From Seed to Harvest: Web Archiving Program Considerations for SUL

Nicholas Taylor
@nullhandle

Stanford University Libraries
April 17, 2013
hello, my name is Nicholas...
The Library of Congress Web Archives (LCWA) is composed of collections of archived web sites selected by subject specialists to represent web-based information on a designated topic. It is part of a continuing effort by the Library to evaluate, select, collect, catalog, provide access to, and preserve digital materials for future generations of researchers.

The early development project for Web archives was called MINERVA.
## Web Archiving Life Cycle Model

### Program Elements
- Vision and Objectives
- Resources and Workflow
- Access / Use / Reuse
- Preservation
- Risk Management

### Workflow Elements
- Appraisal and Selection
- Scoping
- Data Capture
- Storage and Organization
- Quality Assurance and Analysis
Web Archiving

PROGRAM ELEMENTS
web archiving program vision

ePADD Discovery Module

iiif

AIMS Born-Digital Collections

LOCKSS

DPN

blacklight

PASIG

ePADD Discovery Module

Google books

PVW

HighWire

SIRPL

hydra
“The Stanford University Libraries (SUL) is more than a cluster of libraries; it connects people with information by providing diverse resources and services to the academic community.”

“Stanford University Libraries...develops and implements resources and services...that support research and instruction.”
“DLSS is the information technology production arm of the Stanford Libraries; it serves as the digitization, digital preservation and access systems provider for SUL; and it is the research and development unit for new technologies, standards and methodologies related to library systems.”

SUL: “SULAIR Digital Library Systems and Services (DLSS)”

SUL: “New Images of Rare Books and Digitization Devices”
The web archiving program will provide capabilities for the acquisition, preservation, and dissemination of resources that are increasingly and, often, exclusively accessible via the web that are necessary to support University research, instruction, and other purposes.
objectives

• build infrastructure
• develop expertise
• create research collections
• archive records and deprecated content
• mirror government documents
cost modeling
staffing

• service manager
• crawl engineer
• curators
• system administrators
• software engineers
• technical services
• legal counsel

"Digitizing Mark Adams cartoons" by Flickr user suldpq under CC BY-NC-SA 2.0
infrastructure
readily workflow-able

- collection management
- site nomination
- permissions tracking
- crawl scheduling
- data capture
- quality assurance
workflow challenges

- test crawling
- automated QA
- AIP/DIP generation
- SDR ingest
- indexing
- enabling access
- tools testing
access policy

- dark archive
- data redistribution
- embargo
- onsite/offsite replay
- takedown requests
browse and API: Wayback

Internet Archive: “Wayback Machine”

UK Web Archive: “Wayback Machine”
many Wayback Machines

List of Web archiving initiatives

This article uses bare URLs for citations. Please consider adding full citations so that the article remains verifiable. Several templates and the Reflinks tool are available to assist in formatting. (Reflinks documentation) (February 2012)

This page contains a list of Web archiving initiatives worldwide. For easier reading, the information is divided in three tables: Web archiving initiatives, archived data and access methods.

This Wikipedia page was originally generated from the results obtained for the research paper A survey on web archiving initiatives[1] published by the Portuguese Web Archive[2] team.

Contents [hide]
1 Web archiving initiatives
2 Archived data
3 Access methods
4 References

Web archiving initiatives

<table>
<thead>
<tr>
<th>Name</th>
<th>Country</th>
<th>Creation Year</th>
<th>Technologies</th>
<th>Number of Employees</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Australia’s Web Archive[3]</td>
<td>Australia</td>
<td>1996</td>
<td>PANDORA Digital Archiving System (PANDAS), NLA Trove, HTTrack</td>
<td>10</td>
<td>The PANDORA Archive which takes a selective approach is a collaborative program of 11 agencies that provide an estimate average monthly staffing equivalent to around 10 FTE. It is supported by the NLA. 0.25 person-month. Whole domain harvests have been conducted annually since 2005 in collaboration with the Internet Archive using Heritrix, Wayback Machine.</td>
</tr>
<tr>
<td>PageFreezer’s Deep</td>
<td>Canada, US</td>
<td>1996</td>
<td>PageFreezer’s Deep</td>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>
### List of Web archiving initiatives

From Wikipedia, the free encyclopedia

This page contains a list of Web archiving initiatives, archived data and access methods.

This Wikipedia page was originally created by the Portuguese Web Archive (2012) and was updated by several contributors (February 2012).

#### Contents
- Web archiving initiatives
- Archived data
- Access methods
- References

#### Web archiving initiatives

<table>
<thead>
<tr>
<th>Name</th>
<th>Country</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Australia's Web Archive</td>
<td>Australia</td>
<td></td>
</tr>
<tr>
<td>Our digital island, a Tasmanian Web Archive</td>
<td>Australia</td>
<td></td>
</tr>
<tr>
<td>PageFreezer's Deep</td>
<td>Canada, US.</td>
<td></td>
</tr>
<tr>
<td>HTTrack</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Experientially: Web Curator, Heritrix, and Wayback Machine</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Enterprise Class On Demand service to archive and replay web</td>
<td></td>
<td></td>
</tr>
<tr>
<td>blogs, Ajax, Flash, video, audio &amp; social media for litigation protection</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
### Web Archiving
- **Published:** 2006
- **At the Library:** Green » Stacks » (No call number)
- **Online:** [www.springerlink.com + 2 more sources](http://www.springerlink.com)
- **Format:** Book

### California State Government [electronic resource] : ca.gov web archive
- **Search database:** webarchives.cdlib.org
- **Subjects:** Government Information: United States
- **Summary:** The ca.gov domain is the 3rd largest domain in the U.S. government web. A great deal of that content is published by the State of California, and represents a rich source of material for researchers, students, the general public and state agencies themselves. This archive preserves access to hundreds of state agency sites that are captured on a regular basis. State agencies utilize their websites to publish everything from press releases, agendas, minutes, events, reports and statistics. This material is especially volatile as leadership changes or as time sensitive issues are no longer on agendas or in the news. This information while not distributed in print still holds immense social and research value, which we as memory institutions have an obligation to collect and preserve. By collecting and preserving this information, the study of issues in California, the world’s fourth largest economy, will be greatly enhanced. The archive is maintained by government information specialists across several UC campuses and the Stanford University Libraries with support from the California Digital Library.

