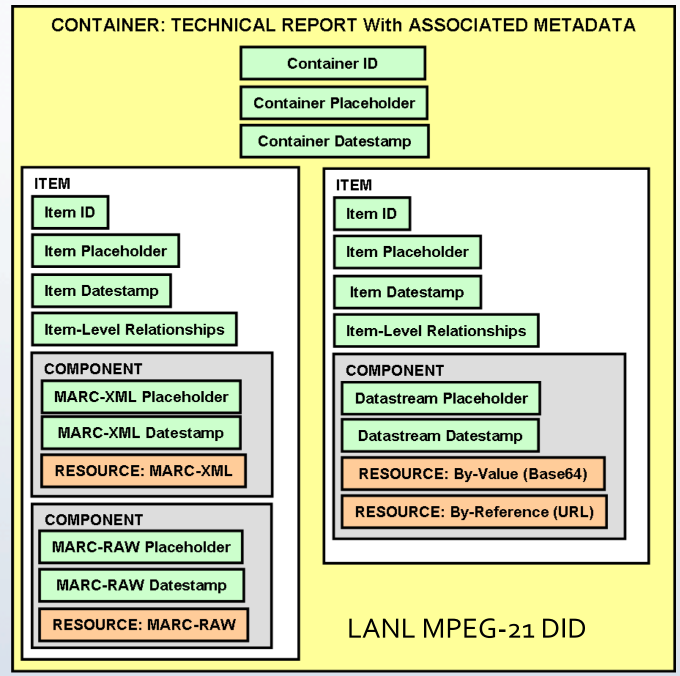


Many Digital Repository Models

Digital Libraries

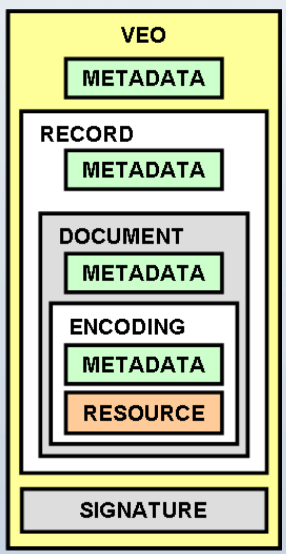


Metadata Encoding & Transmission Standard (USA)



Technical Reports

Official Government Records



Victoria Electronic Records Encapsulated Object (Australia)

Repositories vs Websites

Repositories

- Metadata-Rich
- Organized Content
- Known Resource List
- Refreshed
- Migrated
- Structured Changes
- Professional Backup Strategy
- Professional Recovery Strategy

Harvester Home Companion

Websites

- Minimal Metadata
- Link-dependent Content
- Uncertain Resource Count
- Uncertain Refreshing
- Uncertain Migration
- Unpredictable Changes
- Varying Backup Strategies
- Haphazard Recovery

Crawlapalooza

Outline (2)

① Background: The Challenge of Digital Preservation

② Research Focus: Website Preservation

③ The Counting Problem

④ The Representation Problem

⑤ The CRATE Reference Model

⑥ MODOAI

⑦ Future Work

⑧ Contributions

⑨ Questions & Comments

Website Preservation Approaches

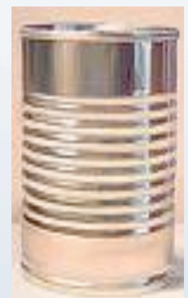
- Luck
  - Not actively trying to preserve
  - Hoping it’s there if you lose it

  → The Lazy Preservation approach (McCown)
- Cleverness
  - Internet Archive: Crawl and Save
  - LOCKSS: Share copies in controlled space

  → Work around what the web server delivers

  Neither enlists the web server as an agent of preservation

2 Problems in Website Preservation: Counting & Representation



The counting problem

How many pages are on that site?

To save it you have to get it

The representation problem

What's that page all about?

Use requires understanding

Preservation & The Counting Problem

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

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

- How big is W?
  - *It depends on whom you ask:*
    - File System
    - Web Server Configuration file
    - Website Links
    - Web Logs

→ HTTP doesn't know!

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:

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

Data Object

Representation Information

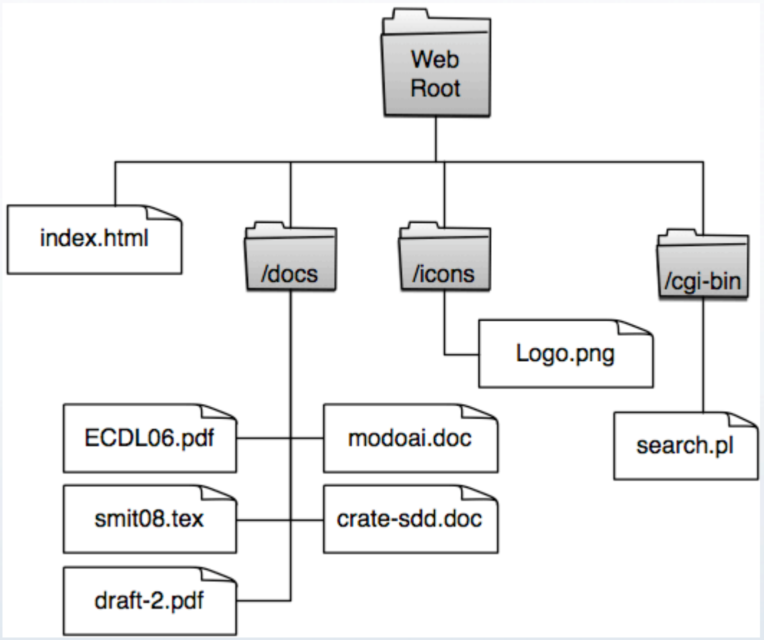
Knowledge Base

Information Object

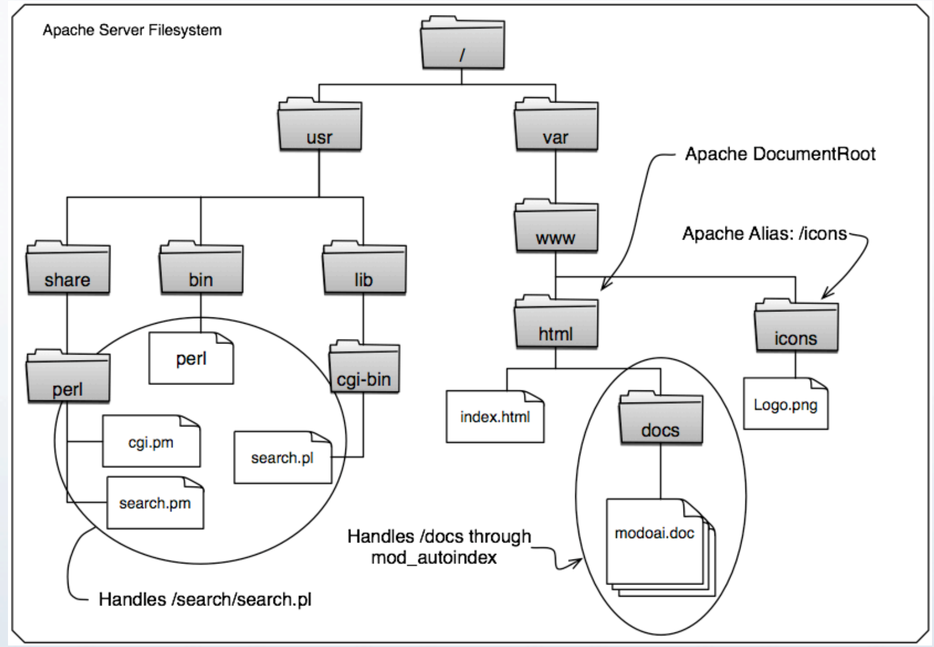
Outline (3)

① Background: The Challenge of Digital Preservation
② Research Focus: Website Preservation
③ The Counting Problem
④ The Representation Problem
⑤ The CRATE Reference Model
⑥ MODOAI: A Technical Implementation of CRATE
⑦ Future Work
⑧ Contributions
⑨ Questions & Comments

Why The Counting Problem Exists



Links <> Website



File System <> Website

http://foo.edu/draft-2.pdf

http://foo.edu/search.pl?name=Maly

A website is more than Links and Files

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

The Counting Problem & HTTP

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

- **HTTP cannot count W**
  - There is no "Select *" in HTTP
  - It is not a Query Language
  - Only GET one resource at a time
  - HTTP cannot give a crawler a list of resources it has
- **HTTP cannot update W**
  - Conditional GET by *datestamp* or *etag* is limited: affects $\mathbf{w_x}$ only
  - Cannot get a list of pages that have been deleted
  - *Each* resource must be requested, one at a time

HTTP alone is insufficient to confidently enumerate a site's resources

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

GET /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)
...

