From Seed to Harvest: Web Archiving Program Considerations for SUL

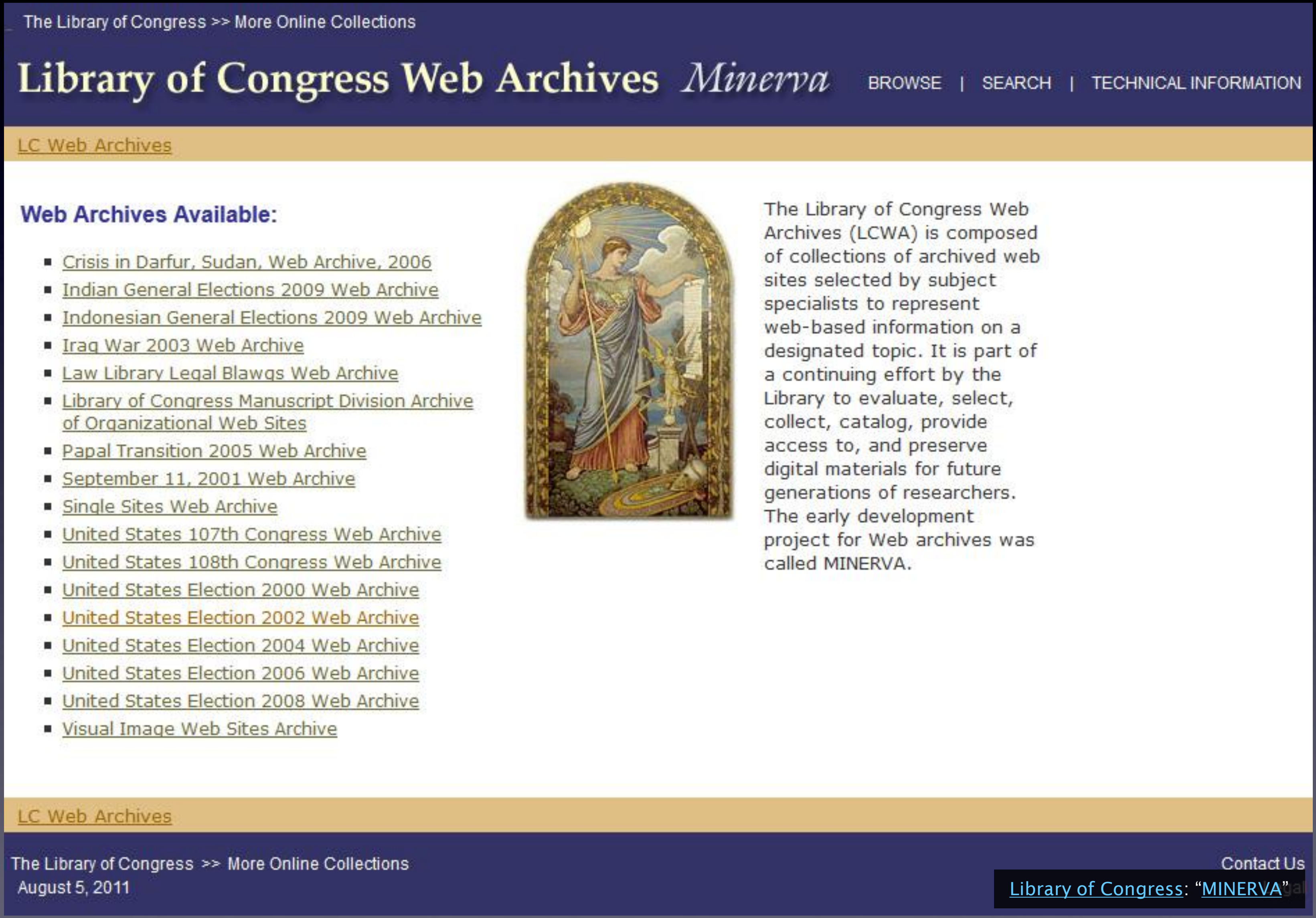
Nicholas Taylor
@nullhandle

Stanford University Libraries
April 17, 2013

"Digital" by Flickr user clickclaker under CC BY-NC-ND 2.0

hello, my name is Nicholas…

Library of Congress Web Archiving



The Library of Congress >> More Online Collections

Library of Congress Web Archives *Minerva*

BROWSE | SEARCH | TECHNICAL INFORMATION

LC Web Archives

Web Archives Available:

- Crisis in Darfur, Sudan, Web Archive, 2006
- Indian General Elections 2009 Web Archive
- Indonesian General Elections 2009 Web Archive
- Iraq War 2003 Web Archive
- Law Library Legal Blawgs Web Archive
- Library of Congress Manuscript Division Archive of Organizational Web Sites
- Papal Transition 2005 Web Archive
- September 11, 2001 Web Archive
- Single Sites Web Archive
- United States 107th Congress Web Archive
- United States 108th Congress Web Archive
- United States Election 2000 Web Archive
- United States Election 2002 Web Archive
- United States Election 2004 Web Archive
- United States Election 2006 Web Archive
- United States Election 2008 Web Archive
- Visual Image Web Sites Archive

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.

LC Web Archives

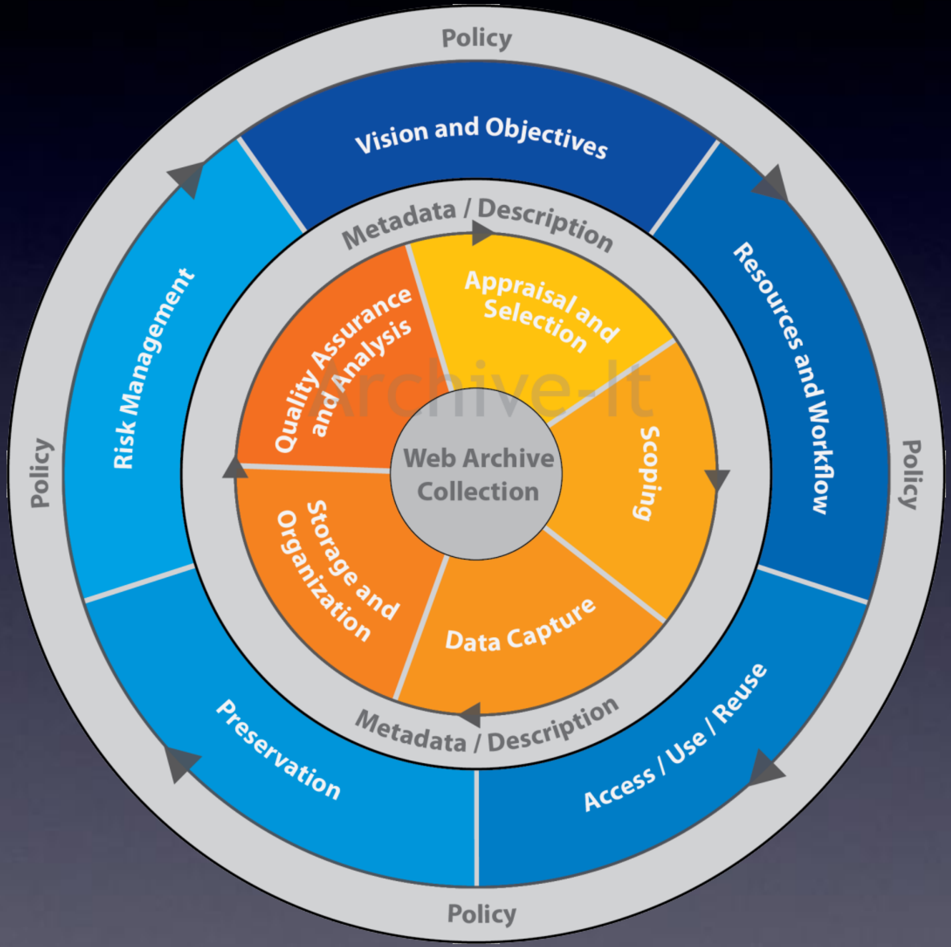
The Library of Congress >> More Online Collections

August 5, 2011

Contact Us

Library of Congress: “MINERVA”

Web Archiving Life Cycle Model



“Web Archiving Life Cycle Model” by M. Bragg, K. Hanna, et al. (2013). Reproduced with permission.

Web Archiving Life Cycle Model

Program Elements

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

Workflow Elements

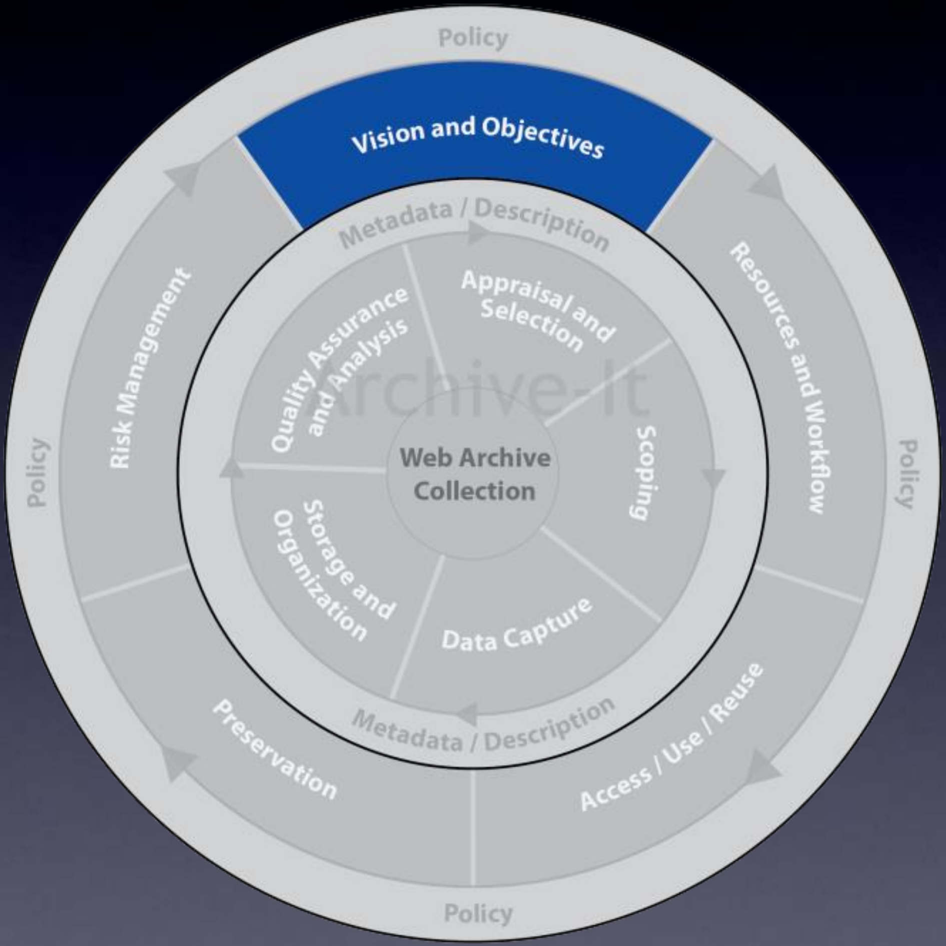
- Appraisal and Selection
- Scoping
- Data Capture
- Storage and Organization
- Quality Assurance and Analysis

Web Archiving

PROGRAM ELEMENTS

“Element Blocks” by Flickr user Asian Art Museum under CC BY-NC-ND 2.0

Vision and Objectives



web archiving program vision

iiif

AIMS Born-Digital Collections

open annotation collaboration

DPN

THE DIGITAL PRESERVATION NETWORK

PASIG

LOCKSS

blacklight

ePADD Discovery Module

Google books

PVW2

HighWire
Stanford University

hydra

SPIRL

SUL mission

"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."*

SUL: "About The Stanford University Libraries"

"Stanford University Libraries...develops and implements resources and services...that ***support research and instruction****."*

