Web Archiving Service (WAS)

An NDIIPP collaboration: LC/CDL/UNT/NYU

9 November 2005

John Kunze, California Digital Library
Summary

- “Web-at-Risk” Project Partners
- Curator-Operated Web Crawls
- W/ARC Storage Format
- ARK Persistent Identifiers
- Geographic Replication just in case
- Data Desiccation just in case
“Web-at-Risk” Project

• Goal: to develop WAS
  – A web archiving service toolset…
  – to be used by curators/selectors in libraries
  – to capture, curate, and preserve collections of web-based government and political information

• Support: Library of Congress NDIIPP grant
  – California Digital Library (CDL) for 10 UC libraries
  – University of North Texas (UNT)
  – New York University (NYU)
Extended Family of Partners

**The Library of Congress**
- Martha Anderson, Agreement Officer
- Technical Representative

**California Digital Library**
- (Lead Partner)
  - Dan Greenstein, P.I.
  - Peter Brantley, Director of Technology
  - Lynne Cameron, Manager, Ingest Group
  - Patricia Cruse, Director
  - John Kunze, Technical Lead
  - Shifra Raffel, Web Archive Programmer
  - Tracy Seneca, Web Archive Coordinator
  - Stu Sugarman, Arch. and Insfrast. Manager

**University of North Texas**
- Cathy Hartman, Fellow, Texas Center for Digital Knowledge
- Kathleen Murray, Assessment Analyst

**New York University**
- David Ackerman, Executive Director of eServices
- Rasan Rasch, Web Archive Programmer

**Selection and Curator Agents**
- Arizona State Library
- New York University
- Stanford University
- **University of California:** UC Berkeley, UC Davis, UC Irvine, UCLA, UC Riverside, UC Santa Barbara, UC Santa Cruz, UC San Diego
- **University of North Texas**
Curator-Operated Web Crawler

• Extends collection-building to the web
• Functional areas
  – Specify seeds, initiate, and monitor crawls
  – Analyze and group crawls in collections
  – Annotate collections, crawls, seeds, pages
  – Search and browse local archive
  – Publish and/or preserve
# Collection Development

<table>
<thead>
<tr>
<th>Policy Setting</th>
<th>Political mandates, organizational mission, financial parameters, and technical capabilities.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Selection</td>
<td>Factors: Focus of the collection, unit of selection, web boundaries, copyright obligations, and authenticity of materials.</td>
</tr>
<tr>
<td>Acquisition</td>
<td>Requirements for crawling tools: Global or selective capture.</td>
</tr>
</tbody>
</table>
| Description    | Baseline metadata: Machine-generated.  
|                | Enriched metadata: Specific to an organization; both human-generated and machine-generated metadata. |
| Organization   | Considerations: Retain or modify the organizational structure of the materials as they existed on the web. |
| Presentation   | Considerations: Mirror the web at the time of their capture or selectively present (searching and browsing). |
| Maintenance    | Functions: Training, hardware and software maintenance, performance optimization, backups, upgrades, and duplicate detection. |
| Deselection    | Reasons: Duplication, errors, legal or social considerations. |
| Preservation   | Challenges: Persistent naming, replication, desiccated data (low-tech, long-lived derivatives), format migration and/or emulation, inventory management, volatility, re-validation, and storage. |
Running Crawls

• Curator specifies crawl:
  – Seedlist (entry point URLs, validated first)
  – When to start
  – How often to re-harvest

• Obtaining status of in-progress crawl
  – Option to abort

• Primary crawler for WAS will be Heritrix
Organizing Crawls

Analyzing and grouping functions:
• Inspect crawl logs and errors
• Create and name collections
• Move crawls between collections
• Add metadata to collections, crawls, seeds, and pages
• Keep log of rights clearance discussion
• Delete, deselect, abandon
Search and Browse Functions

Inspecting locally held harvested data by
• Browsing from a given point in archive
• Searching an archive or collection
  – Only latest crawled page visible (google-ish)
  – All crawls of a page visible (time-dimension)

Likely tools:
  – Indexes by XTF or NutchWAX (Lucene-based)
  – Query and navigation by WERA
  – Thanks to IA and IIPC for these and for Heritrix
Finally, Publish and/or Preserve

- Publish: ready for public to see
  - Questions: publish but let perish? Never publish?
- Preserve: steps to improve redundancy that imply extra storage and management
  - Geographic replication
  - Deriving desiccated (low-tech, long-lived) formats

That completes the curatorial tool view.
What’s digital preservation?

• Stored objects that remain usable and faithful to the creators’ original intention

• How? By safeguarding information’s …
  – Viability (intact bit streams)
  – Renderability (by machines)
  – Understandability (by humans)

• Viability to be addressed by periodic checksum audit and media refresh
  – Damaged objects to be healed by access to redundant, geographically distributed holdings
W/ARC Storage Format

- WARC = Web ARChive file format
- Next generation of ARC, called for by IIPC
  - ARC format created by the Internet Archive
  - Over 600TB of ARCs gathered since 1996
- An ARC or WARC file is a simple sequence of content blocks, each introduced by a small text header
  - ARCs for crawlers to write captured content easily
  - WARC for captured and related content blocks
- Support in Heritrix; later (?) Alexa, HTTrack
W/ARC Status

• WARC web archiving extensions to ARC on ISO’s radar
• WARC remains simple, open, fast, and general-purpose
  – E.g., LANL journal archiving
• Work in progress:
  http://cvs.sourceforge.net/viewcvs.py/archive-access/archive-access/src/docs/warc/
• Co-authors: Allan Arvidson, John Kunze, Gordon Mohr, Michael Stack; comments to
  – jak@ucop.edu
• Support first in Heritrix; later (?) Alexa, HTTrack
W/ARC File Anatomy