GET /images/jas2000.jpg HTTP/1.1
Host: www.joanasmith.com
If-Modified-Since Sat, 03 May 2008 19:43:31 GMT
```

The Counting Problem: How Do Crawlers Build **W**?

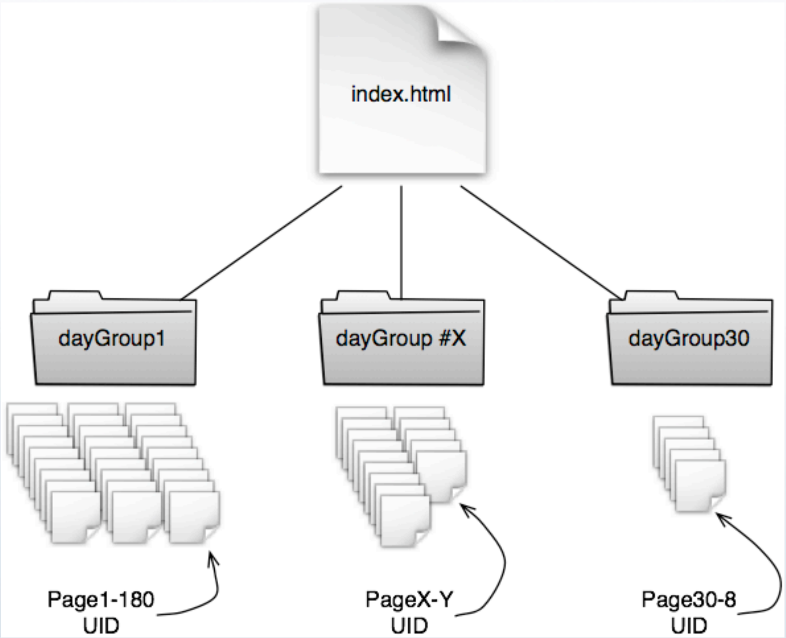
$$W = \{w_1, w_2, w_3, w_4 \ldots w_n\}$$

- Website preservation depends mainly on crawlers
- How do crawlers behave on a site?
- What impact does site design have on crawlers?
- Are there limits to crawls (depth, breadth)?
- How does resource removal affect crawlers?
- Series of *Crawler Observation* experiments
    1. Longitudinal investigations (5 to 13 months each)
    2. Decaying Website Experiments
    3. Buffet & Bread Crumb Experiments

The Counting Experiments: Decaying Websites

- Each website 30 directories wide
- 954 resources (HTML, PDF, images)
- *Unique* content on each site
- Daily removal of 1+ resources
- Eventually, only root is left
- Logs harvested for daily crawler activity (5 months)
- Patterns graphed
- Cache behavior monitored

(McCown – Lazy Preserveration)



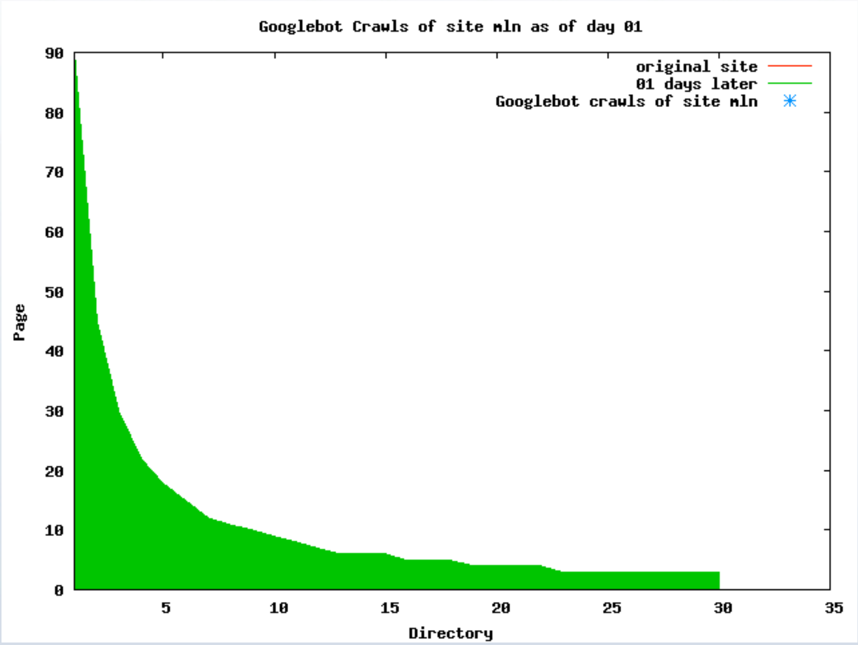
4 Unique Websites

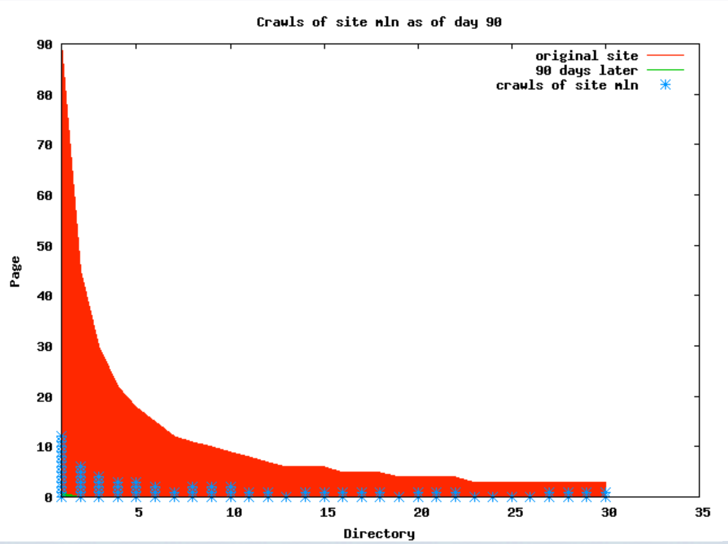
W = Links

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

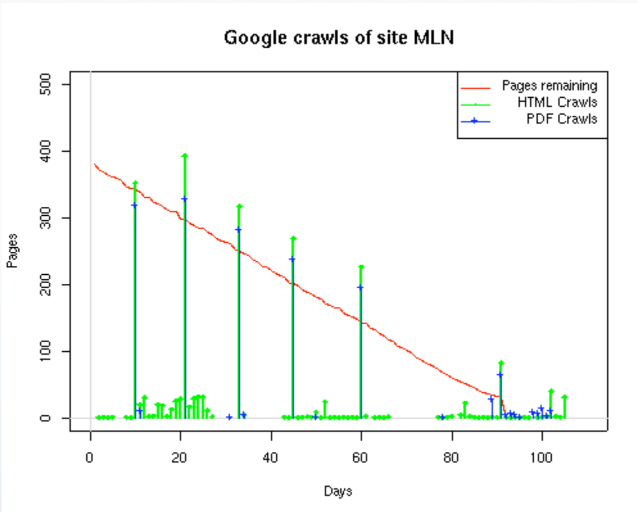
Results: How Crawlers Count Resources on Decaying Websites

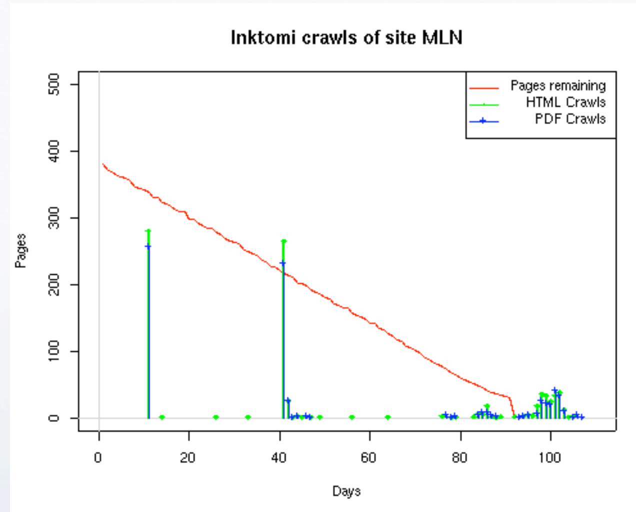
W => is decaying



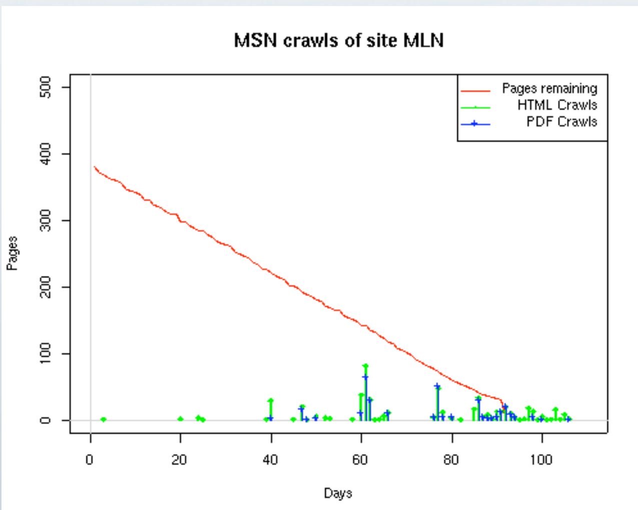


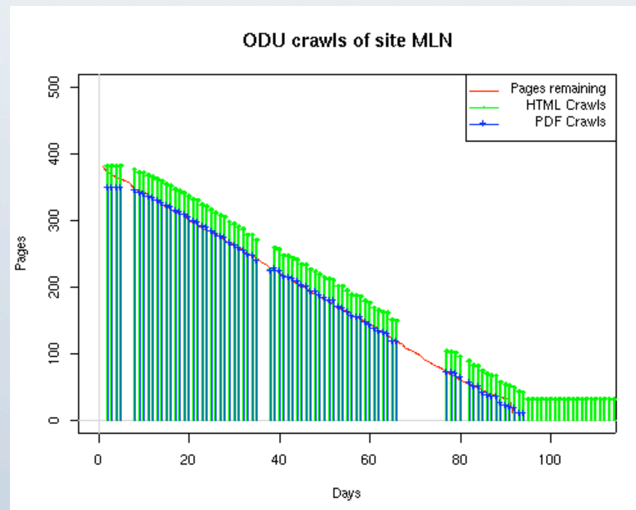
Counting Experiment #1 Summary: Decaying Websites





- Insight into crawler methods to count W
- Crawlers miss parts of W
- ➢ Note heavy load on server by ODU crawler →



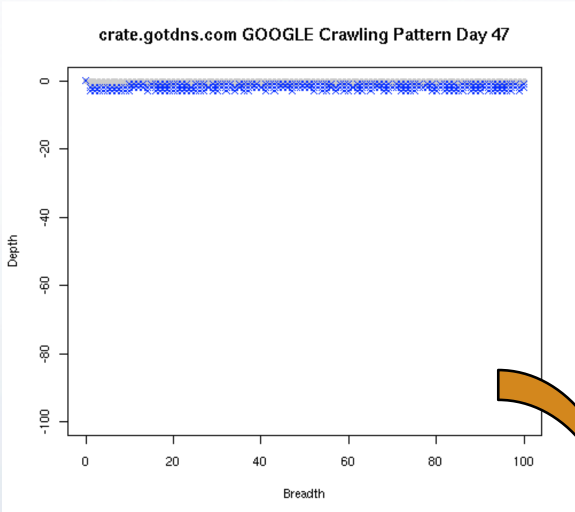


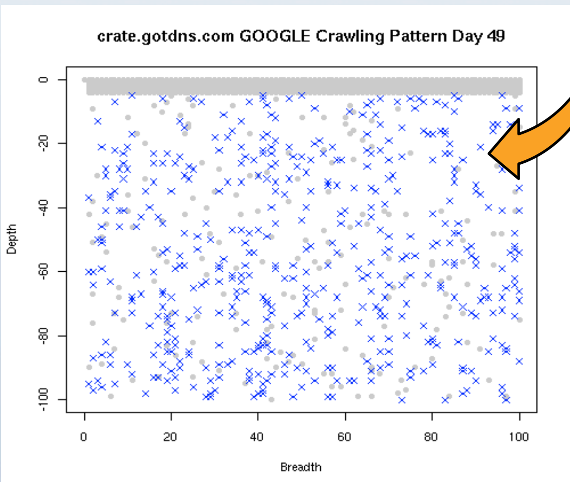
The Buffet & Bread Crumb Sites: Counting Experiment #2

- 4 very wide, very deep websites
- Over 20,000 resources on each site
- Unique, English-language content
- Two distinct website structures
  - Buffet example: http://crate.gotdns.com/
  - Bread Crumb example: http://oducrate.gotdns.com/
- Installed and monitored for 13 months
- Logs harvested for daily crawler activity & patterns
- *Static content* for entire experiment
- Two domains: dot-com and dot-edu

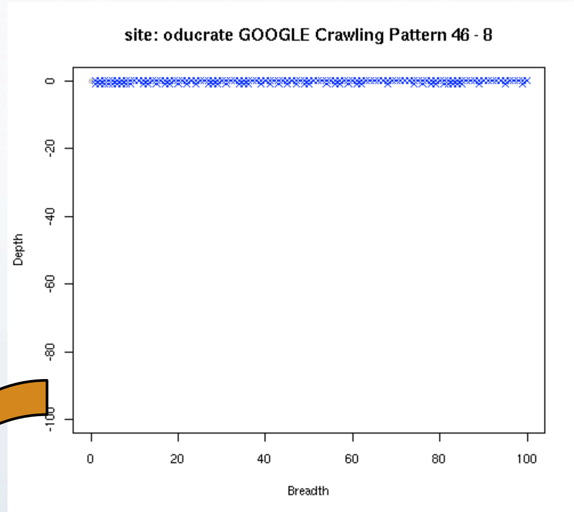
Counting Patterns: Buffet & Bread Crumb Sites

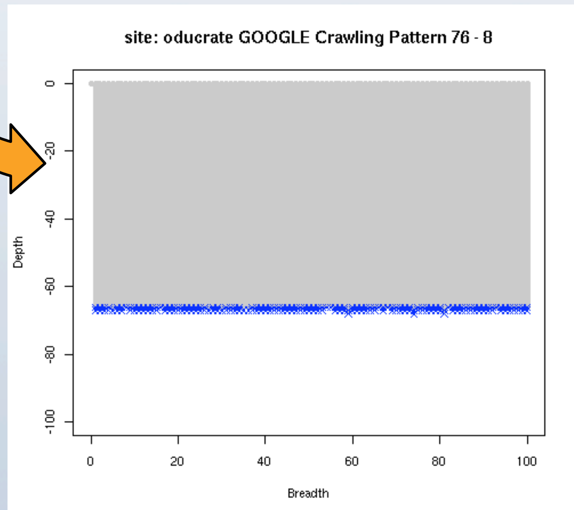
W = Buffet Links





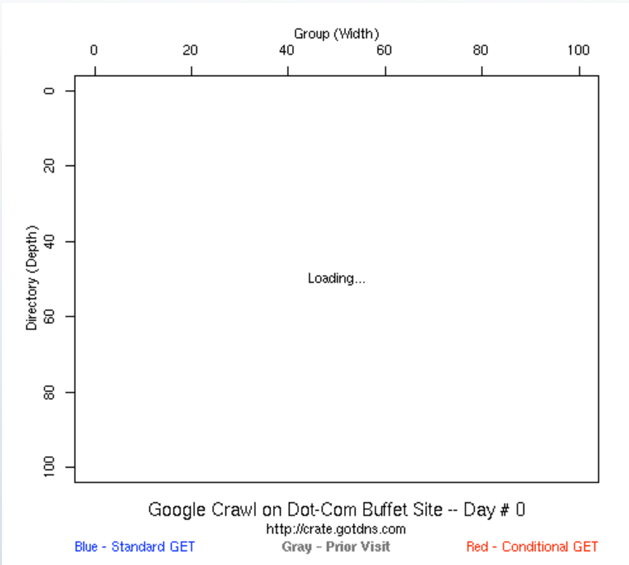
W = Bread Crumb Links





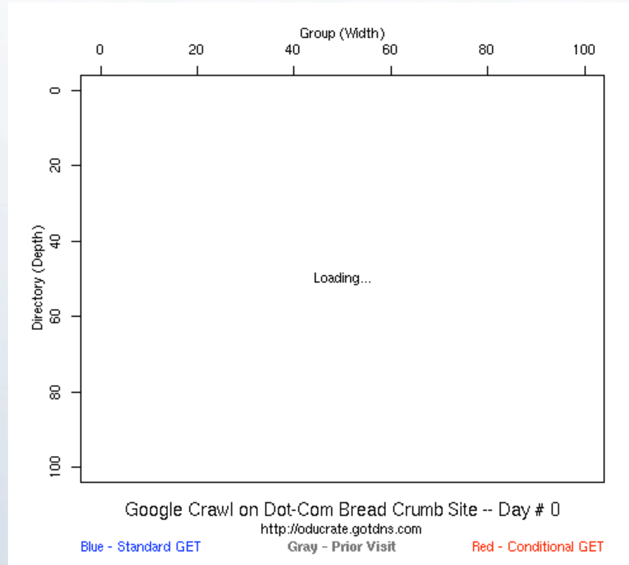
Counting Experiment #2: GoogleBot At Work

Buffet Site



W = Buffet Links

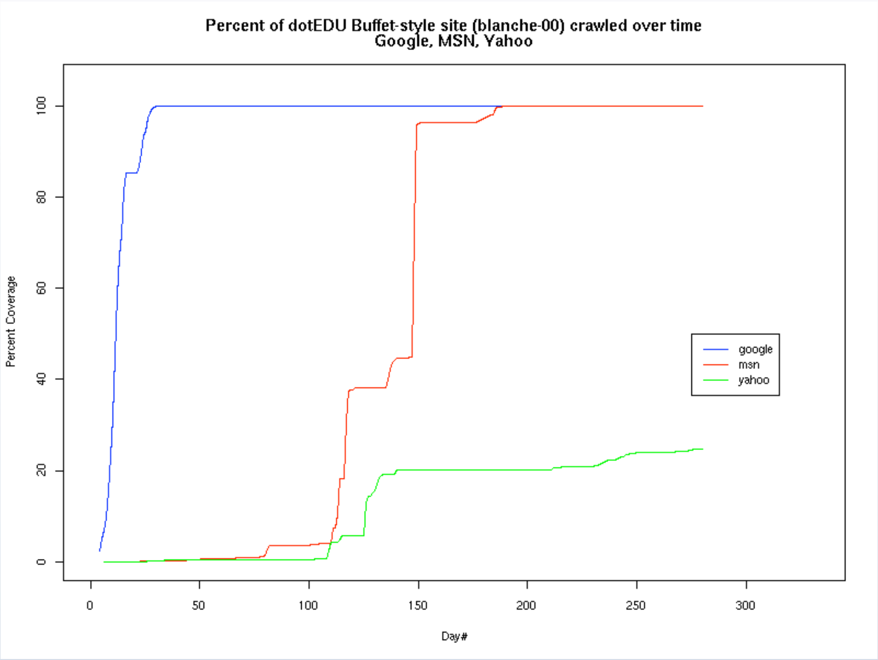
Bread Crumb Site

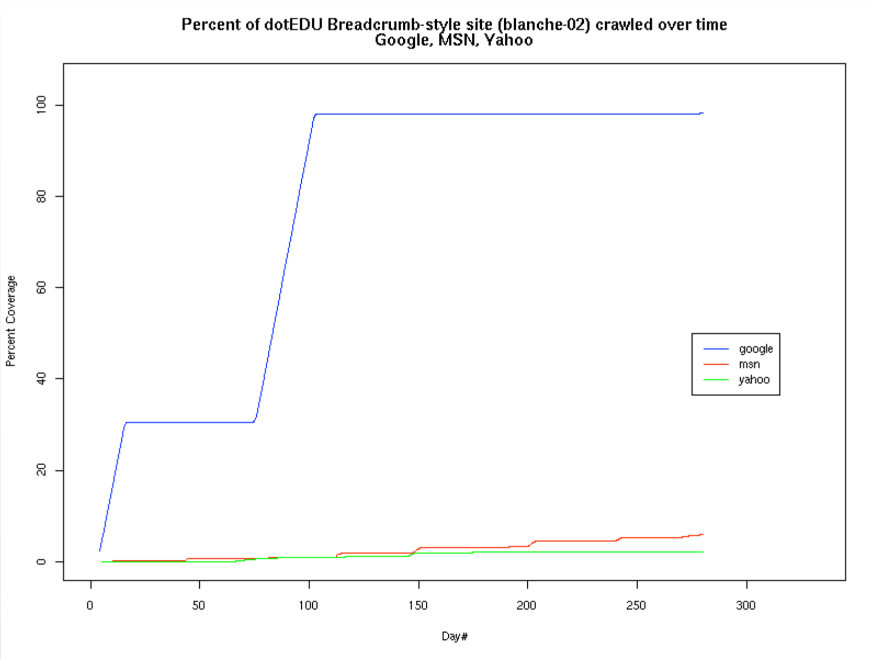


W = Bread Crumb Links

Counting Experiment #2 Summary: Buffet & Bread Crumb

View of W can depend on how the website is designed





W = Buffet Links

W = Bread Crumb Links

Counting Experiment #3: Solving The Counting Problem At Home

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

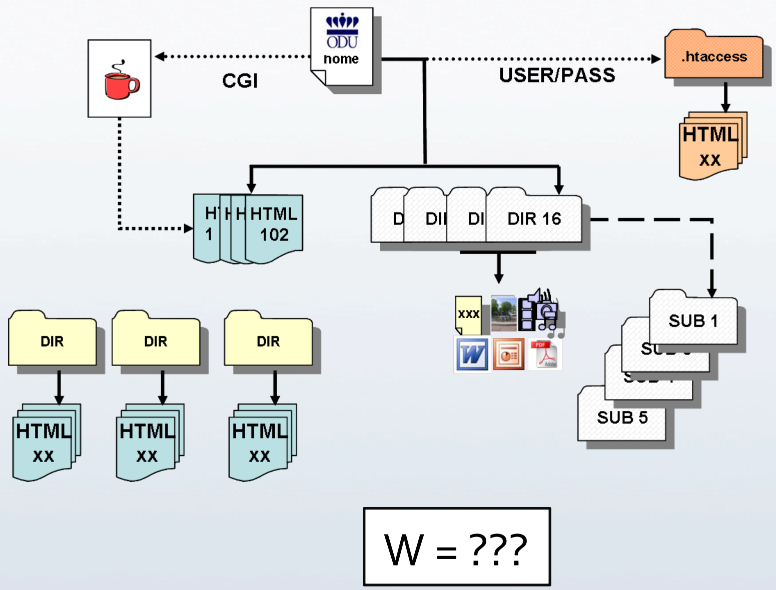
How should the Webmaster build W?

- Three main sources of website content listing:
  - Web Server File System
  - Website Links (Crawls)
  - Web Server Logs
- Each has limitations
- Each has unique insights
- *The "Count" depends on the strategy used*

W = ???

Counting Problem Experiment #3: A Real-World Example

- CS Dept Website Snapshot 6 June 2006
- Logs from 11/2005 – 01/2008
- No "tilde" sections
- CGI not in snapshot
- Miscellaneous resources missing
- Typical "Real World" website
- Real sites offer complexity not usually found on synthetic sites