SUL: "SULAIR Brief Guide"

SUL: "Stanford University Libraries on Vimeo"

DLSS mission

*"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"

proposed program mission

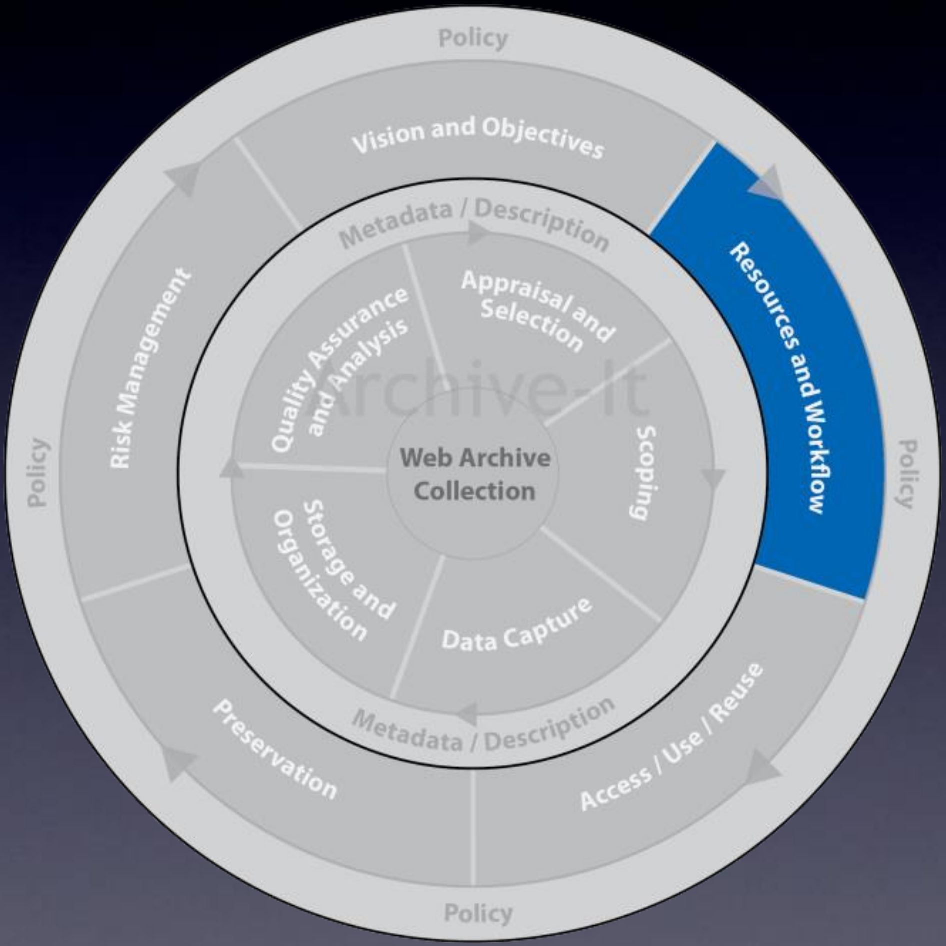
“The web archiving program will provide capabilities for the acquisition, preservation, and dissemination of resources that are increasingly and, often, exclusively accessible via the web that are necessary to support University research, instruction, and other purposes.”

objectives

- build infrastructure
- develop expertise
- create research collections
- archive records and deprecated content
- mirror government documents

"Objective" by Flickr user Pedro J. Ferreira under CC BY-NC-ND 2.0

Resources and Workflow



cost modeling

“dollar butterfly (2)” by Flickr user eikosi under CC BY-SA 2.0

staffing

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

"Digitizing Mark Adams cartoons" by Flickr user suldpg under CC BY-NC-SA 2.0

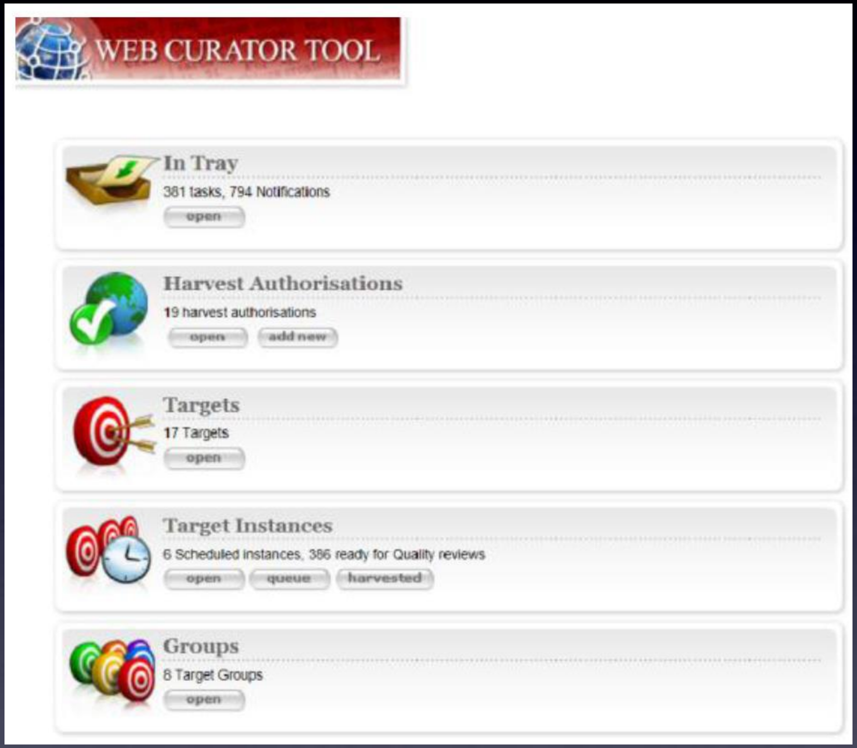
infrastructure

Google Storage Server
Professor Joshua Lederberg, Nobel laureate in Genetics at Stanford,
having used the ARPAnet, foresaw the promise of computer-
mediated communication in IEEE 1978: Digital Communications
and the Conduct of Science: the New Literacy.
It took time for technology to match Lederberg's vision.
Between 1981 and 1990 NSF supported research networks, combining
the ARPAnet and other services into the Internet. In 1991
Maria Zemankova, Ed Fox, and Michael Lesk issued a report on
Digital Libraries. An Interagency Digital Library Initiative
started in 1993 with funding by

"Google Storage Server" by Flickr user Kazuya (Kaz) Yokohama under CC BY-NC-ND 2.0

readily workflow-able

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



WEB CURATOR TOOL

In Tray

381 tasks, 794 Notifications

Harvest Authorisations

19 harvest authorisations

Targets

17 Targets

Target Instances

6 Scheduled instances, 386 ready for Quality reviews

Groups

8 Target Groups

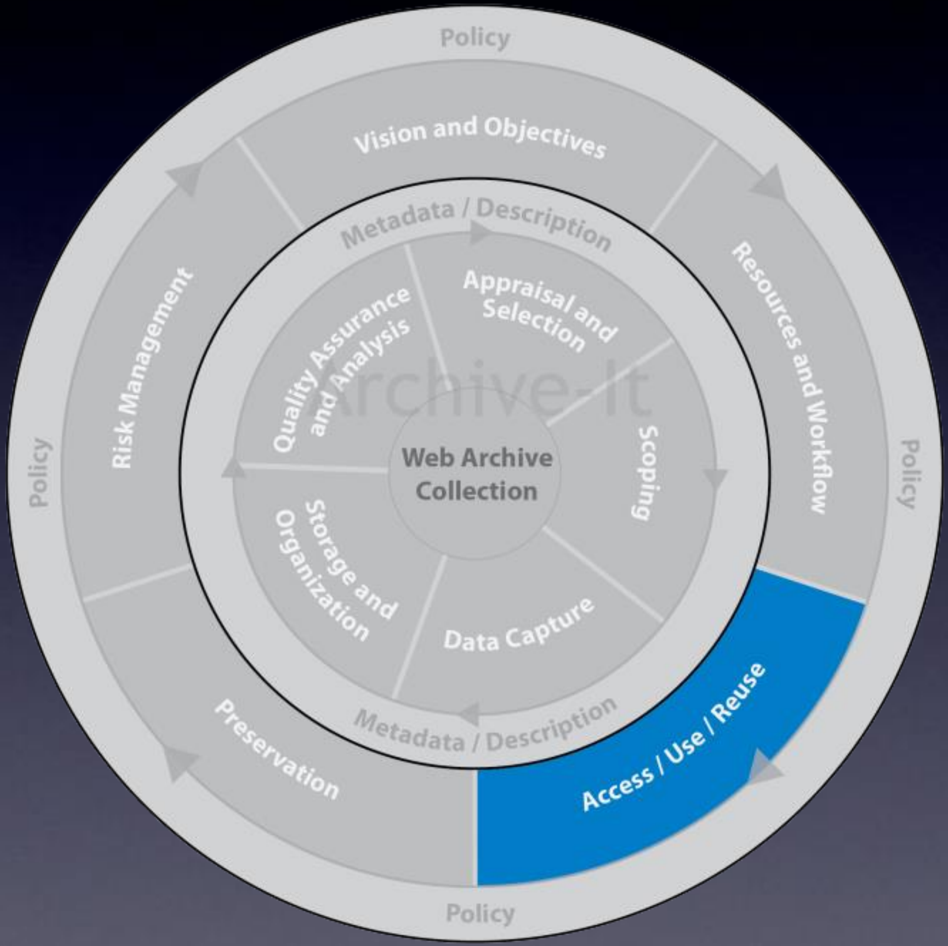
“Web Curator Tool User Manual Version 1.5.2”

workflow challenges

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

"Salmon Ladder at Bonneville Dam" by Flickr user Serolynne under CC BY-NC-ND 2.0

Access / Use / Reuse

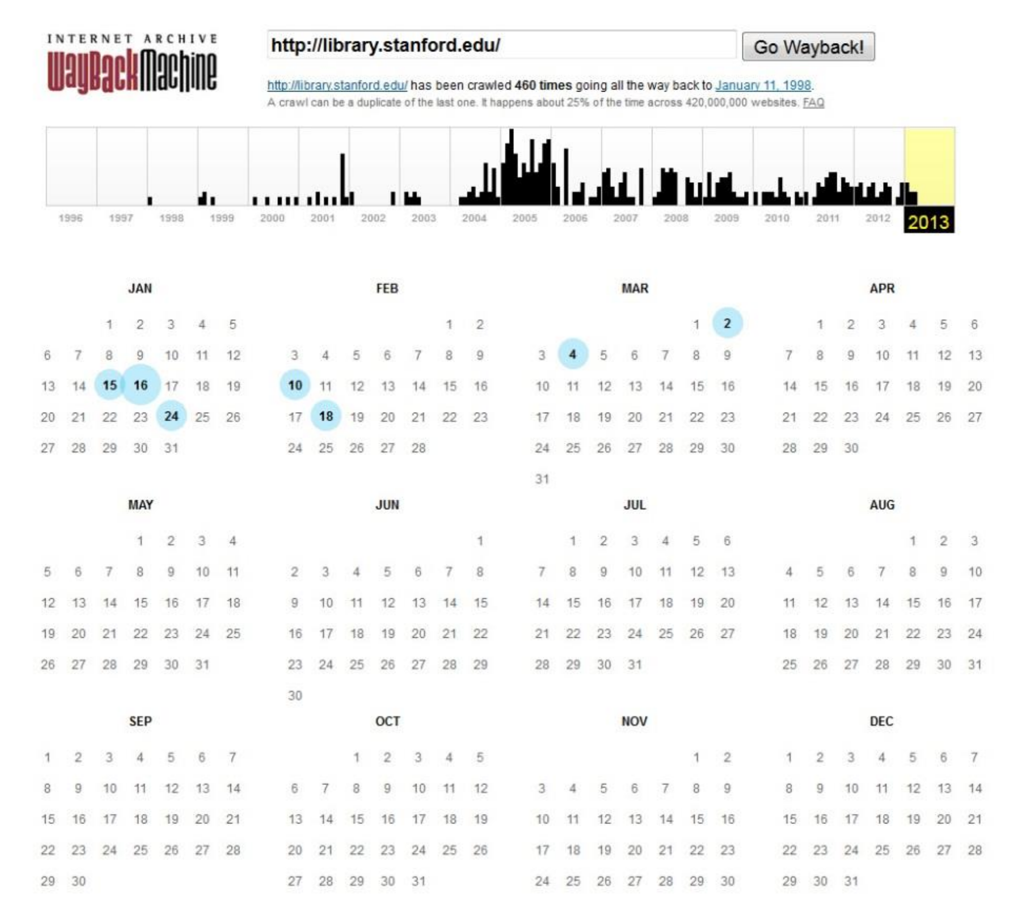


access policy

- dark archive
- data redistribution
- embargo
- onsite/offsite replay
- takedown requests