### Archiving websites : a practical guide for information management professionals
- **Author:** Brown, Adrian, 1959-
- **Published:** 2006
- **At the Library:** SAL3 » Stacks » Z701.3.W43 876 2006
- **Format:** Book
full-text search: Solr
Preservation
bit preservation
preservation engineering
Risk Management
Risk Management

- “appified” web
- copyright
- ephemeral web
- financial sustainability
- fostering use
Policy
copyright

- § 108 (library exceptions)
- fair use
- notification vs. permission
- opt-out / takedown
- robots.txt
- third-party sites
- exceptions?

“Noria con Copyrights” by Flickr user Alex Novoa under CC BY-NC-ND 2.0
collection development
Web Archiving

WORKFLOW ELEMENTS
Appraisal and Selection
informing selection

• value
• risk
• size
• extent to which archived

“Fruit market-Barcelona” by Flickr user Marcel Theisen under CC BY-NC-SA 2.0
Welcome

Selection of web archive is a manual process that relies upon people to select quality sites and nominate or submit them to web archives for inclusion. Past (national and/or international) efforts to automate selection have either failed to convince staff that automation was identifying quality resources, or have focused on providing an interface for selectors to submit sites rather than carry out the selection per se. Selection has thus remained, for most institutions, an element of the workflow wholly dependent on contributions from externals. The number of selectors providing sites is typically small and their contributions are inevitably subjective. The resulting collections, whilst immensely valuable, are therefore mostly representative of the expertly selected sites and do not fully represent the sites frequently used in a more social setting.

Crowd sourcing is an opportunity to develop a new approach to this problem. It taps into the growth of social networks to outsource tasks typically performed by an employee or contractor, to an undefined, large group of people or community (a ‘crowd’) (Wikipedia, 2011). It is a particularly attractive option in the current economic climate, where we are all being asked to ‘do more with less’. A number of cultural heritage institutions and/or projects have already begun to leverage the power of the crowd for digitised collections, including the National Library of Australia (through Trove), the Transcribe Bentham project at ULCC, and the National Library of Finland (through the DigitalKoot program). No such projects have yet been launched for web archives.

This project will develop an automated approach to selection for web archiving based on the principles of crowd sourcing. It has been awarded partial funding from the International Internet Preservation Consortium and supports the forthcoming web archiving strategy by increasing the number of selections and automating the selection process.
Wikipedia Live Monitor

Articles Edited Right Now:

Merging Right Now:

None yet.
Monitoring...

Article Clusters Repeatedly Edited In Short Intervals:

None yet.
Monitoring...

Breaking News Candidate Article Clusters Right Now:

None yet.
Monitoring...

Proudly Developed in Europe By Thomas Steiner (tomas@google.com)
It's Open Source and Apache 2.0 Licensed (Fork This Project Or Github)
List of think tanks in the United States

This is a list of think tanks in the United States.

This is an incomplete list, which may never be able to satisfy particular standards for completeness. You can help by expanding it with reliably sourced entries.

A–M

- Acton Institute
- Alexis de Tocqueville Institution
- Allegheny Institute for Public Policy
- American Action Forum
- American Civil Rights Union
- American Consumer Institute
- American Enterprise Institute
- American Foreign Policy Council
- American Institute for Economic Research
- American–Iranian Council
- American Israel Public Affairs Committee
- American Security Council Foundation
- Americans for the Arts
- Analysis Group
- Askew Institute on Politics and Society
- Aspen Institute
- Atlantic Council
- -
Active Projects
End of Term 2012 - Bulk Lists
Papal Transition 2013
Past Projects
Election 2012 Collaborative Web Archive Collection
End of Term Presidential Harvest 2008
End of Term Presidential Harvest 2012
Gulf Oil Spill
Olympics 2010
Olympics 2012
Paralympics 2012
Resignations of Gretchen Batalis
Vaclav Havel

Request a Project
If you would like to start a nomination project, please send an email to laurenm@unt.edu to initiate the setup process.
Scoping
the purpose of scoping
Data Capture
other data capture tools

Dan Chudnov and Laura Wrubel: "social feed manager"

Archive Team: "Wget with WARC output"

Mat Kelly: "WAIL"

Dan Chudnov and Laura Wrubel: "social feed manager"
the elusive web

“Light Writing - Spider Web” by Flickr user forcefeed:swede under CC BY-ND 2.0
scale
Storage and Organization
packages and their contents
Quality Assurance and Analysis
QA before, after, during
Hello! My URL is... www.elogxation.com

"Hello! My URL Is..." by Flickr user vasta under CC BY-NC-ND 2.0
Considerations

BEYOND THE MODEL

“My donut” by Flickr user Molemaster under CC BY-NC-SA 2.0
other program requirements

- marketing/outreach
- performance metrics
- service level definitions
- service roadmap
- training
- user documentation

“Sticky notes” by Flickr user Kris Krug under CC BY-SA 2.0
incorporating existing projects

- plan capacity
- normalize data
- ingest into SDR
- seek permissions
- process
- catalog
- enable access
community engagement
the web changes
Nicholas Taylor
@nullhandle