Counting Experiment #3: Methods

- Search Engine Standard: Build A Sitemap.

  → But HOW?

1. Self-crawl (wget)
2. External Crawlers (Sitemap-Tool websites)
3. File System List
4. Log Harvesting
5. Local Sitemap script (access to local resources)

→ Experiment #3 Compares Results: $W = \{w_1, w_2, w_3, w_4 \ldots w_n\}$

Counting Experiment #3: Crawl Results

- Self-crawls and remote crawls produced similar results
  - Some are self-limiting: <= 300 URIs, e.g.
  - Some can convert fully-qualified internal links
  - External crawls returned "mailto" links
  - Google's script (combined self-crawl) choked on malformed log data
- Some Rewrite rules inferred from a comparison of links and file system (confirmed by Sys Adm)
- All required manual cleanup to merge information

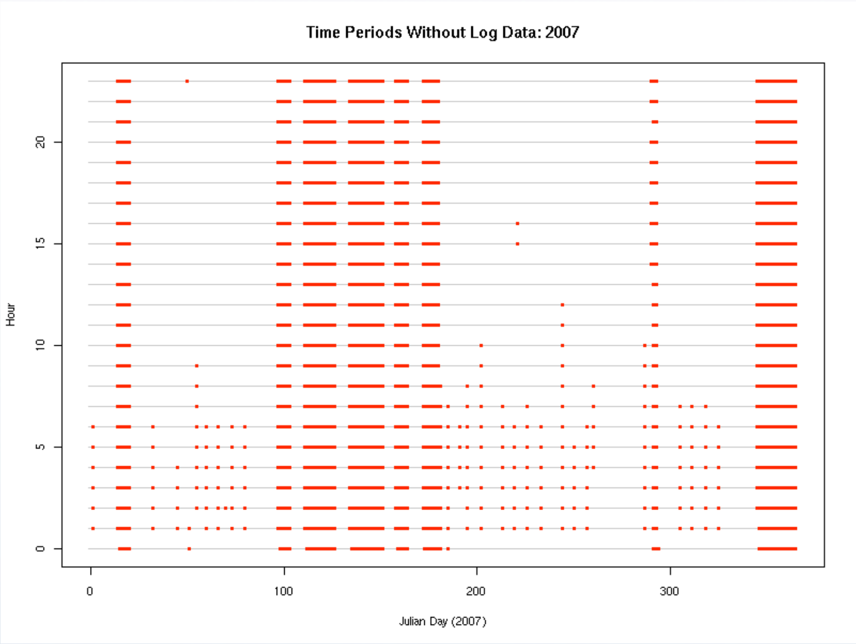
Source	Files	URLs
Self-Crawl	406	538
External Crawl	406	761
File System	*2,052*	*2,052**

*Canonical URLs

Counting Experiment #3: Log Data

26% Hours Covered

37% Hours Covered

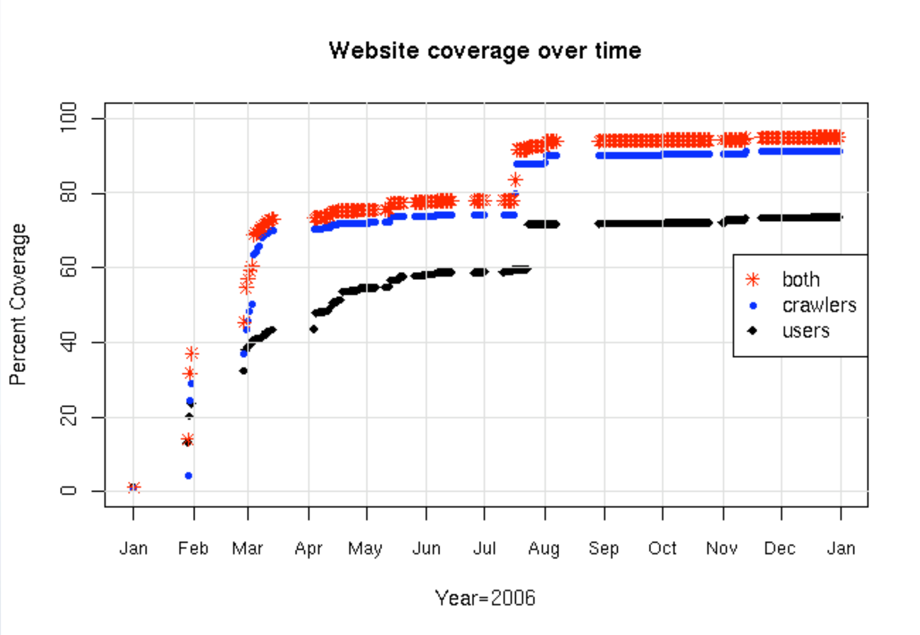


- Sparse but large: > 50M entries/year!
- Numerous Log-Entry Errors
- Extensive manual refinement of data

Sample Log Entry:

164.106.195.133 - - [01/Jun/2006:11:10:20 -0400] "GET /files/gfx-logo-odu-crown.gif HTTP/1.1" 304 -

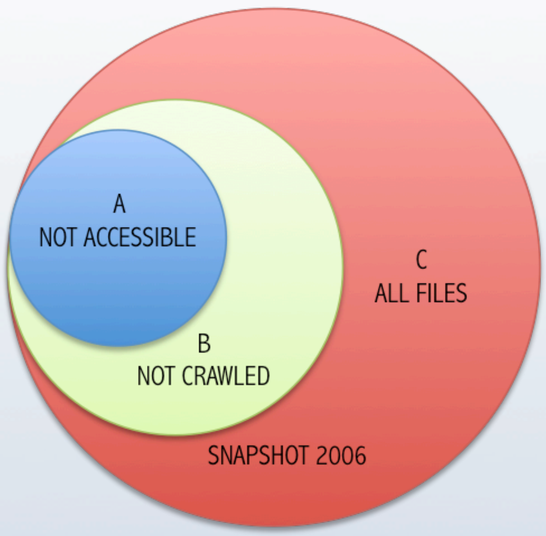
Counting Experiment #3: Log Analysis Results



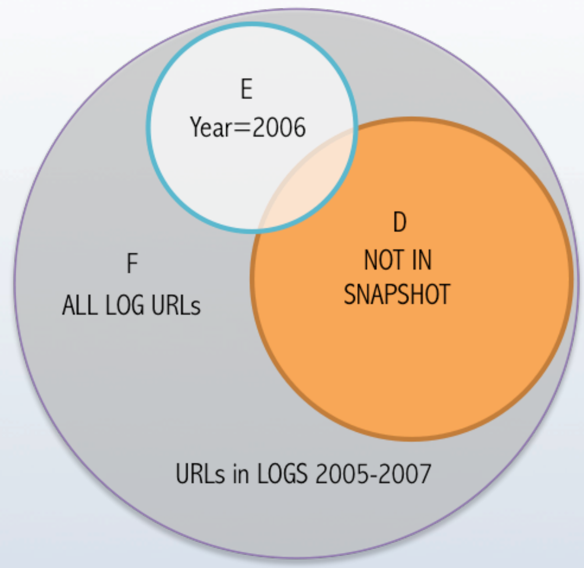
It takes a long time to cover the website!

- Coverage excludes protected areas
- Backup files also excluded, even if accessible
- Coverage = Relative to file system listing
- About 98% Coverage: $W = \{w_1, w_2, w_3, w_4 \ldots w_n\}$

Snapshot:
File System vs. Crawlers & Logs

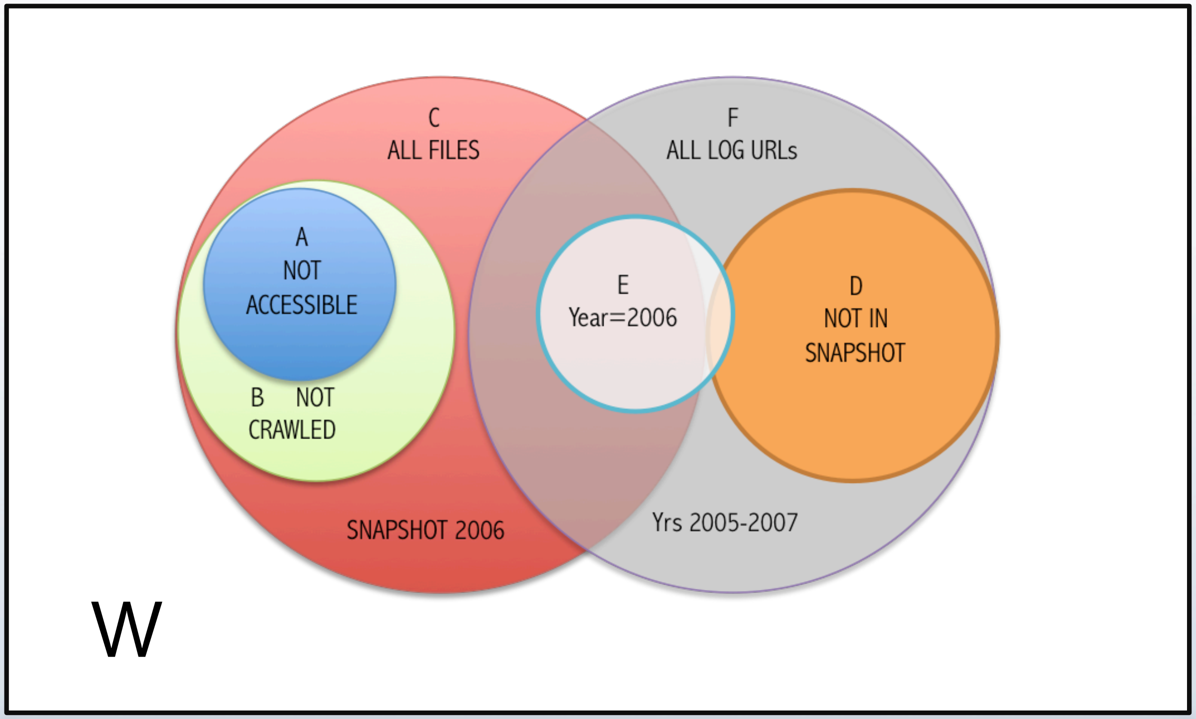


File System View of W



Web Logs View of W

Counting Experiment #3 Results: Combination of Methods is Best



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

Counting Experiment #3 Conclusion: Strategies for Optimizing Resource Count

- Log entries can give a reasonably complete picture
- Best results come from combination of methods
- Continuous update required as site evolves
- **Sitemap** produced from combination will give most complete picture to harvesting agent

http://www.cratemodel.org/sitemap.xml

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

BUT: Crawling is Still *Inefficient*

- Sitemaps only ***list*** resources, not send them
- Crawlers must still ***visit each*** one
- Update semantics are inefficient
- Selective harvesting not an option
- Sites change --> Sitemaps follow: *Race condition*

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

Outline (4)

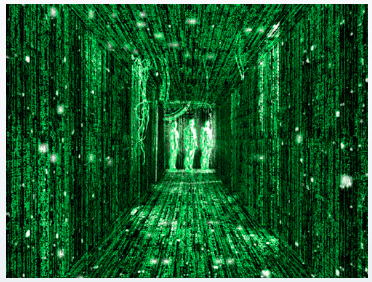
① Background: The Challenge of Digital Preservation
② Research Focus: Website Preservation
③ The Counting Problem
④ The Representation Problem
⑤ The CRATE Reference Model
⑥ MODOAI
⑦ Future Work
⑧ Contributions
⑨ Questions & Comments

Why The Representation Problem Exists

Data Object

Here are some bits

Representation Information



What do I know about it?

Knowledge Base

What is the Context?

Information Object

So this is what it should look like

Do I know enough to represent this correctly?

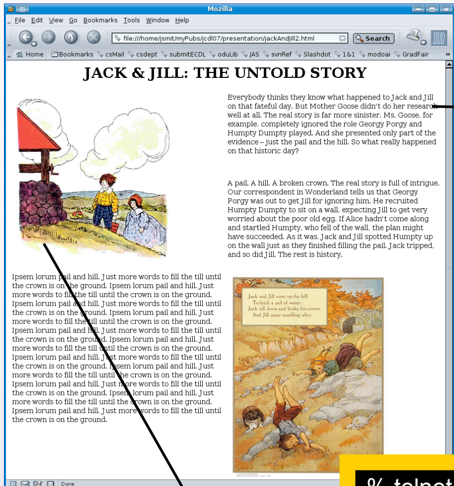
W alone is insufficient

Preservation Function P(W) is required

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

$P(\ w_1\ \ w_n\ \ w_2\ )\ \} \ P(W) = W$

Web Sites: Metadata Challenged



HTML metadata