“DO NOT DUPLICATE” by Flickr user Sam UL under CC BY-NC-SA 2.0

browse and API: Wayback



INTERNET ARCHIVE
WayBackMachine

http://library.stanford.edu/ | Go Wayback!

http://library.stanford.edu/ has been crawled **460 times** going all the way back to January 11, 1998.
A crawl can be a duplicate of the last one. It happens about 25% of the time across 420,000,000 websites. FAQ

JAN

		1	2	3	4	5
6	7	8	9	10	11	12
13	14	15	16	17	18	19
20	21	22	23	24	25	26
27	28	29	30	31		

FEB

					1	2
3	4	5	6	7	8	9
10	11	12	13	14	15	16
17	18	19	20	21	22	23
24	25	26	27	28		

MAR

					1	2
3	4	5	6	7	8	9
10	11	12	13	14	15	16
17	18	19	20	21	22	23
24	25	26	27	28	29	30
31						

APR

	1	2	3	4	5	6
7	8	9	10	11	12	13
14	15	16	17	18	19	20
21	22	23	24	25	26	27
28	29	30				

MAY

			1	2	3	4
5	6	7	8	9	10	11
12	13	14	15	16	17	18
19	20	21	22	23	24	25
26	27	28	29	30	31	

JUN

						1
2	3	4	5	6	7	8
9	10	11	12	13	14	15
16	17	18	19	20	21	22
23	24	25	26	27	28	29
30						

JUL

	1	2	3	4	5	6
7	8	9	10	11	12	13
14	15	16	17	18	19	20
21	22	23	24	25	26	27
28	29	30	31			

AUG

				1	2	3
4	5	6	7	8	9	10
11	12	13	14	15	16	17
18	19	20	21	22	23	24
25	26	27	28	29	30	31

SEP

1	2	3	4	5	6	7
8	9	10	11	12	13	14
15	16	17	18	19	20	21
22	23	24	25	26	27	28
29	30					

OCT

		1	2	3	4	5
6	7	8	9	10	11	12
13	14	15	16	17	18	19
20	21	22	23	24	25	26
27	28	29	30	31		

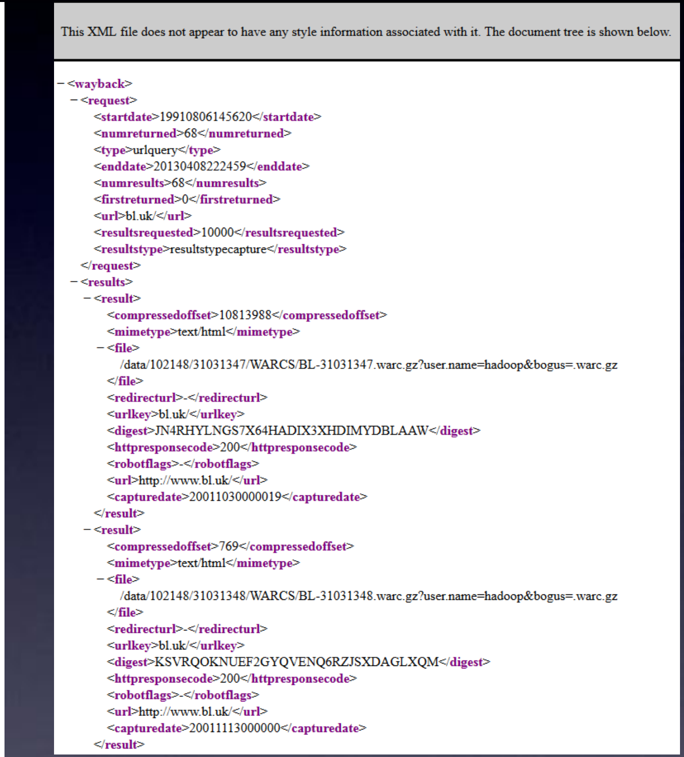
NOV

					1	2
3	4	5	6	7	8	9
10	11	12	13	14	15	16
17	18	19	20	21	22	23
24	25	26	27	28	29	30

DEC

1	2	3	4	5	6	7
8	9	10	11	12	13	14
15	16	17	18	19	20	21
22	23	24	25	26	27	28
29	30	31				

Internet Archive: “Wayback Machine”

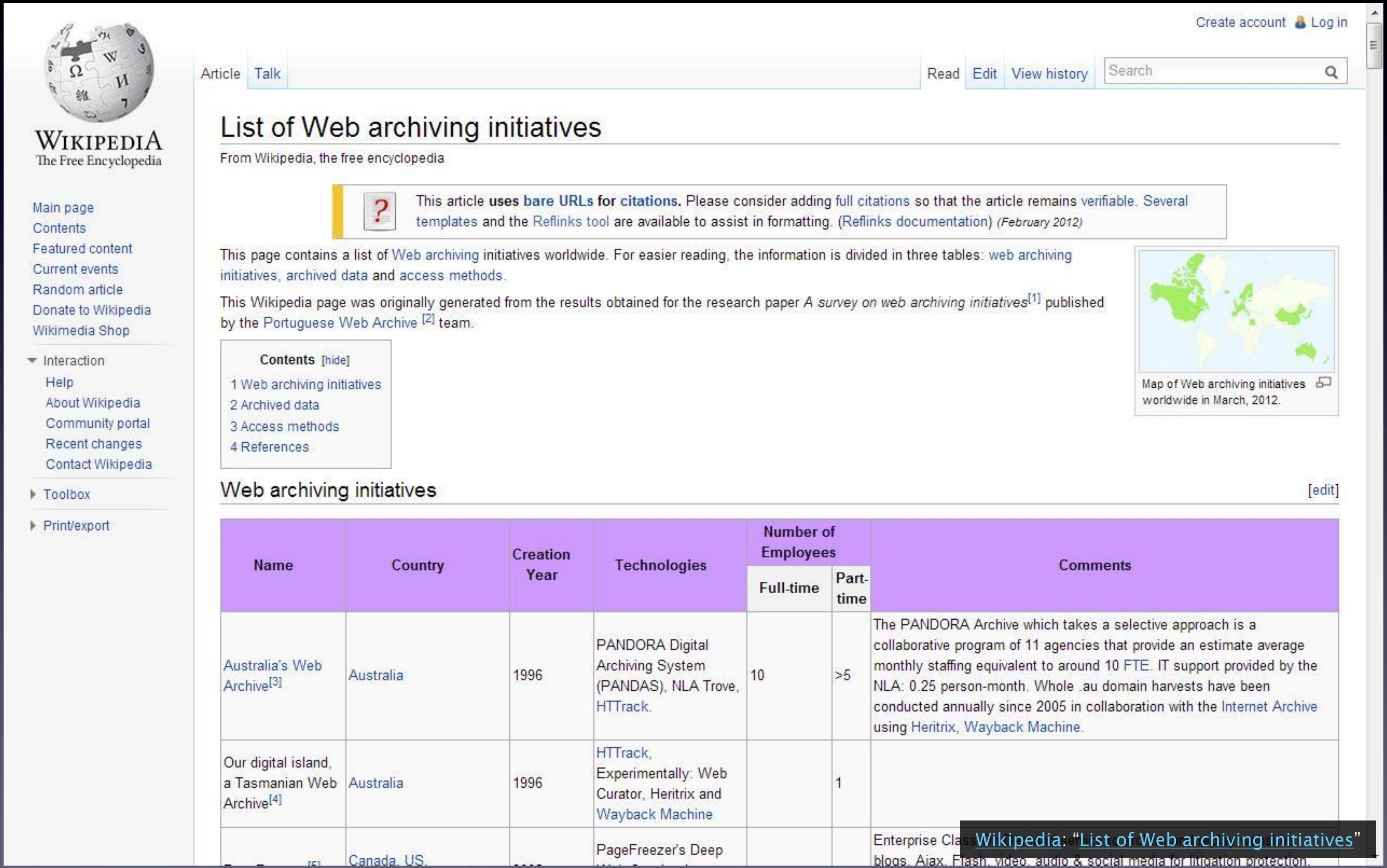


This XML file does not appear to have any style information associated with it. The document tree is shown below.

```
− <wayback>
  − <request>
      <startdate>19910806145620</startdate>
      <numreturned>68</numreturned>
      <type>urlquery</type>
      <enddate>20130408222459</enddate>
      <numresults>68</numresults>
      <firstreturned>0</firstreturned>
      <url>bl.uk/</url>
      <resultsrequested>10000</resultsrequested>
      <resultstype>resultstypecapture</resultstype>
    </request>
  − <results>
    − <result>
        <compressedoffset>10813988</compressedoffset>
        <mimetype>text/html</mimetype>
      − <file>
          /data/102148/31031347/WARCS/BL-31031347.warc.gz?user.name=hadoop&bogus=.warc.gz
        </file>
        <redirecturl>-</redirecturl>
        <urlkey>bl.uk/</urlkey>
        <digest>JN4RHYLNGS7X64HADIX3XHDIMYDBLAAW</digest>
        <httpresponsecode>200</httpresponsecode>
        <robotflags>-</robotflags>
        <url>http://www.bl.uk/</url>
        <capturedate>20011030000019</capturedate>
      </result>
    − <result>
        <compressedoffset>769</compressedoffset>
        <mimetype>text/html</mimetype>
      − <file>
          /data/102148/31031348/WARCS/BL-31031348.warc.gz?user.name=hadoop&bogus=.warc.gz
        </file>
        <redirecturl>-</redirecturl>
        <urlkey>bl.uk/</urlkey>
        <digest>KSVRQOKNUEF2GYQVENQ6RZJSXDAGLXQM</digest>
        <httpresponsecode>200</httpresponsecode>
        <robotflags>-</robotflags>
        <url>http://www.bl.uk/</url>
        <capturedate>20011113000000</capturedate>
      </result>
```

UK Web Archive: “Wayback Machine”

many Wayback Machines



List of Web archiving initiatives

From Wikipedia, the free encyclopedia

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.

Map of Web archiving initiatives worldwide in March, 2012.