W/ARC File

W/ARC Record

Text header

Content block

Length, also source URI, date, type, …

E.g., HTTP response headers and length bytes of HTML, GIF, PDF, …

Append at will
W/ARCs and Crawling

• One crawl often spans multiple W/ARCs
• W/ARC: records are order-independent
  – File can be exploded and recombined easily
  – File can be used as a container for anything
• Typical start of a W/ARC:
  – file-descriptive record; dns:foo.bar; http://foo.bar/robots.txt; and then first interesting content
How WARC improves on ARC

- Stored metadata linked to other stored data (e.g., subject classifier, discovered language, encoding)
- Per-record compression and integrity checks
- Stored request protocol information
- Stored results of data migrations
- Stored revisit events (for duplicate reduction)
- Handling of long records (e.g., truncation, segmentation)
- **The glue**: stored globally unique record identifiers
ARK Persistent Identifiers

• ARK = Archival Resource Key
  – A URL-based scheme for high service quality names

• ARK is *nothing* to do with ARC, but will be our WARC glue
  – The key link from one WARC record to another

• An ARK gives access to not one, but to three things:
  – The object, its metadata, and the provider’s commitment

• Open and lightweight, it tries to minimize dependencies on special-purpose, complex infrastructure and technology
ARK Anatomy

http://foobar.zaf.org/ark:/12025/654xz321/s3/f8.05v.tiff
\___________________/ \__/ \___/ \______/ \____________/
(replaceable) | | | | 4 Qualifier
| ARK Label | | (NMA-supported)
| | | |
1 Name Mapping Authority | 3 Name (NAA-assigned)
Hostport (NMAH) |

2 Name Assigning Authority Number (NAAN)

1 = current service provider; identity inert; replaceable
2 = organization that originally assigned the id
3 = name originally assigned to the abstract object, often opaque
4 = extension disclosing object hierarchy & variants, often non-opaque
ARK Usage

Two ARKs accessing the same thing
http://loc.gov/ark:/12025/654xz321
http://rutgers.edu/ark:/12025/654xz321

Access to metadata -- add a ‘?’
http://loc.gov/ark:/12025/654xz321?

Access to support statement -- add ‘??’
http://loc.gov/ark:/12025/654xz321??

Two reserved characters:
. and /
Opaque Identifiers

- Persistence cannot be improved by syntax
  - But can be worsened by semantic rot and user irritation
  - Problems are political and social, not technical
- NOID (Nice Opaque IDentifiers) can be used to generate identifiers that minimize these problems
- Mints in random or sequential order, with or without a check character
  - Can be used with any identifier scheme (eg, Handles)
  - Open-source, available through CDL
Some ARK Namespaces

12025 National Library of Medicine
12026 Library of Congress
12027 National Agriculture Library
13030 California Digital Library
13038 World Intellectual Property Organization
20775 University of California San Diego
29114 University of California San Francisco
28722 University of California Berkeley
15230 Rutgers University Libraries
13960 Internet Archive
64269 Digital Curation Centre
62624 New York University Libraries
67531 University of North Texas Libraries
27927 Portico
12148 National Library of France

To reserve a namespace, email to ark@cdlib.org
Geographic Replication

- A web crawl to be stored in CDL’s repository system as an aggregate of intact WARC$s and crawl tailings
- Currently using SRB for storage
  - Storage Resource Broker system (SDSC)
  - Ready to configure for replicate-on-write
- Have not yet committed to a replication strategy
Data Desiccation

- Always save the original object
- As a hedge, we are exploring the derivation of simple, low-tech, desiccated formats, each of which has an excellent preservation outlook
- We may never in the future have time, tools, or funding to convert most objects
  - Rendering tools near peak functionality can derive simple text and image formats upon ingest
  - Should the original ever fail, the alternates may yet retain most of the original’s cultural value
Migration and Emulation

• Migration problems
  – Unknown costs, human review, format errors

• Emulation problems
  – Unknown costs, human review, software IP

• Two approaches to the same problem:
  – Trying to maintain a match between an object and a technical context in which the object is still renderable and understandable

• Can reducing the technical context will make preserving the object easier?
The Lesson from Paper

• As a recording and display device
  – Can last for 1000 years

• Why this astonishing performance?
  – No technical intermediation required

• So when intermediation *is* required…
  – The simplest technologies to maintain and understand today are the simplest to carry forward and to recreate in the future
Low-Tech Dependencies

• Language as technology
  – Loss is inevitable due to linguistic shifts
  – Compare to loss in sensitive hi-tech media

• Low-tech dependencies
  – Paper: light source (sun, candle)
  – Microfilm: light + lens (500-years of optics)
Desiccated Data

• Lesson from the longest-lived online digital format
  – Plain text archives of IETF internet RFCs
  – High in value, low in features

• Preservation through “desiccation”
  – No fonts, graphics, colors, diacritics, etc.
  – But essential scholarly value retained
Hedging our Bets

• Always save the original format
• In addition, derive desiccated formats in case the original format ever fails
• Extra storage cost may be incurred anyway if your access system requires a plain text derivative for search indexing
• Question: what about Latin-1 support
• Question: surfacing hidden features
Next Lowest-Tech Technology

• Raster image as alternate desiccated format
  – Rectangular grid of picture elements
• Rendering tools will never be better than at peak of format’s popularity
  – Very common malformed format instances
• Additional fall back format in case the original and plain text versions fail
• Question: pressure to compress
• Question: surfacing hidden data
Conclusion

• We’re building curator-operated web crawling tools to enable libraries to extend their historic collection-building roles into the web
• The WARC file format promises a batch of new and quite general features for archives
• ARK persistent identifiers can be created easily with the open-source NOID software
• Replication and desiccation are steps that we can take today