```
<!DOCTYPE HTML PUBLIC "-//W3C//DTD HTML 4.0 Transitional//EN">
<HTML>
<HEAD>
        <META HTTP-EQUIV="CONTENT-TYPE" CONTENT="text/html; charset=utf-8">
        <TITLE></TITLE>
        <META NAME="GENERATOR" CONTENT="OpenOffice.org 2.0  (Linux)">
</HEAD>
<BODY LANG="en-US" DIR="LTR">
<H1 CLASS="western" ALIGN=CENTER>JACK &amp; JILL: THE UNTOLD STORY</H1>
<TABLE WIDTH=100% BORDER=0 CELLPADDING=4 CELLSPACING=0>
        <COL WIDTH=128*>
        <COL WIDTH=128*>
        <TR VALIGN=TOP>
                <TD WIDTH=50%>
                        <P><IMG SRC="jackAndJill_html_m28b86d66.jpg" NAME="graphics1" ALIGN=LEFT
WIDTH=325 HEIGHT=354 BORDER=0><BR>
                        </P>
                </TD>
                <TD WIDTH=50%>
                        <P>     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. Ms. Goose, for
                        example, completely ignored the role Georgy Porgy and Humpty
                        Dumpty played. And she presented only part of the evidence –
                        just the pail and the hill. So what really happened on that
                        historic day?</P>
                        <P><BR>
                        </P>
                        <P>A pail. A hill. A broken crown. The real story is full of
                        intrigue. Our correspondent in Wonderland tells us that Georgy
                        Porgy 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 of 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 pail. Jack tripped, and so did Jill. The rest is
                        history.</P>
                </TD>
        </TR>
        <TR VALIGN=TOP>
                <TD WIDTH=50%>
                        <P>Ipsem lorum pail and hill. Just more words to fill the till
                        until the crown is on the ground.  Ipsem lorum pail and hill. Just
                        more words to fill the till until the crown is on the ground.
                        Ipsem lorum pail and hill. Just more words to fill the till until
                        the crown is on the ground.  Ipsem lorum pail and hill. Just more
                        words to fill the till until the crown is on the ground.  Ipsem
                        lorum pail and hill. Just more words to fill the till until the
                        crown is on the ground.  Ipsem lorum pail and hill. Just more
                        words to fill the till until the crown is on the ground.  Ipsem
                        lorum pail and hill. Just more words to fill the till until the
                        crown is on the ground.  Ipsem lorum pail and hill. Just more
                        words to fill the till until the crown is on the ground.  Ipsem
                        lorum pail and hill. Just more words to fill the till until the
                        crown is on the ground.  Ipsem lorum pail and hill. Just more
                        words to fill the till until the crown is on the ground.  Ipsem
                        lorum pail and hill. Just more words to fill the till until the
                        crown is on the ground.
                        </P>
                </TD>
```

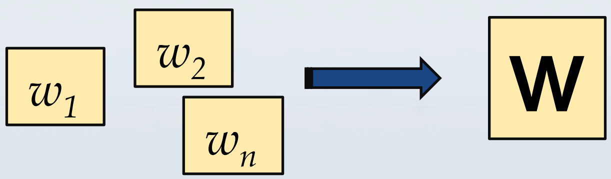
HTTP metadata