Contents [hide]

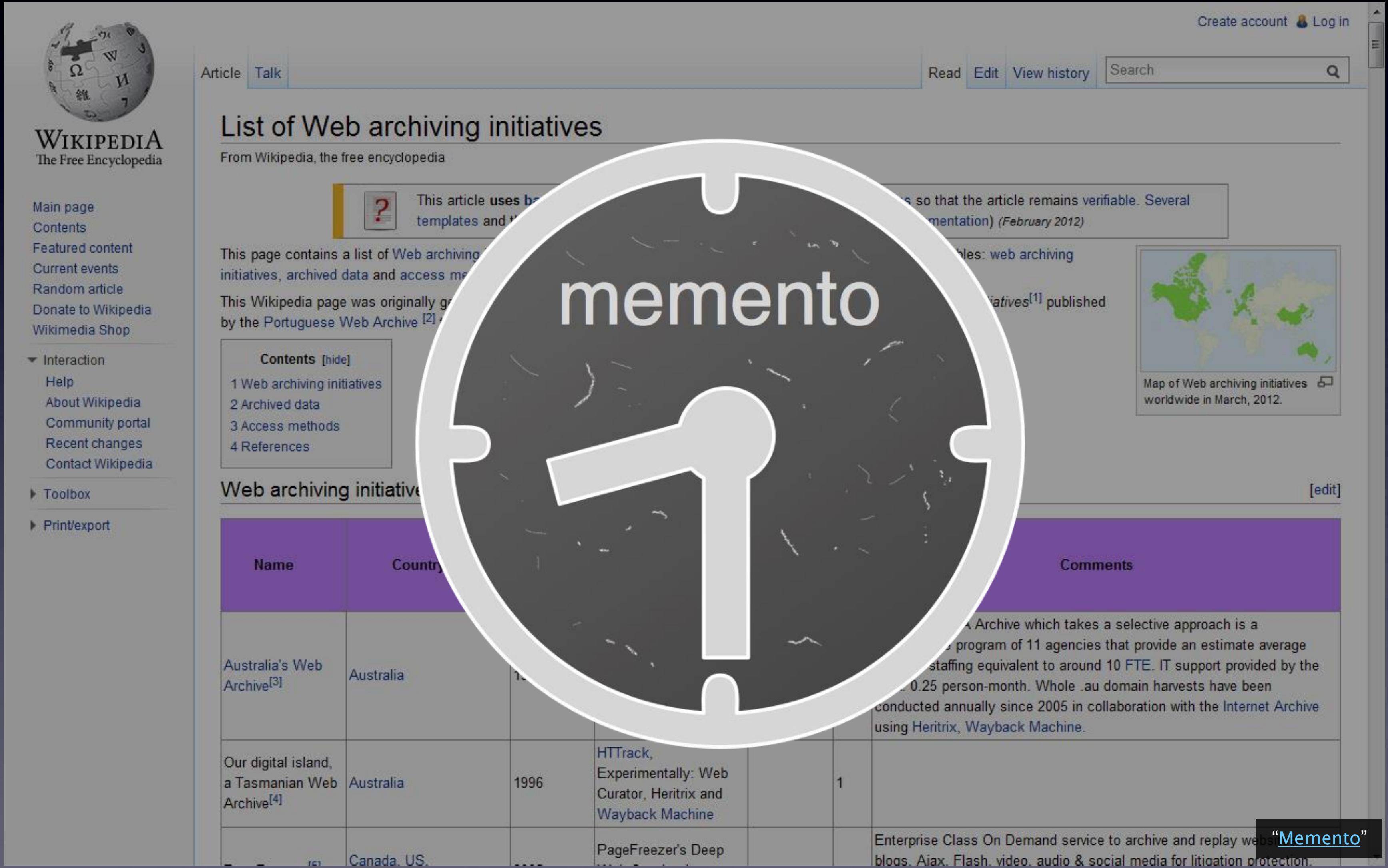
Web archiving initiatives

[edit]

Name	Country	Creation Year	Technologies	Number of Employees		Comments
				Full-time	Part-time	
Australia's Web Archive[3]	Australia	1996	PANDORA Digital Archiving System (PANDAS), NLA Trove, HTTrack.	10	>5	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 support provided by the NLA: 0.25 person-month. Whole .au domain harvests have been conducted annually since 2005 in collaboration with the Internet Archive using Heritrix, Wayback Machine.
Our digital island, a Tasmanian Web Archive[4]	Australia	1996	HTTrack, Experimentally: Web Curator, Heritrix and Wayback Machine		1	
[5]	Canada, US		PageFreezer's Deep			Enterprise Cla… blogs, Ajax, Flash, video, audio & social media for litigation protection.

Wikipedia: "List of Web archiving initiatives"

discovery: Memento



List of Web archiving initiatives

From Wikipedia, the free encyclopedia

This article uses ba… templates and t… …s so that the article remains verifiable. Several …mentation) (February 2012)

This page contains a list of Web archiving [illegible] initiatives, archived data and access me [illegible] bles: web archiving

This Wikipedia page was originally g… …iatives[1] published by the Portuguese Web Archive [2] …

Map of Web archiving initiatives worldwide in March, 2012.

Contents [hide]

1 Web archiving initiatives
2 Archived data
3 Access methods
4 References

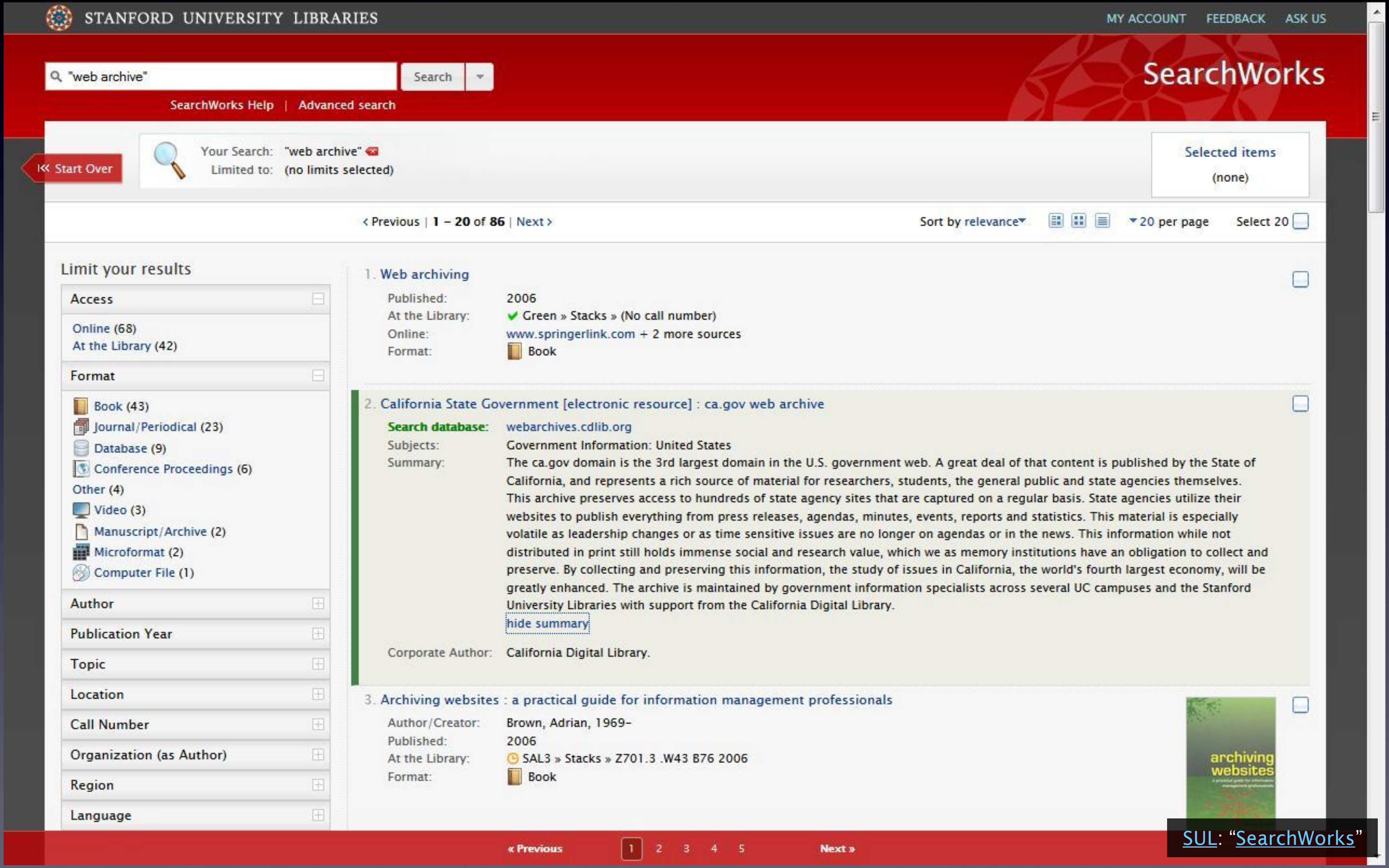
Web archiving initiativ…

[edit]

Name	Country					Comments
Australia's Web Archive[3]	Australia	1…				…A Archive which takes a selective approach is a … program of 11 agencies that provide an estimate average … staffing equivalent to around 10 FTE. IT support provided by the … 0.25 person-month. Whole .au domain harvests have been conducted annually since 2005 in collaboration with the Internet Archive using Heritrix, Wayback Machine.
Our digital island, a Tasmanian Web Archive[4]	Australia	1996	HTTrack, Experimentally: Web Curator, Heritrix and Wayback Machine		1	
[5]	Canada, US,		PageFreezer's Deep			Enterprise Class On Demand service to archive and replay websi… blogs, Ajax, Flash, video, audio & social media for litigation protection.

“Memento”

discovery: SearchWorks



Your Search: "web archive"
Limited to: (no limits selected)

Selected items
(none)

‹ Previous | **1 – 20** of **86** | Next ›   Sort by relevance   20 per page   Select 20 □

Limit your results

Access
- Online (68)
- At the Library (42)

Format
- Book (43)
- Journal/Periodical (23)
- Database (9)
- Conference Proceedings (6)
- Other (4)
- Video (3)
- Manuscript/Archive (2)
- Microformat (2)
- Computer File (1)

Author
Publication Year
Topic
Location
Call Number
Organization (as Author)
Region
Language

1. Web archiving □

Published:	2006
At the Library:	✓ Green » Stacks » (No call number)
Online:	www.springerlink.com + 2 more sources
Format:	Book

2. California State Government [electronic resource] : ca.gov web archive □

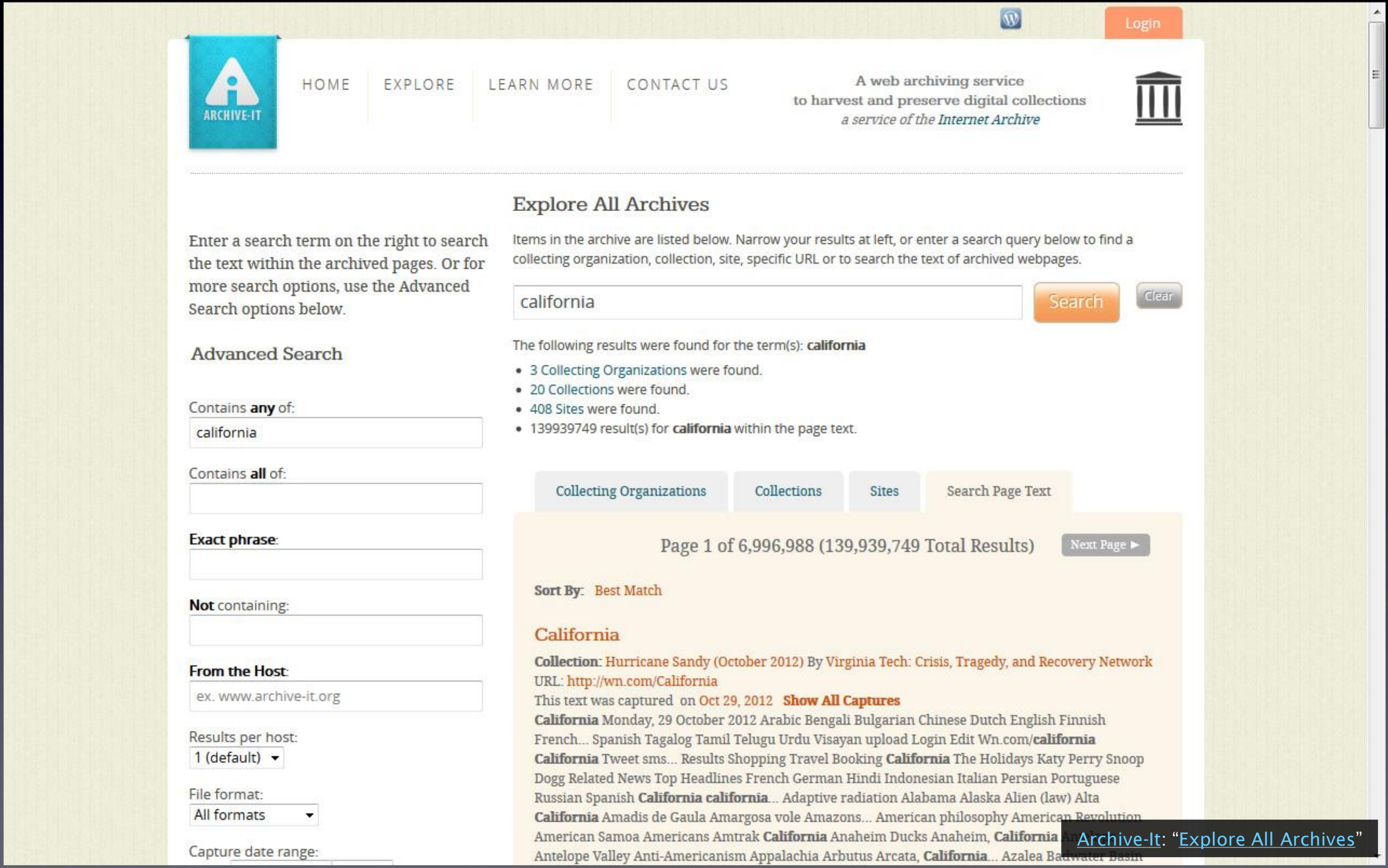
Search database:	webarchives.cdlib.org
Subjects:	Government Information: United States
Summary:	The ca.gov domain is the 3rd largest domain in the U.S. government web. A great deal of that content is published by the State of California, and represents a rich source of material for researchers, students, 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. hide summary
Corporate Author:	California Digital Library.

3. Archiving websites : a practical guide for information management professionals □

Author/Creator:	Brown, Adrian, 1969-
Published:	2006
At the Library:	SAL3 » Stacks » Z701.3 .W43 B76 2006
Format:	Book

SUL: "SearchWorks"

full-text search: Solr



A web archiving service
to harvest and preserve digital collections
a service of the Internet Archive

Enter a search term on the right to search the text within the archived pages. Or for more search options, use the Advanced Search options below.

Advanced Search

Contains **any** of: california

Contains **all** of:

Exact phrase:

Not containing:

From the Host: ex. www.archive-it.org

Results per host: 1 (default)

File format: All formats

Capture date range:

Explore All Archives

Items in the archive are listed below. Narrow your results at left, or enter a search query below to find a collecting organization, collection, site, specific URL or to search the text of archived webpages.

california

The following results were found for the term(s): **california**

- 3 Collecting Organizations were found.
- 20 Collections were found.
- 408 Sites were found.
- 139939749 result(s) for **california** within the page text.

Page 1 of 6,996,988 (139,939,749 Total Results)

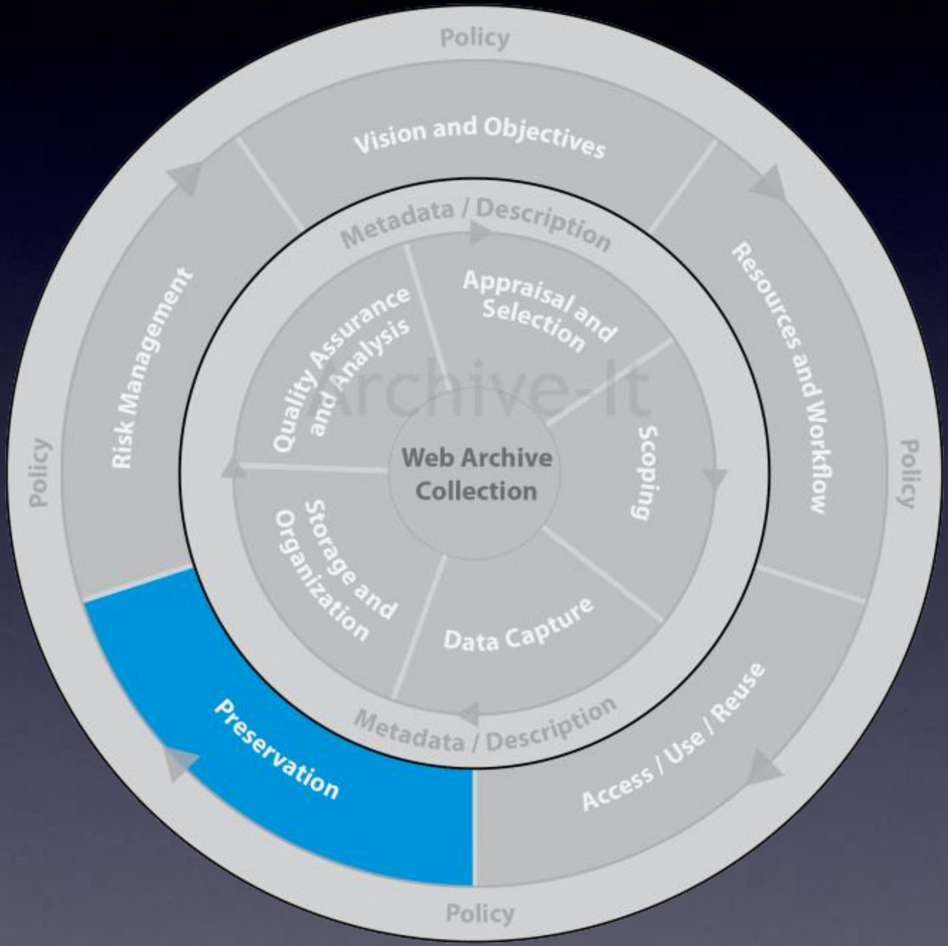
Sort By: Best Match

California

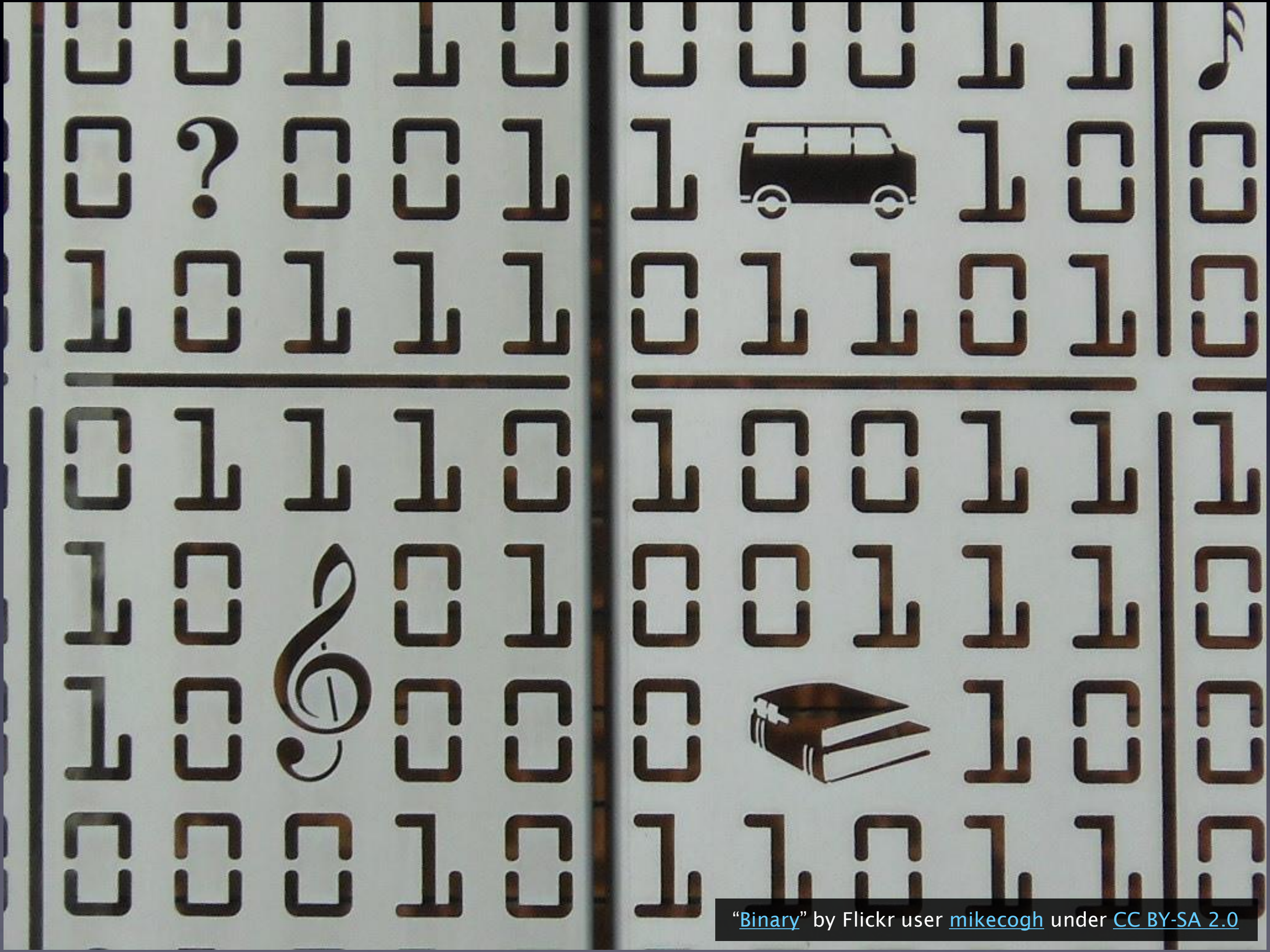
Collection: Hurricane Sandy (October 2012) By Virginia Tech: Crisis, Tragedy, and Recovery Network
URL: http://wn.com/California
This text was captured on Oct 29, 2012 **Show All Captures**
California Monday, 29 October 2012 Arabic Bengali Bulgarian Chinese Dutch English Finnish French... Spanish Tagalog Tamil Telugu Urdu Visayan upload Login Edit Wn.com/**california** **California** Tweet sms... Results Shopping Travel Booking **California** The Holidays Katy Perry Snoop Dogg Related News Top Headlines French German Hindi Indonesian Italian Persian Portuguese Russian Spanish **California california**... Adaptive radiation Alabama Alaska Alien (law) Alta **California** Amadis de Gaula Amargosa vole Amazons... American philosophy American Revolution American Samoa Americans Amtrak **California** Anaheim Ducks Anaheim, **California** Antelope Valley Anti-Americanism Appalachia Arbutus Arcata, **California**... Azalea Badwater Basin

Archive-It: “Explore All Archives”