```
% 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"¢ÃÒÿÄ
```

Solving the Representation Problem

- Must be automatic
- Utilities need batch or command-line mode
- Should examine resource beyond simple MIME information
- Complex object response

Create *self-describing* resources



- Package resource + metadata together
- ASCII, Base64 encoding (avoid non-standard character sets)

What is a "Self-Describing" Resource?

Standard HTTP Headers --

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

***PLUS**: Output from built-in utilities:*

EXIF TOOL:

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 TOOL:
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

w_i

Wrapped *together with the resource* in simple XML

Example Metadata Utilities

Name	Description
Jhove	Image analysis
Kea	Key-phrase extraction
OTS	Open Text Summarizer
ExifTool	Image/video metadata extractor
Pdflib	Extract PDF metadata
MP3-Tag	Extract audio file tags
Essence	Customized information extraction
Droid	MIME++

Representation Experiment #1

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

- Synthetic website from earlier web experiments
  - Combination HTML, PDF, Plain Text, Images Files
  - Less than 100 resources
- Proof-of-Concept
  - Simple utilities installed on web server
  - Per-resource configuration:
    - Jhove-PDF HUL for PDF files
    - Jhove-JPEG HUL for JPEG files
    - Exif for JPEG files
    - Open Text Summarizer for Text files
- Installed on web server
  - Used earlier version of MODOAI
  - Harvested using OAI-PMH ListRecords

Proof-Of-Concept Evaluation

*The web server **can** provide Representation Information*

P(W) = W

Utilize Web Server:

- Integrated, easy-to-implement option for improving web resource metadata
- Per-resource configurable
- Compatible with Java, C, Perl, utilities – any command-line version metadata tool
- Utility speed reflects non-web-server speed
- Produces XML/Complex-Object Response

Get Best-Effort Metadata:

- *Unverified*
  - Utility results are not cross-checked
  - Conflicting information left as is
- *Undifferentiated*
  - No categorization of output
  - No specific ordering of metadata
- *Extemporaneous*
  - Generated at time of dissemination
  - Processed by each appropriate utility upon request

Outline (5)

① Background: The Challenge of Digital Preservation
② Research Focus: Website Preservation
③ The Counting Problem
④ The Representation Problem
⑤ The CRATE Reference Model
⑥ MODOAI
⑦ Future Work
⑧ Contributions
⑨ Questions & Comments

The CRATE Reference Model

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

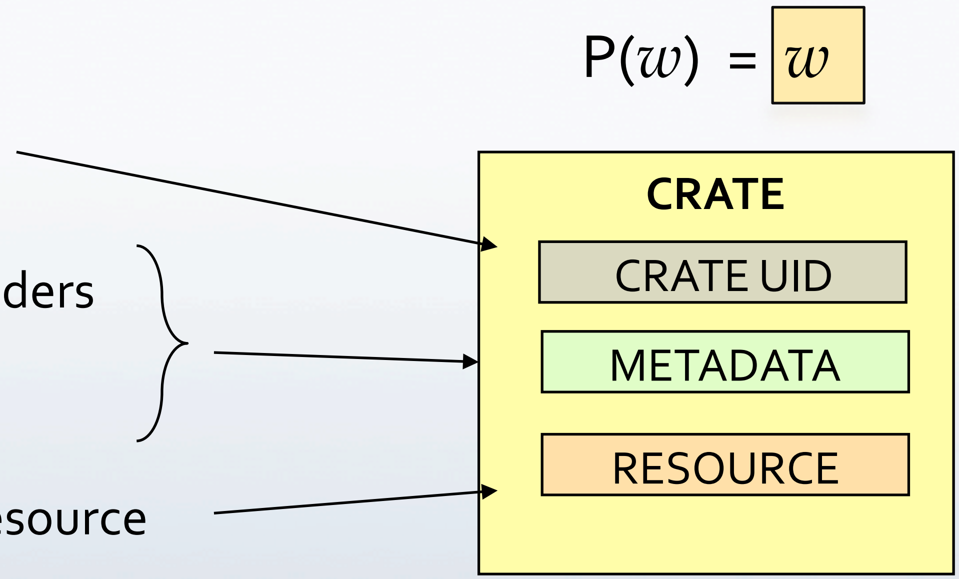
CRATE Addresses the Representation Problem

- How to represent
  - Use metadata utilities
  - Applied per-resource at time of dissemination
  - Undifferentiated, unverified, extemporaneous metadata
- How to package for archive submission
  - Complex-object response
  - XML, Simple encoding

UID + RESOURCE + METADATA = Preservation-Ready Resource

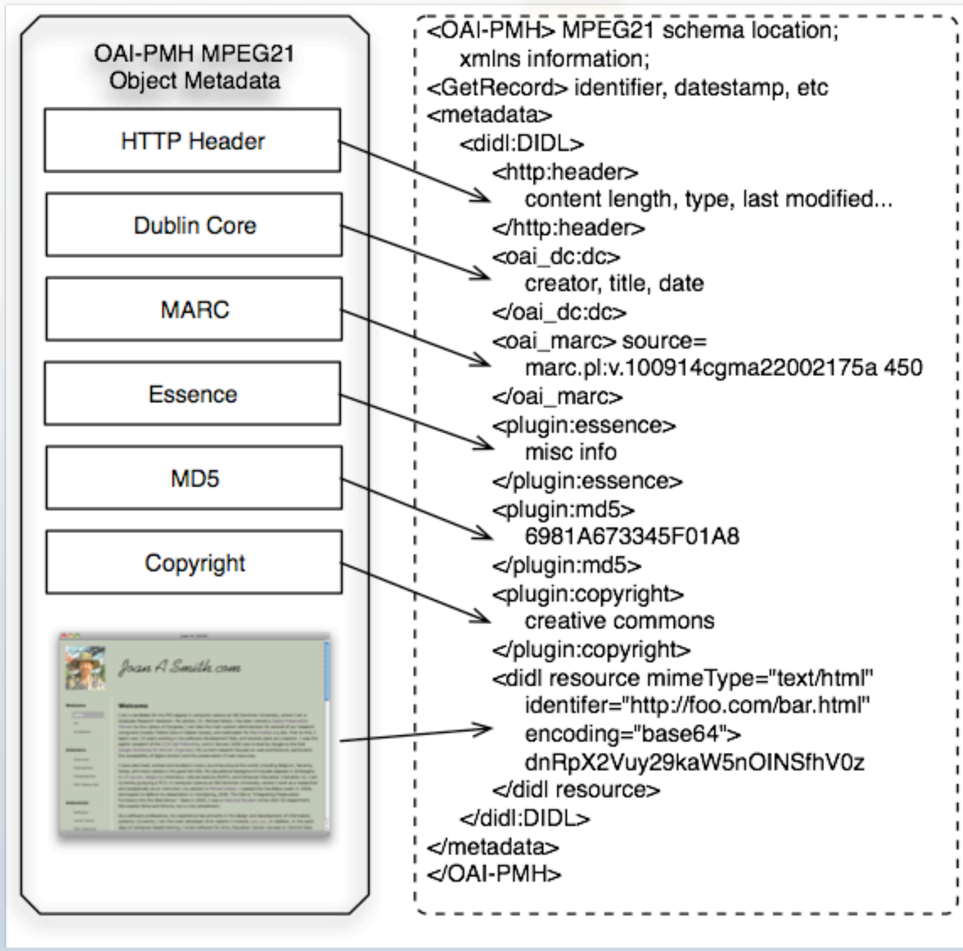
CRATE Elements

- URI, UUID
- Standard HTTP Headers
- Plug-In Metadata
- Base64-Encoded Resource

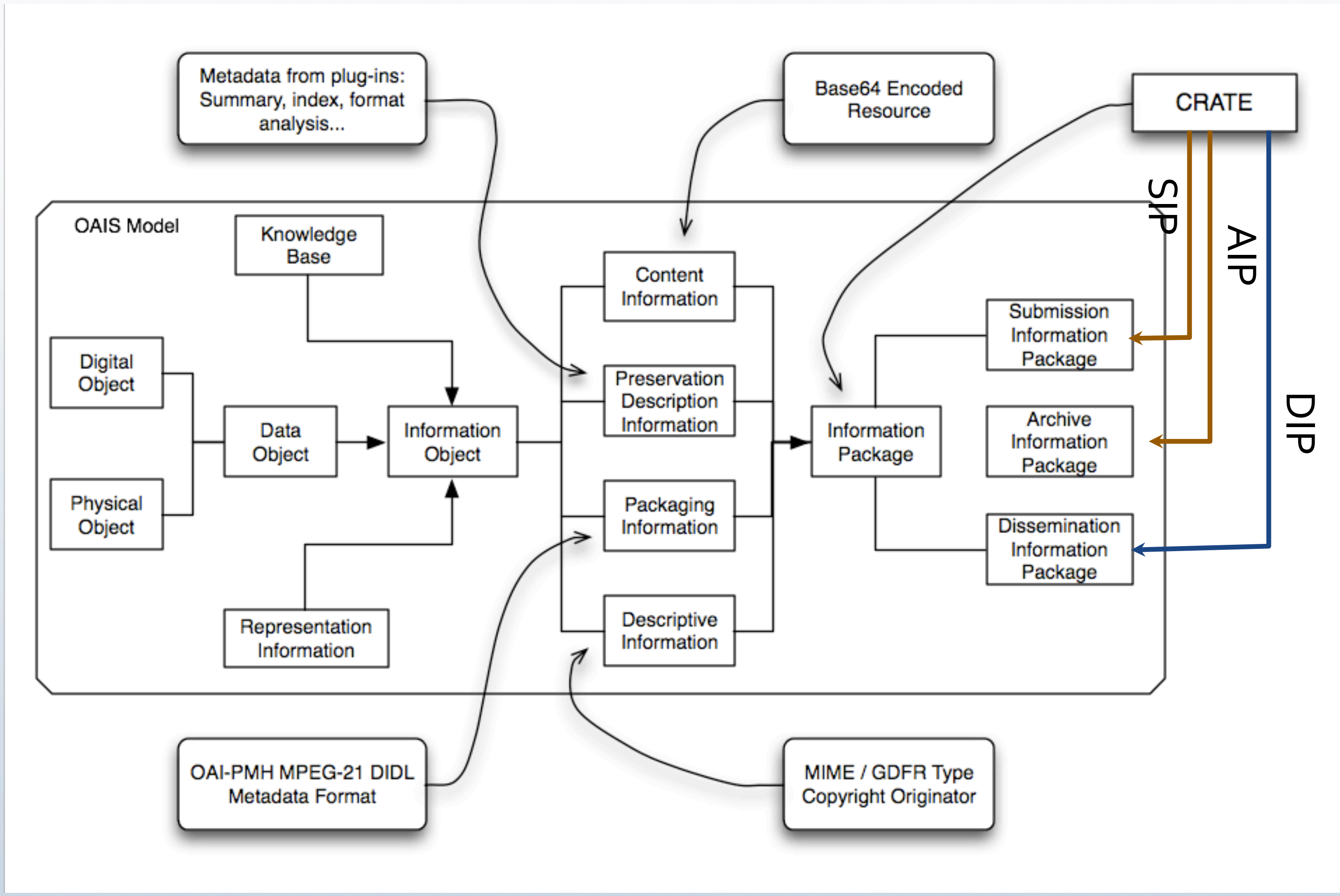


CRATE Example in MPEG-21 DID Format

$P(w) = w$



CRATE in the OAIS Model



Outline (6)

Pulling It All Together with MODOAI: Counting, Representation & Preservation

1. Addresses Counting Problem
   - Sitemap file provides resource listing
   - Implements OAI-PMH Protocol

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

2. Addresses Representation Problem
   - Metadata utilities specified in web server configuration file
   - "Regex" style assignment of utility to resource types
3. Creates Preservation-Ready Resources (DIP)
   - Complex-object response in MPEG-21 DIDL format
   - Whole website or single resource

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

MODOAI: INTEGRATING PRESERVATION & MORE

- Completely rewritten and extended from original prototype
- Apache 2.2 web server module
- Linux (Red Hat, Fedora, Debian) & Mac OS X
- Plugin architecture
- Not just for preservation…

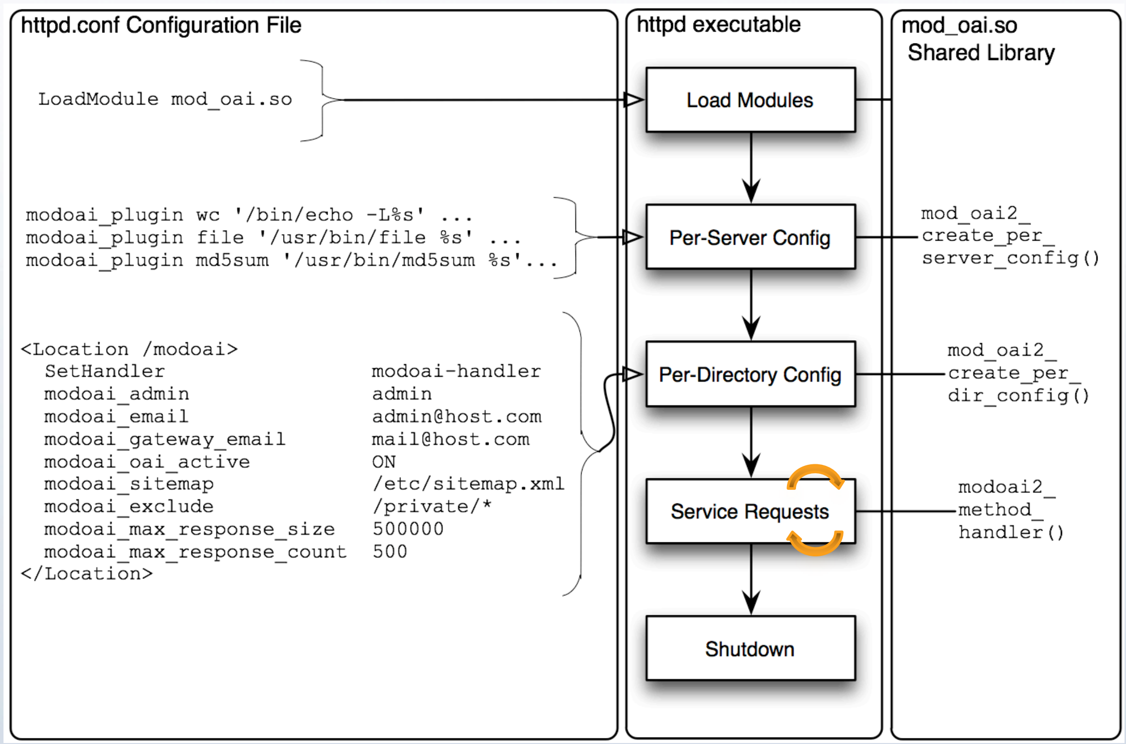