Preservation



bit preservation

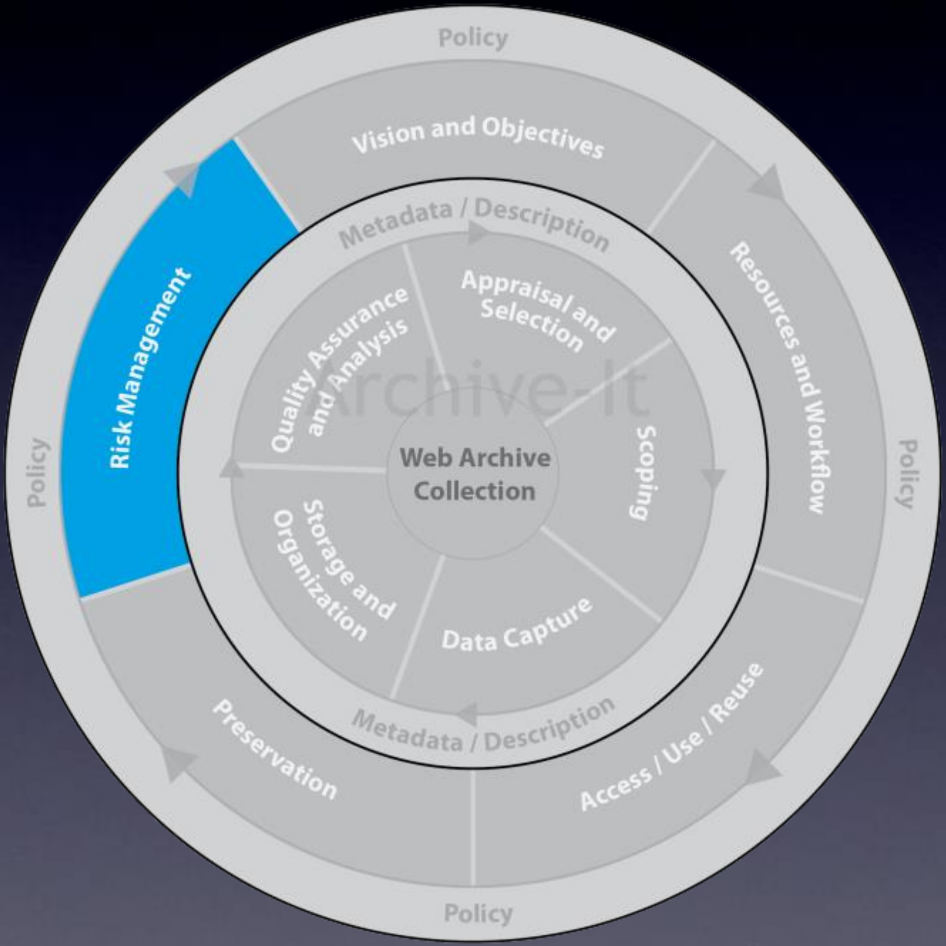


"Binary" by Flickr user mikecogh under CC BY-SA 2.0

preservation engineering

"Máquina de Rube Goldberg en la base del Alinghi" by Flickr user freshwater2006 under CC BY-NC 2.0

Risk Management

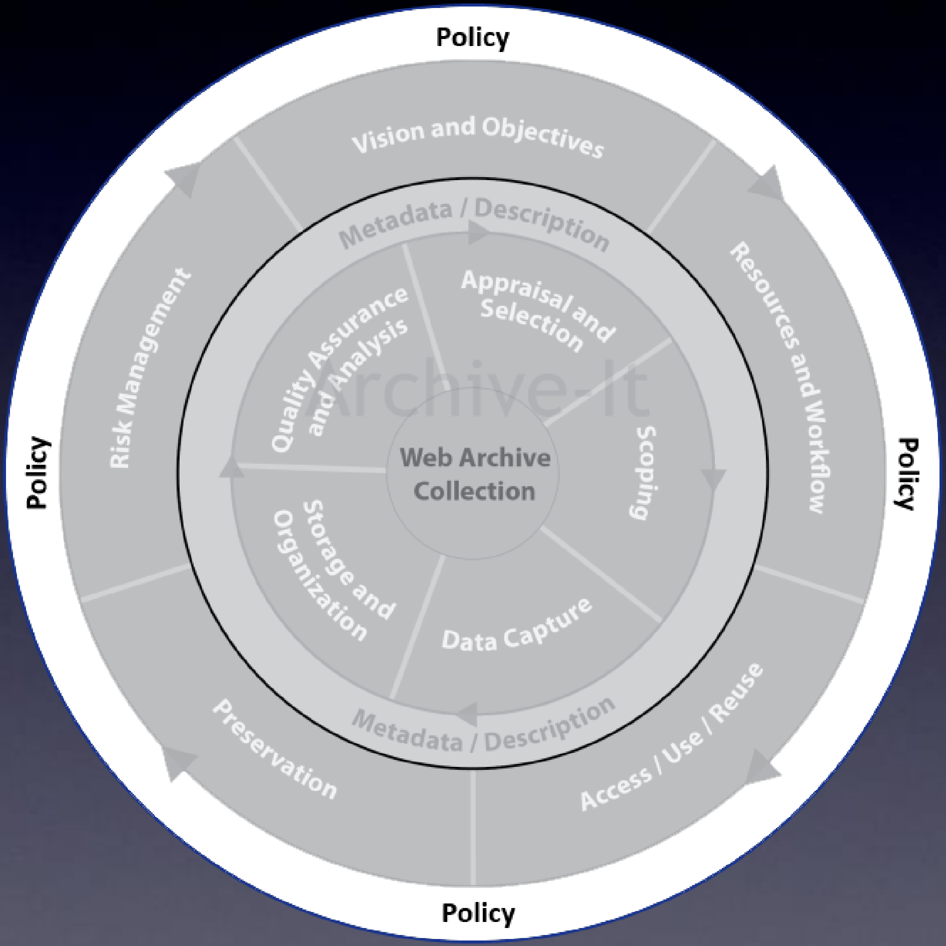


Risk Management

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

“Zombie Awareness - Extinguisher” by Flickr user Spiffy0777 under CC BY-NC-SA 2.0

Policy


Policy
Vision and Objectives
Resources and Workflow
Policy
Access / Use / Reuse
Policy
Preservation
Risk Management
Policy
Metadata / Description
Appraisal and Selection
Scoping
Data Capture
Storage and Organization
Quality Assurance and Analysis
Metadata / Description
Web Archive Collection
Archive-It