OAI-PMH: Efficient, Automatic Harvesting

Better than just a Sitemap!

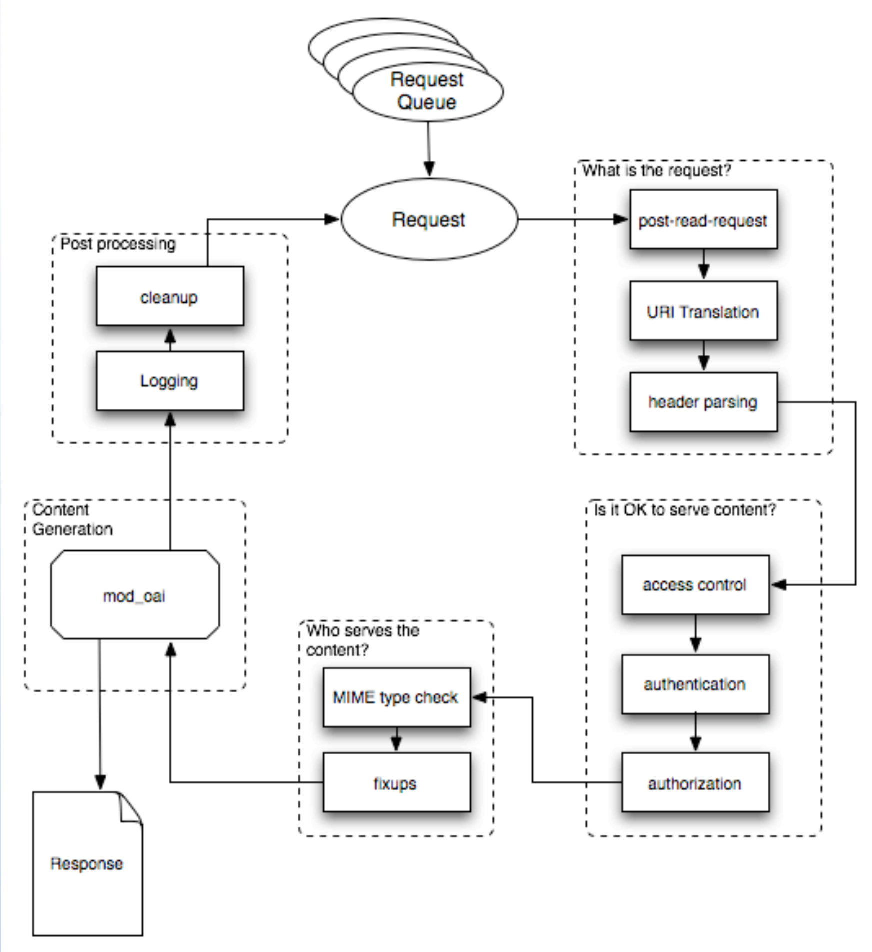
- 6 Verbs of OAI-PMH
  1. Identify
  2. ListIdentifiers
  3. ListRecords
  4. ListSets
  5. GetRecord
  6. ListMetadataFormats
- Efficient Update Semantics
  1. By Date Range
  2. By Set (MIME Types)
- Allows Complex Object Response
  1. Metadata + Resource
  2. CRATE & Other Types

- http://www.cratemodel.org/modoai?verb=Identify
- http://www.cratemodel.org/modoai?verb=ListIdentifiers
- http://www.cratemodel.org/modoai?verb=ListRecords&metadataPrefix=oai_didl
- http://www.cratemodel.org/modoai?verb=ListIdentifiers&from=2000-01-01&until=2005-12-12
- http://www.cratemodel.org/modoai?verb=ListIdentifiers&set=mime:text:html
- http://www.cratemodel.org/modoai?verb=ListSets

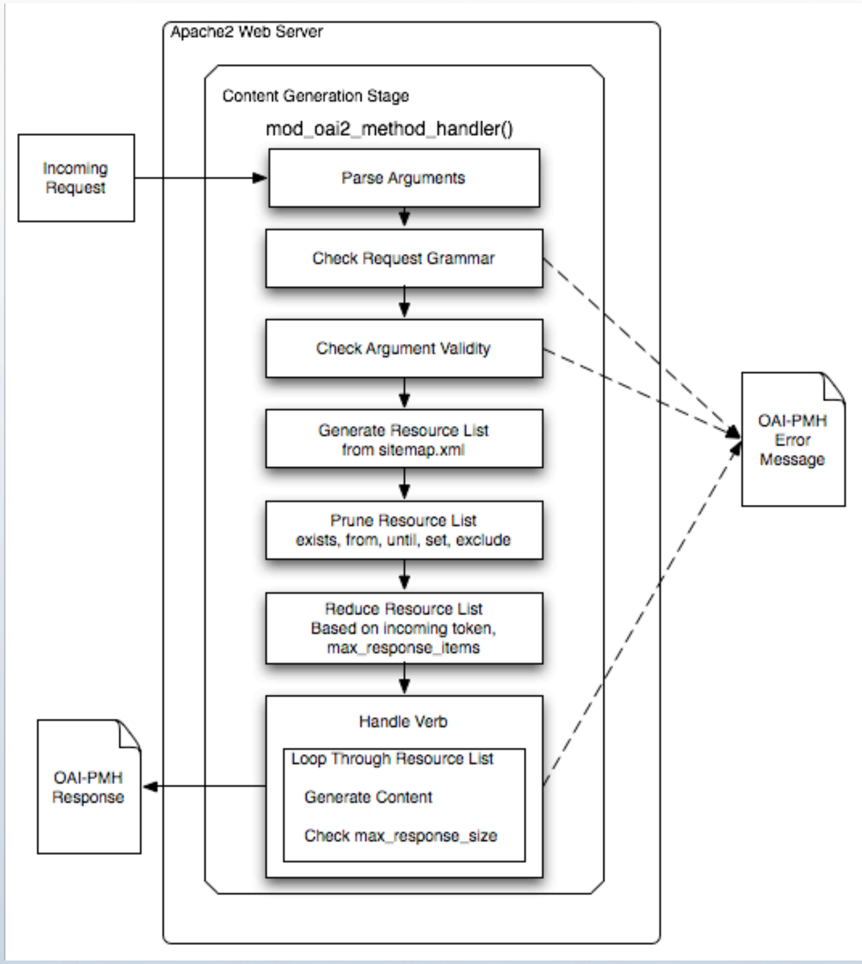
MODOAI in Apache



Apache Request Processing



MODOAI Request Processing



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

Quantitative Evaluation

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

Request Parameters	Active Utilities	Response Time in Min:Sec By Server Load 0 %	50 %	100%	Response Size (Bytes)
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

Evaluation of CRATE Using MODOAI

- No significant impact to web server: *fast* response
- Most utilities well-behaved
- One utility problematic
- Performance is additive
- Size is additive

- CRATE is feasible for web server use
- Web servers are I/O bound, not CPU bound (usually)
- Utilities occupy otherwise-unused CPU cycles
- Compiled utilities are much faster
- Individual test run on any new utility should be performed before inclusion as part of a CRATE solution

Outline (7)

① Background: The Challenge of Digital Preservation
② Research Focus: Website Preservation
③ The Counting Problem
④ The Representation Problem
⑤ The CRATE Reference Model
⑥ MODOAI
⑦ Future Work
⑧ Contributions
⑨ Questions & Comments

Future Work

- HTTP content negotiation
  - Accept-Encoding, q values: CRATE-type content
- Atom or other RSS-type feed
  - Other non-OAI-PMH technology for CRATE-type output
- MODOAI enhancements
  - Basic metadata utility set
  - Improve documentation
  - Build user community

Outline (8)

① Background: The Challenge of Digital Preservation
② Research Focus: Website Preservation
③ The Counting Problem
④ The Representation Problem
⑤ The CRATE Reference Model
⑥ MODOAI
⑦ Future Work
⑧ Contributions
⑨ Questions & Comments

Contributions

- Solve the Counting Problem with Sitemaps

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

- Solve the Representation Problem with CRATE

$$P(W) = W$$

- Solve Crawling Inefficiencies with OAI-PMH

Give me all JPEGs newer than 2008-06-06

➢ *Solve All 3 with MODOAI*

Contributions

- Cardinality of W best determined by combining listing sources
- Novel approach to website preservation: use the web server!
  - Metadata utilities installed on web server
  - Resource + metadata packaged together at dissemination time
  - Data-centric not Tool-centric
- CRATE
  - Ontologically agnostic
  - Simplest model possible
  - UID + Undifferentiated Metadata + Base64 Encoded Resource
  - Compatible with other archive models
- MODOAI
  - New, ROBUST extensible, plugin architecture
  - Can implement public and private Sitemaps
  - GPL-2 release at http://code.google.com/p/modoai/
  - Demo at http://cratemodel.org/

Next Stop: Emory University

Outline (9)

1. Background: The Challenge of Digital Preservation
2. Research Focus: Website Preservation
3. The Counting Problem
4. The Representation Problem
5. The CRATE Reference Model
6. MODOAI
7. Future Work
8. Contributions
9. Questions & Comments

Backup Slides

Resumption Tokens in MODOAI

- Server-throttled response rate
- Variable by number, size, both
- Determined in configuration file (modoai.conf)

modoai_max_response_size

- ➢ Number of Bytes that trigger a Resumption Token
- ➢ Initial response-part completes a Record (no partial-record)

modoai_max_response_count

- ➢ Number of Records that trigger a Resumption Token
- ➢ By record-counter in Sitemap file, after pruning (sets, dates, previous Resumption Tokens)

Summary

CRATE:

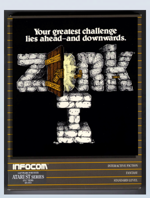
- "Pretty Good" Preservation
- Enlists web server as preservation agent
- Feasible & Practical for webmasters to use
- Significantly improves likelihood of long-term preservation
- Self-describing resources
- Addresses *Counting* & *Representation* Problems

MODOAI:

- Addresses Counting Problem through Sitemap
- Simple to install & configure
- Facilitates crawling
- Improves update semantics over standard HTTP
- More than just a preservation tool

Fragility of Digital Data

Durable



Fragile

"Digital information lasts forever --
or 5 years, whichever comes first"

-- Jeff Rothenberg

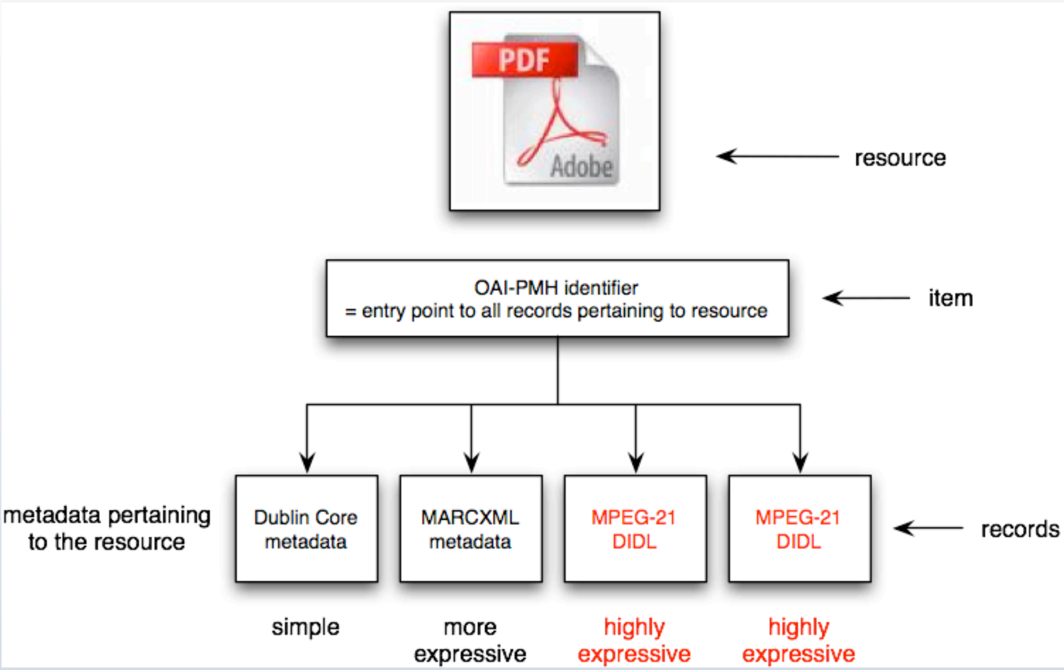
- Do you still have a copy of your first email?
- Can you still compile and run the first program you ever wrote? BASIC compilers are hard to find these days…
- If lightning fried your computer, how much information would you have lost?
- How many versions of your website have you made? How many do you still have?

Digital information is very fragile

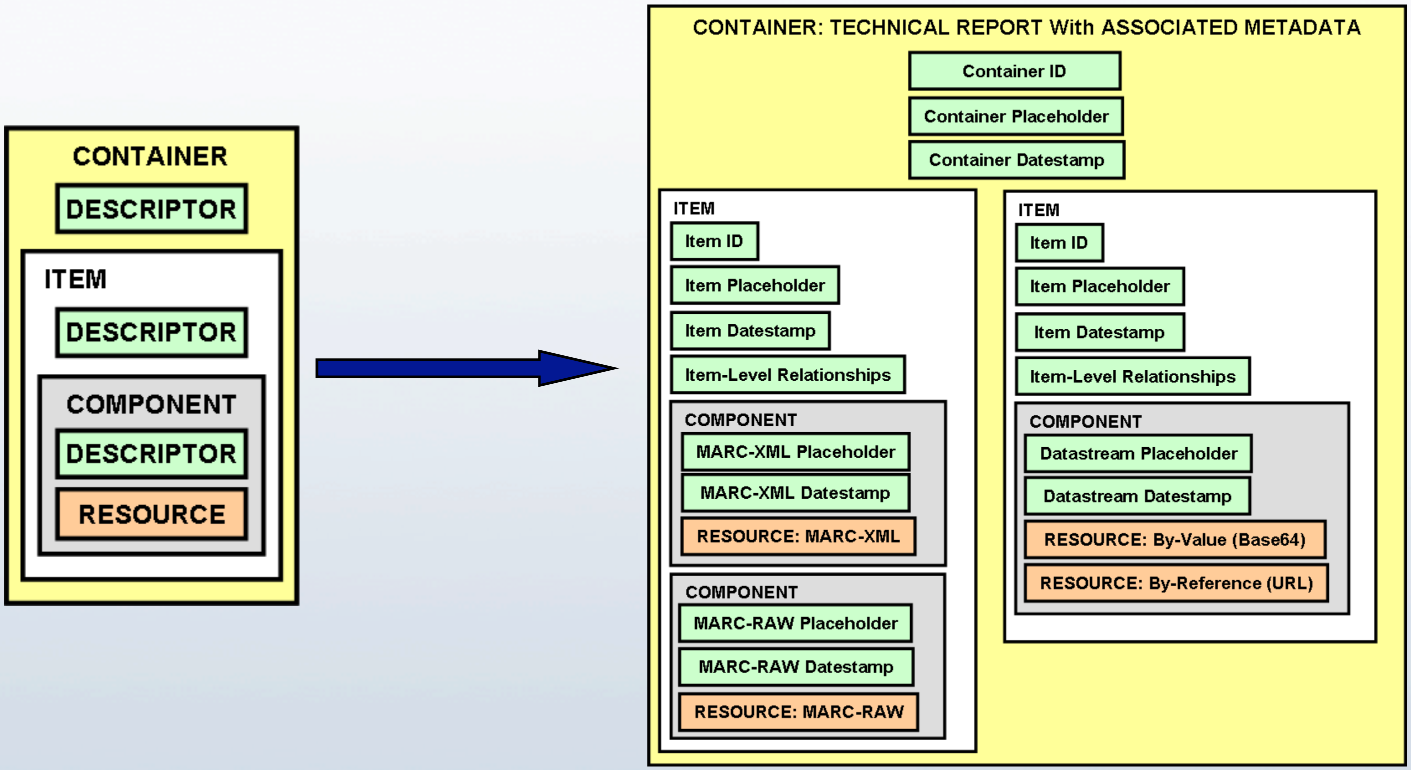
Formal Models & Implementations

- The Ultimate Standard: The OAIS Model
- Publisher-driven: LOCKSS Caches
- Official Records: VERS
- METS & PREMIS
- Complex Objects
- OAI-PMH
- LANL's MPEG-21 DID

OAI-PMH Data Model



The MPEG-21 DIDL Model

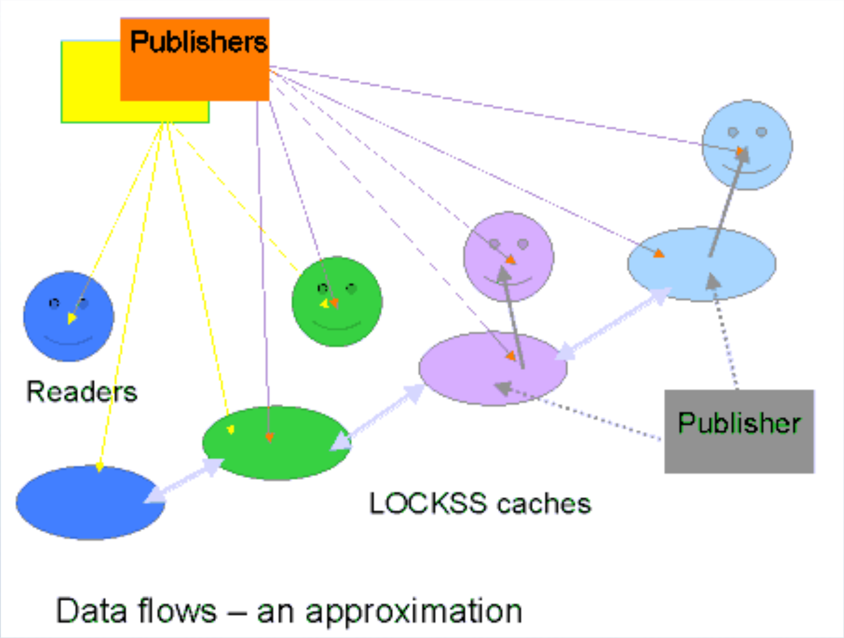


DP Strategy Example: LOCKSS Caches

- The point of LOCKSS is to ensure long-term availability of digital publications *even if the publisher goes out of business*
- Peer-to-peer network is used to maintain and repair content
- Ensures content is only available to authorized subscribers

3 Goals of LOCKSS:

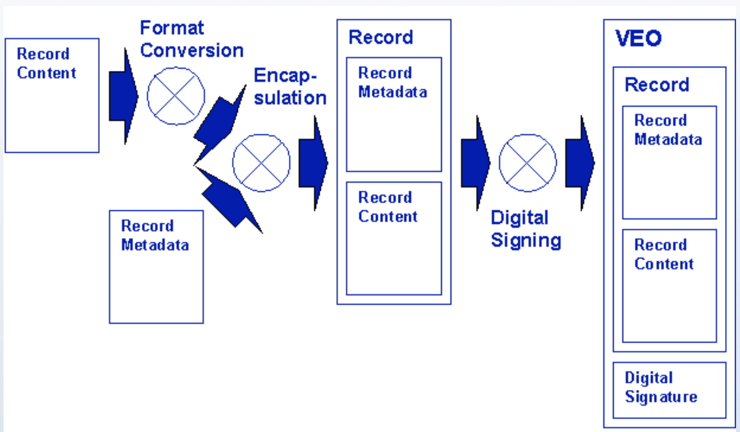
1. Preserve content (bits)
2. Preserve access (to bits)
3. Preserve understanding of bits (as journal content)



In this example, each LOCKSS cache (oval) collects journal content from the publisher's web site as it is published. Readers (circles) can get content from the publisher site. When the publisher's web site is not available (gray) to a local community, readers from that community get content from their local institution's cache. The caches "talk" to each other to maintain the content's integrity over time .

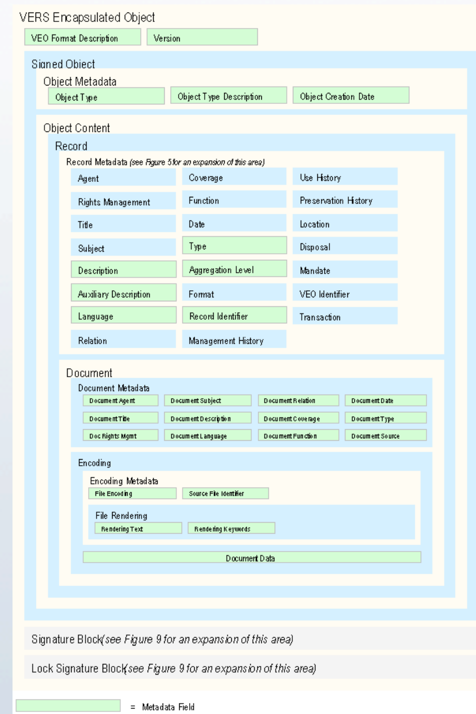
DP Strategy Example: VERS

VERS Process



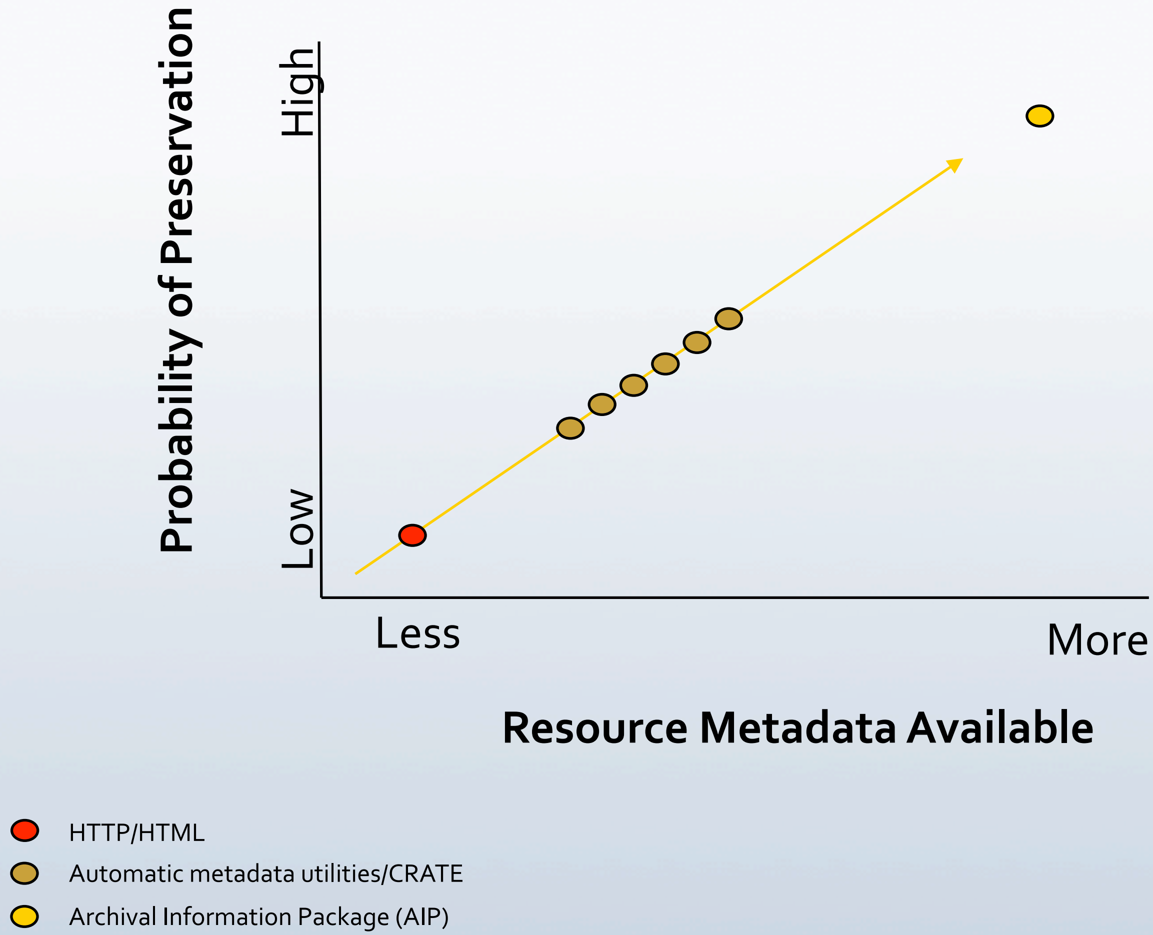
Note the emphasis on **digital signatures**: A key element of official records

VERS Objects



The final object contains a wealth of **metadata**

Preservation & Metadata



Strategies for Optimizing Resource Description

- Use metadata utilities
- Use complex-object technology
- Get the web server to help:
  - Create self-describing resources
  - Complex-Object Response

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

→ A New Type of Website Preservation

- *Integrated* Preservation
- Web server + Metadata Utilities + Sitemap
- Born Digital → now Born *Archival*
- ASCII
- XML
- Undifferentiated, Unverified, Extemporaneous metadata

3 Types of Website Preservation

1. Incidental
   - Casual enthusiast
   - Longer than short-term, shorter than long-term
   - Examples:
     - Usenet restoration by Google
     - User-downloaded site replicas
     - PC Games conversions (pseudo-migration)
2. Intentional
   - Passive
   - Interactive
   - Examples:
     - Internet Archive (Passive) & WARC (Passive-Interactive)
     - Library of Congress (NDIIPP)
     - International Efforts (WARP, European Archive, etc.)
3. Lazy
   - Crawler-dependent
   - Web-Infrastructure dependent
   - Examples:
     - McCown's Warrick Tools
     - NNTP/SMTP Experiments

Search Engine Crawlers & Preservation

- Agent of Preservation: Caches
- Agent of Refreshing: Update Crawls
- Agent of Migration: PDF → Cached view
- Lazy Preservation Agent

- Design influences crawling pattern
- Not all crawlers find all resources, even when fully linked

Limitations of HTTP

- Minimal Headers designed for immediate file transfer
- Content-negotiation between client and server
- MIME is main resource-typing tool
- Here-and-now MIME type, not yesterday-tomorrow