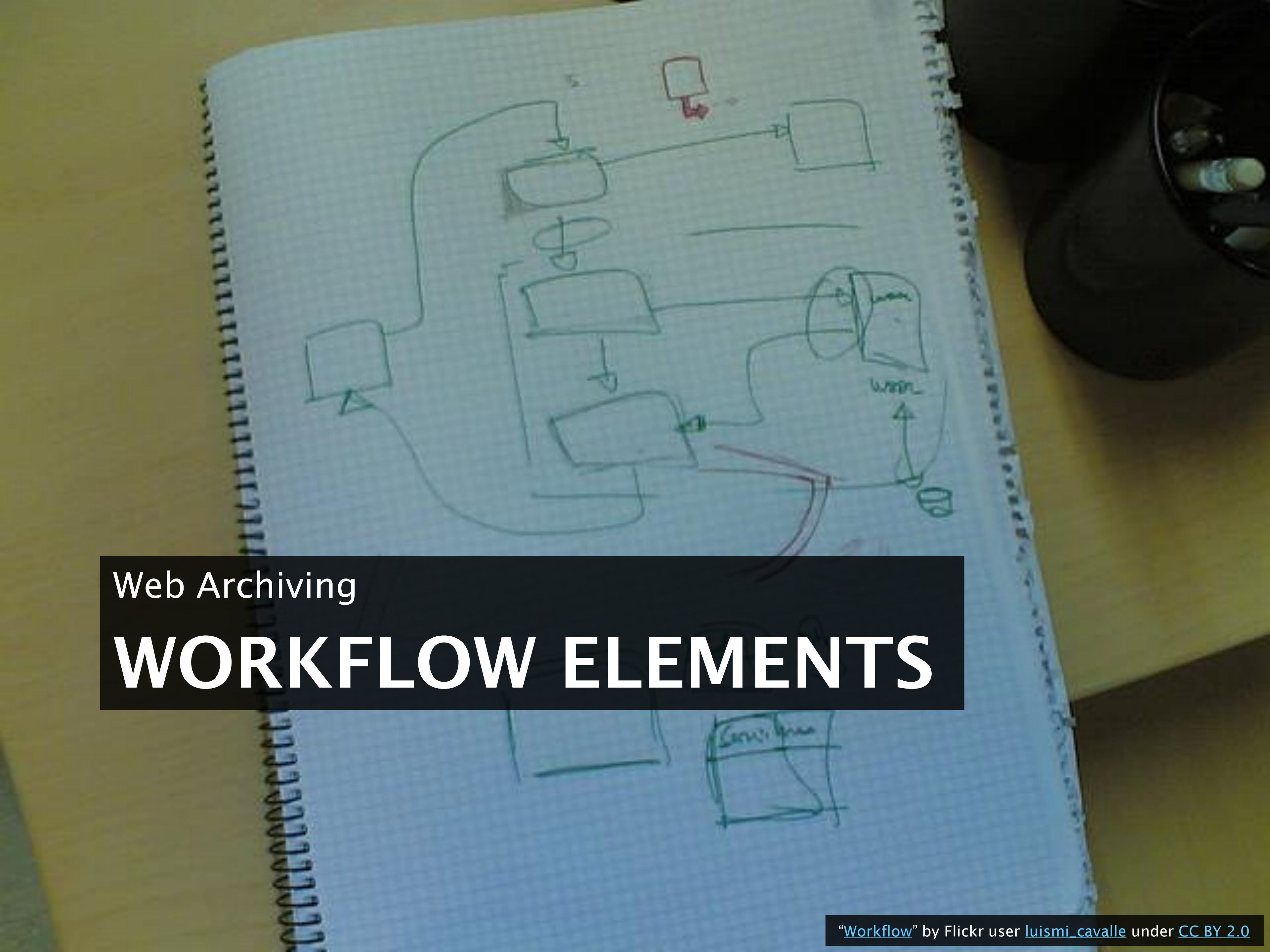
copyright

- § 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

“leaf-cutter ants” by Flickr user Vilseskogen under CC BY-NC-SA 2.0

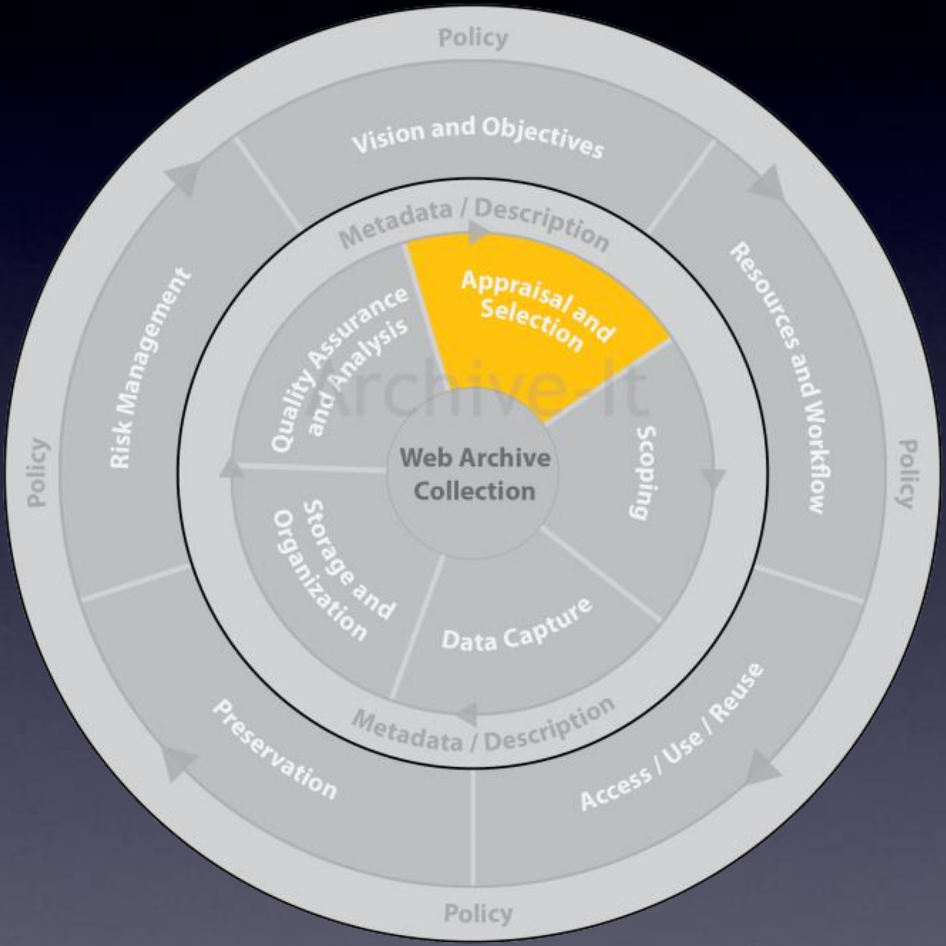


Web Archiving

WORKFLOW ELEMENTS

“Workflow” by Flickr user luismi_cavalle under CC BY 2.0

Appraisal and Selection

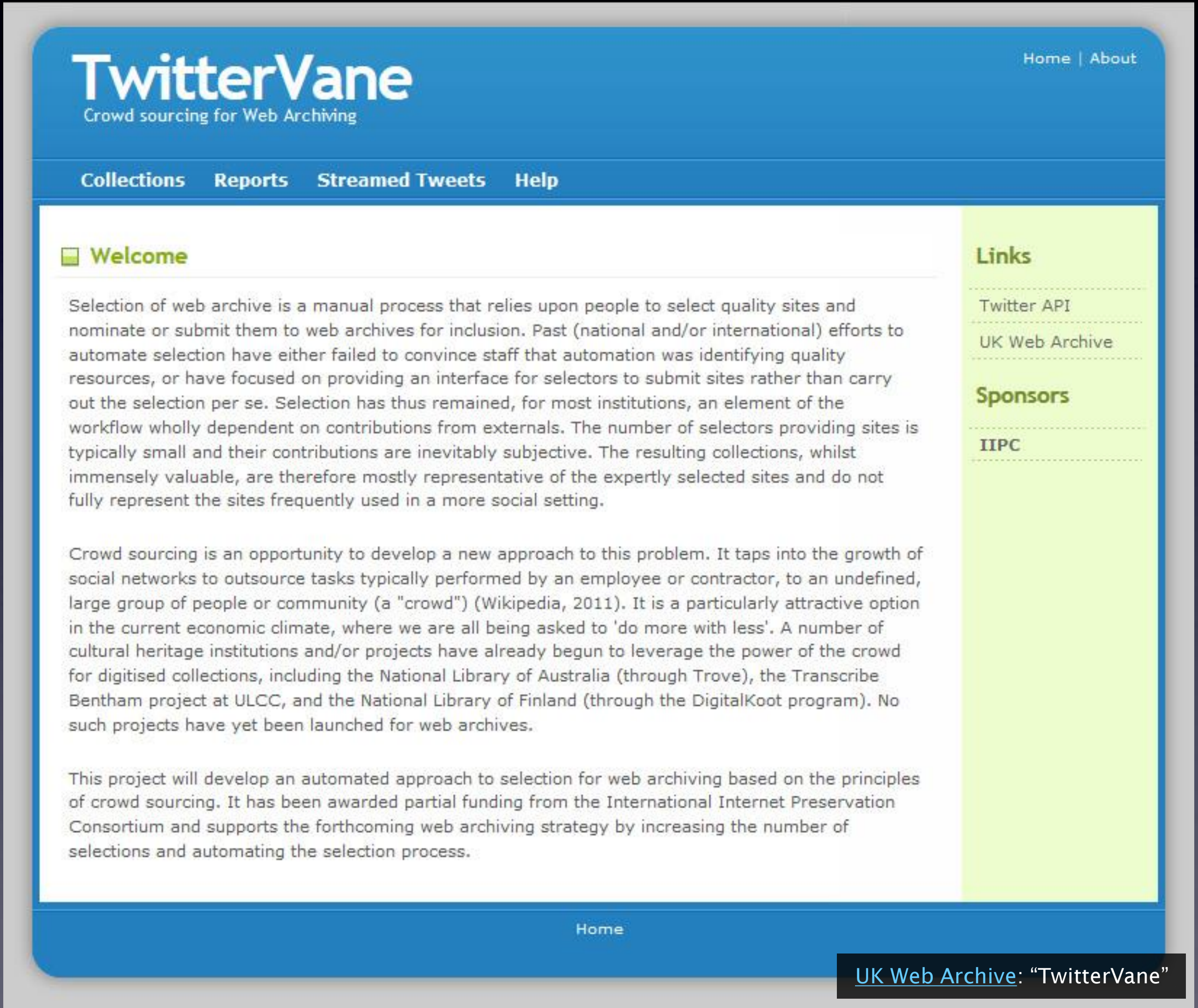


informing selection

- value
- risk
- size
- extent to which archived

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

TwitterVane



TwitterVane
Crowd sourcing for Web Archiving

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.

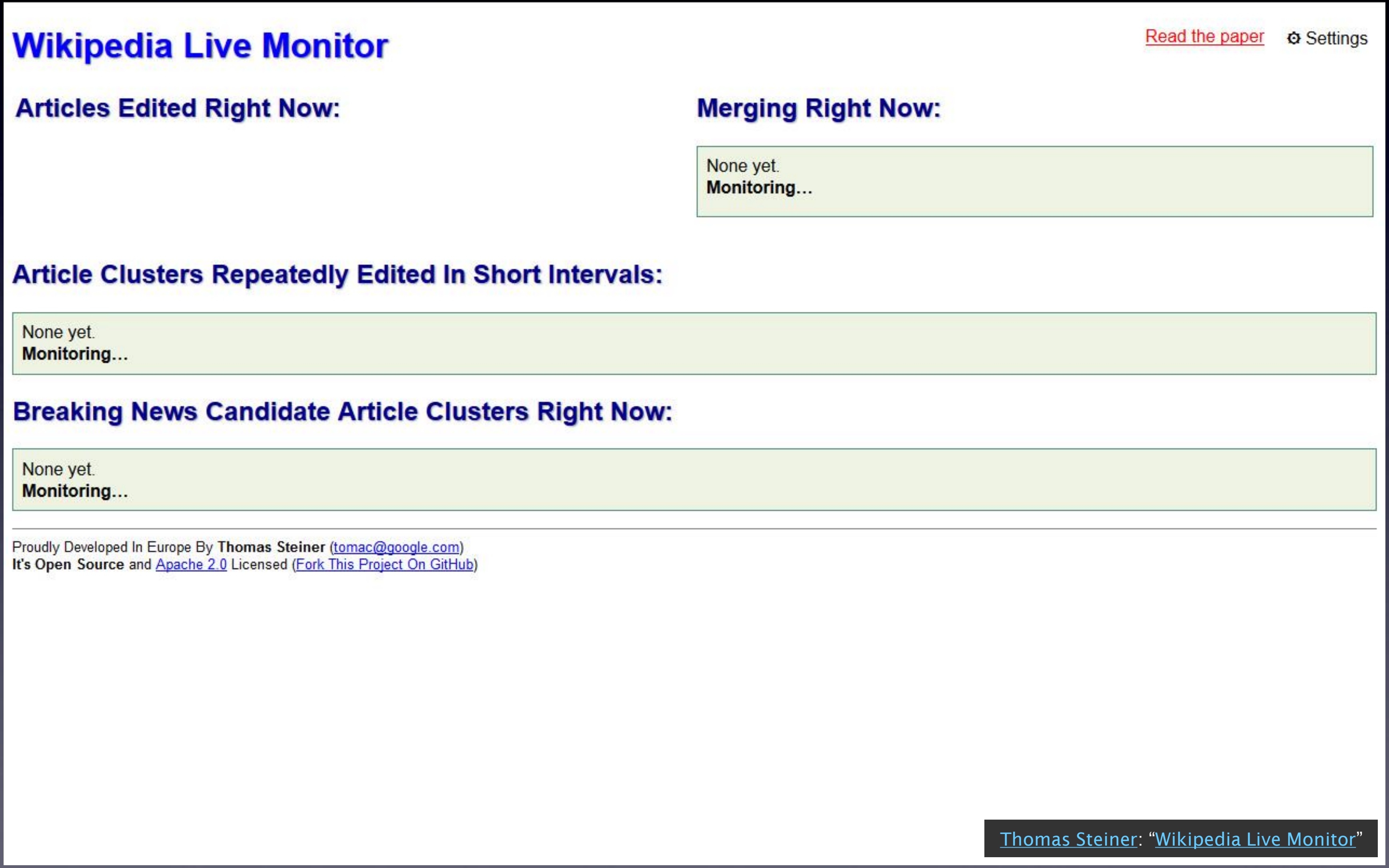
This project will develop an automated approach to selection for web archiving based on the principles of crowd sourcing. It has been awarded partial funding from the International Internet Preservation Consortium and supports the forthcoming web archiving strategy by increasing the number of selections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

Links
- Twitter API
- UK Web Archive

Sponsors
- IIPC

UK Web Archive: “TwitterVane”

Wikipedia Live Monitor



Wikipedia Live Monitor

Read the paper ⚙ Settings

Articles Edited Right Now:

Merging Right Now:

None yet.
Monitoring…

Article Clusters Repeatedly Edited In Short Intervals:

None yet.
Monitoring…

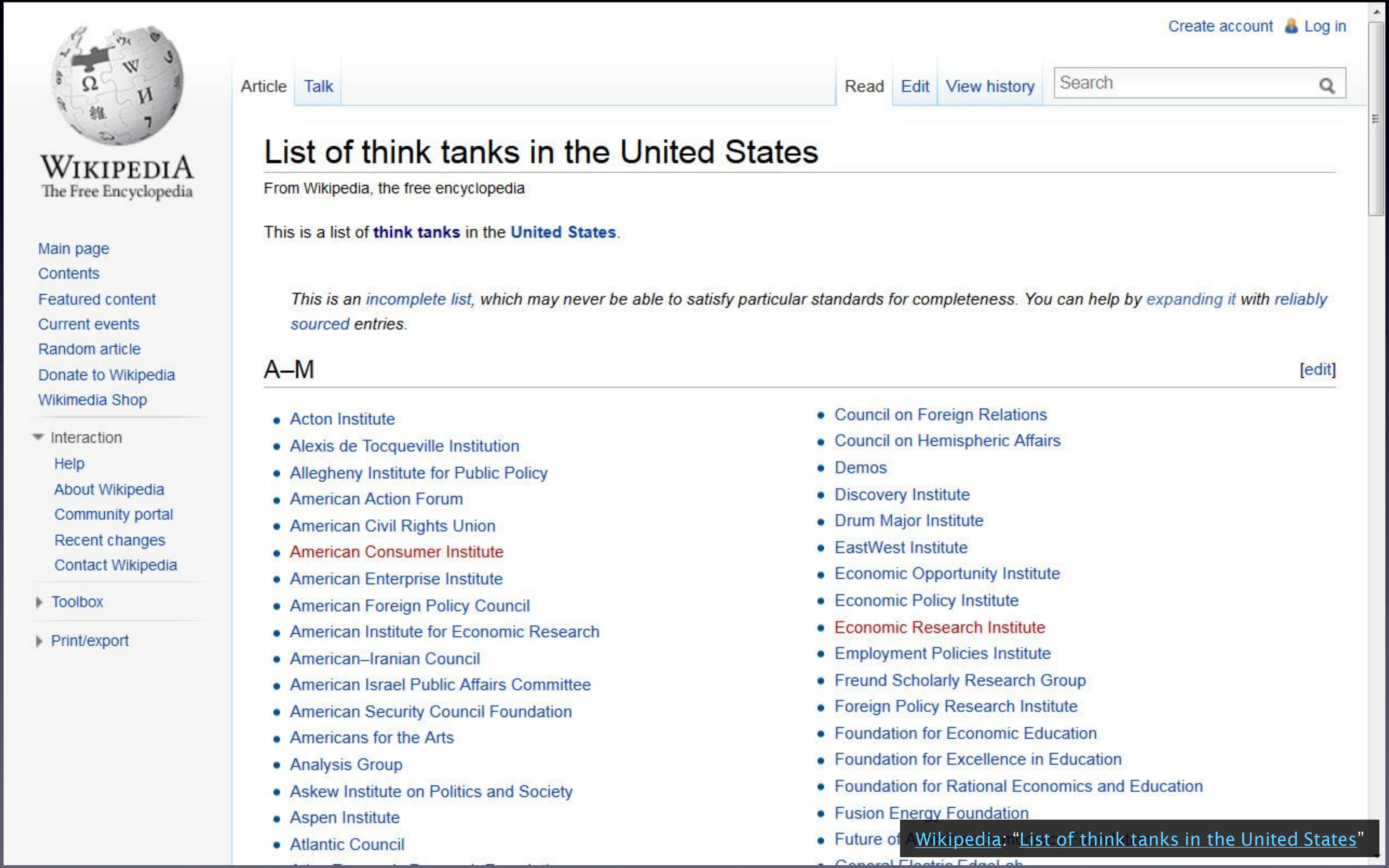
Breaking News Candidate Article Clusters Right Now:

None yet.
Monitoring…

Proudly Developed In Europe By **Thomas Steiner** (tomac@google.com)
It's Open Source and Apache 2.0 Licensed (Fork This Project On GitHub)

Thomas Steiner: “Wikipedia Live Monitor”

Wikipedia articles



List of think tanks in the United States

From Wikipedia, the free encyclopedia

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.