```
% 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.
```

OAI-PMH: Empowering HTTP

We said we need a way to

- Get a list of all URLs for the site
- Get a list of changes (new, gone, altered) since last visit
- Get a list by some grouping we specify (e.g., MIME)

OAI-PMH gives us these options

- Works a lot like CGI-style URLs you may see:

  http://www.foo.org/ask.php?pid=3244&uid=jsmith (PHP-enabled web server)

  http://www.foo.org/oaiserver?verb=Identify (OAI-PMH-enabled web server)
- It is designed for the robot, not the browser
  - Gives back valid, XML-formatted response
- **mod_oai** is an Apache 2 module that allows OAI-PMH verbs to be used on the web site

Web Preservation: Assumptions & Caveats

- It is worthwhile
  - "Everyday" web sites matter
  - Culturally important content exists at many levels
- It is difficult
  - Wide variety of content types create migration issues
  - Hidden resource paths (accidental/ intentional)
- Solutions are needed
  - No standard exists
  - Existing repository approaches are impractical
- Sites will cooperate
  - Webmasters will try a reasonable solution if it is easy
  - Site owners want content preserved
- Backup ≠ Preservation
  - Backups are duplicates for near-term recovery
  - Preservation means long term (more than 5 years!)
- Text-Based protocols & encodings will survive
  - HTTP/XML/ASCII/Base64
  - Human-readable and machine-useable = durable
  - UTF-8 or ASCII

Lessons from the Search Engines: Make It Easy

Evolution of Search on the web

1. Hard to use/Poor results ↔ Few users (think: alta-vista)
2. Easier to use/OK results ↔ More users (think: Ask Jeeves)
3. Simple to use/Great results ↔ Everybody Googles

Search Engines turbo-charge the internet

- At-Your-Fingertips browsing = immediate user benefit
- Search Engines are successful (finally)
- Search Engines are *easy*

Digital Preservation is not like Search Engines

- Digital preservation requires heroic effort & constant vigilance
- Benefits usually accrue only *after* a disaster

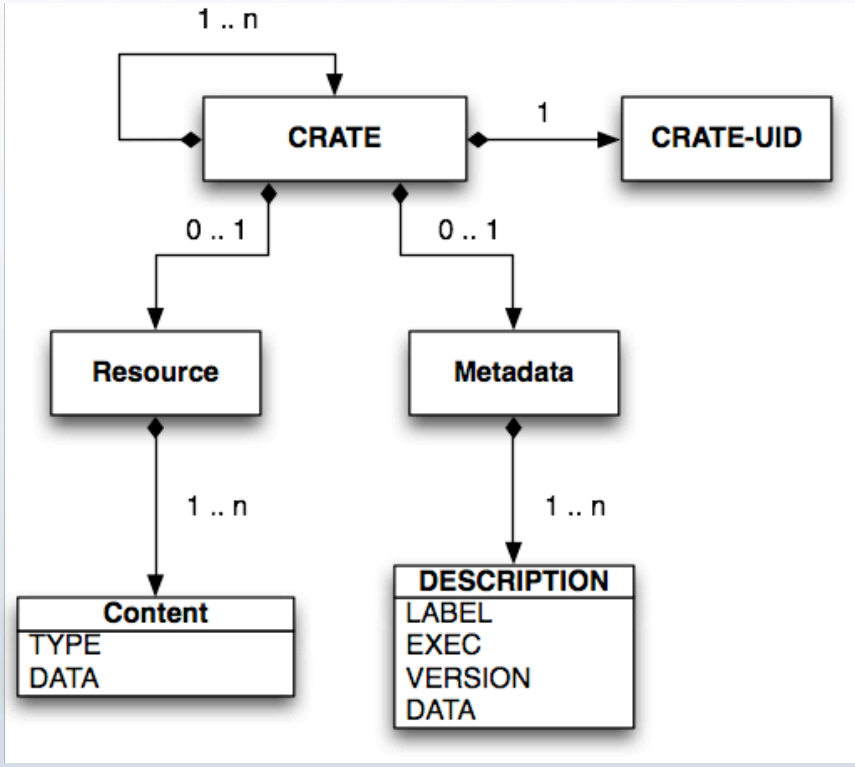
How can we make preservation easy?

We need to find resources

We need to package resources

Why not use the web server itself?

CRATE UML Diagram



Conclusions

- 3 Existing Preservation Approaches
  - Incidental
    - Long-term backup by enthusiasts
    - Migration from A to B
  - Intentional
    - Passive
    - Interactive
  - Lazy
    - Crawler-dependent
    - Web-Infrastructure-based
- **New Approach: Integrated**
  - Counting Problem
  - Representation Problem
  - Resources re-born archival
  - Enhances existing approaches
    - Casual archivist
    - Professional preservationist
    - Lazy webmaster