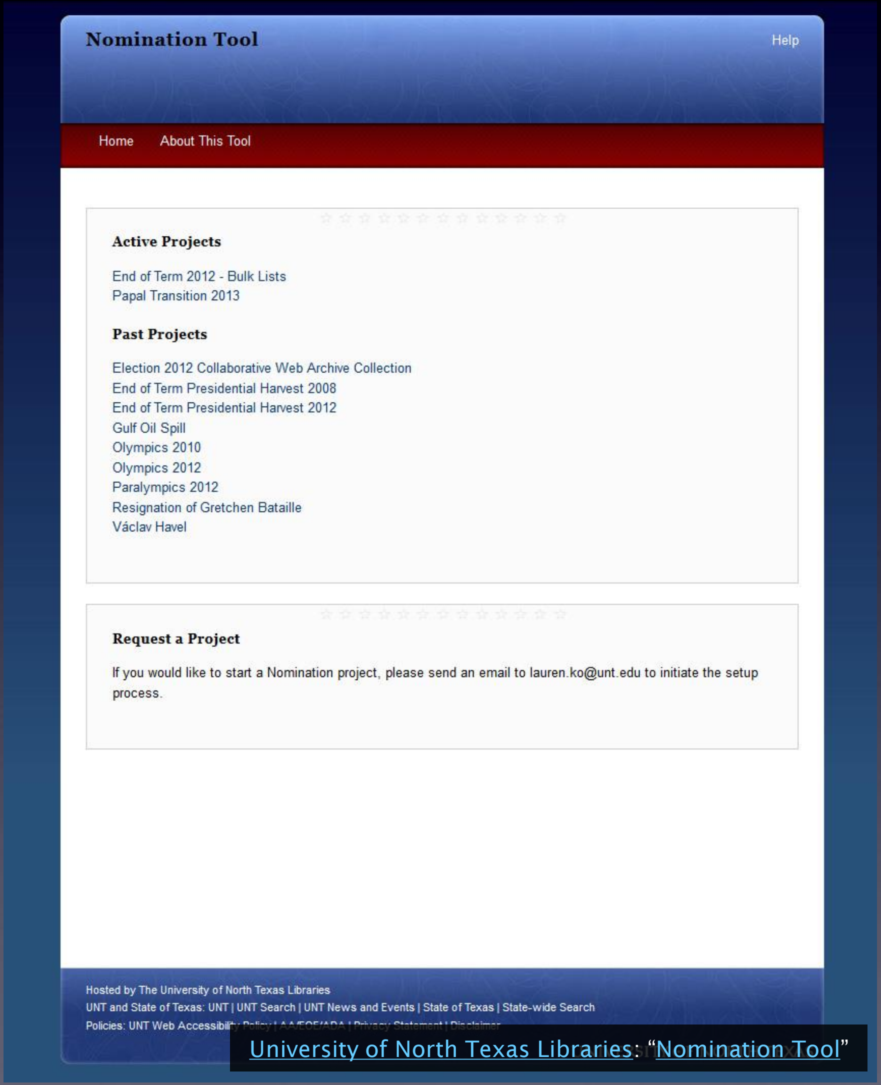
A–M

[edit]

- 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
- Council on Foreign Relations
- Council on Hemispheric Affairs
- Demos
- Discovery Institute
- Drum Major Institute
- EastWest Institute
- Economic Opportunity Institute
- Economic Policy Institute
- Economic Research Institute
- Employment Policies Institute
- Freund Scholarly Research Group
- Foreign Policy Research Institute
- Foundation for Economic Education
- Foundation for Excellence in Education
- Foundation for Rational Economics and Education
- Fusion Energy Foundation
- Future of A

Wikipedia: "List of think tanks in the United States"

UNT Nomination Tool



Nomination Tool

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
Resignation of Gretchen Bataille
Václav Havel

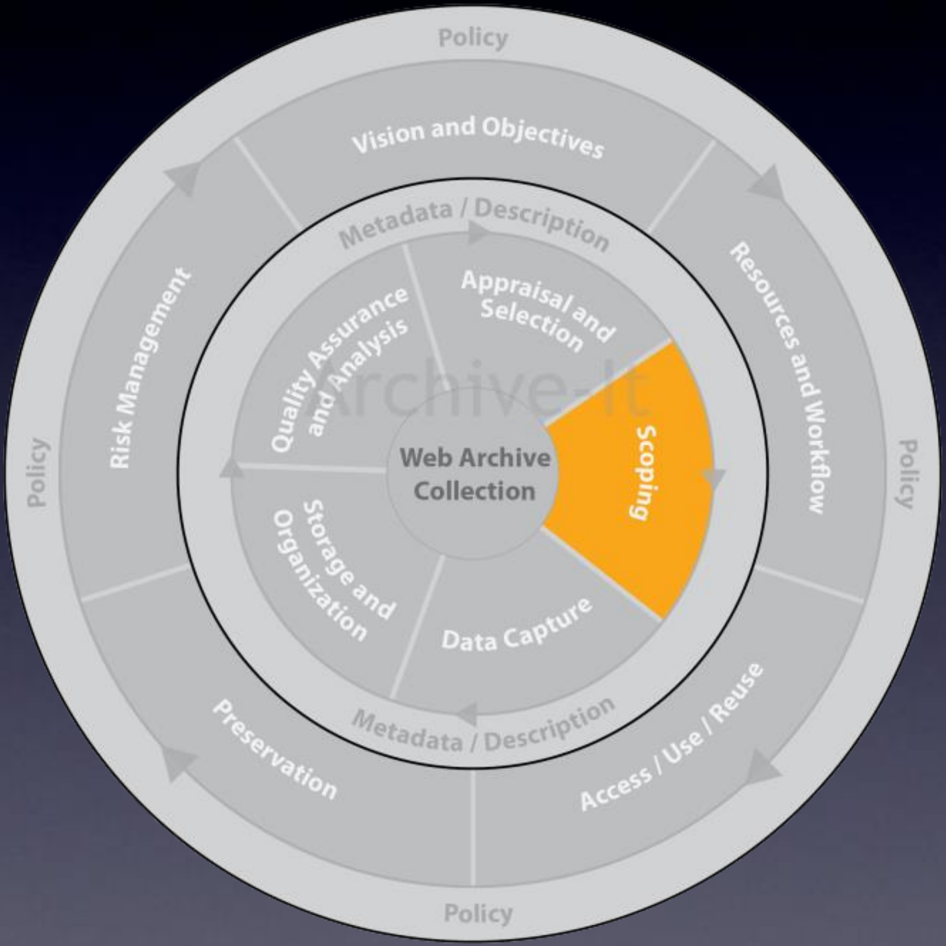
Request a Project

If you would like to start a Nomination project, please send an email to lauren.ko@unt.edu to initiate the setup process.

Hosted by The University of North Texas Libraries
UNT and State of Texas: UNT | UNT Search | UNT News and Events | State of Texas | State-wide Search
Policies: UNT Web Accessibility Policy | AA/EOE/ADA | Privacy Statement | Disclaimer

University of North Texas Libraries: "Nomination Tool"

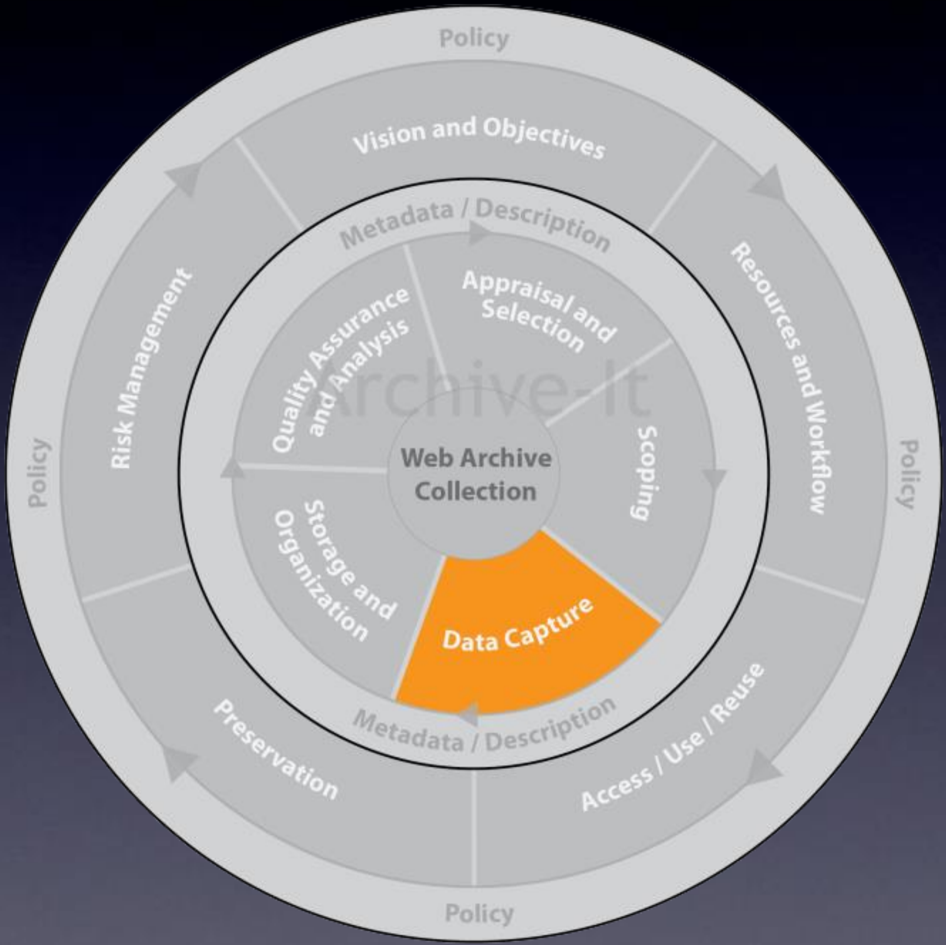
Scoping


Policy
Vision and Objectives
Resources and Workflow
Policy
Access / Use / Reuse
Policy
Preservation
Policy
Risk Management
Metadata / Description
Appraisal and Selection
Scoping
Data Capture
Metadata / Description
Storage and Organization
Quality Assurance and Analysis
Web Archive Collection
Archive-It

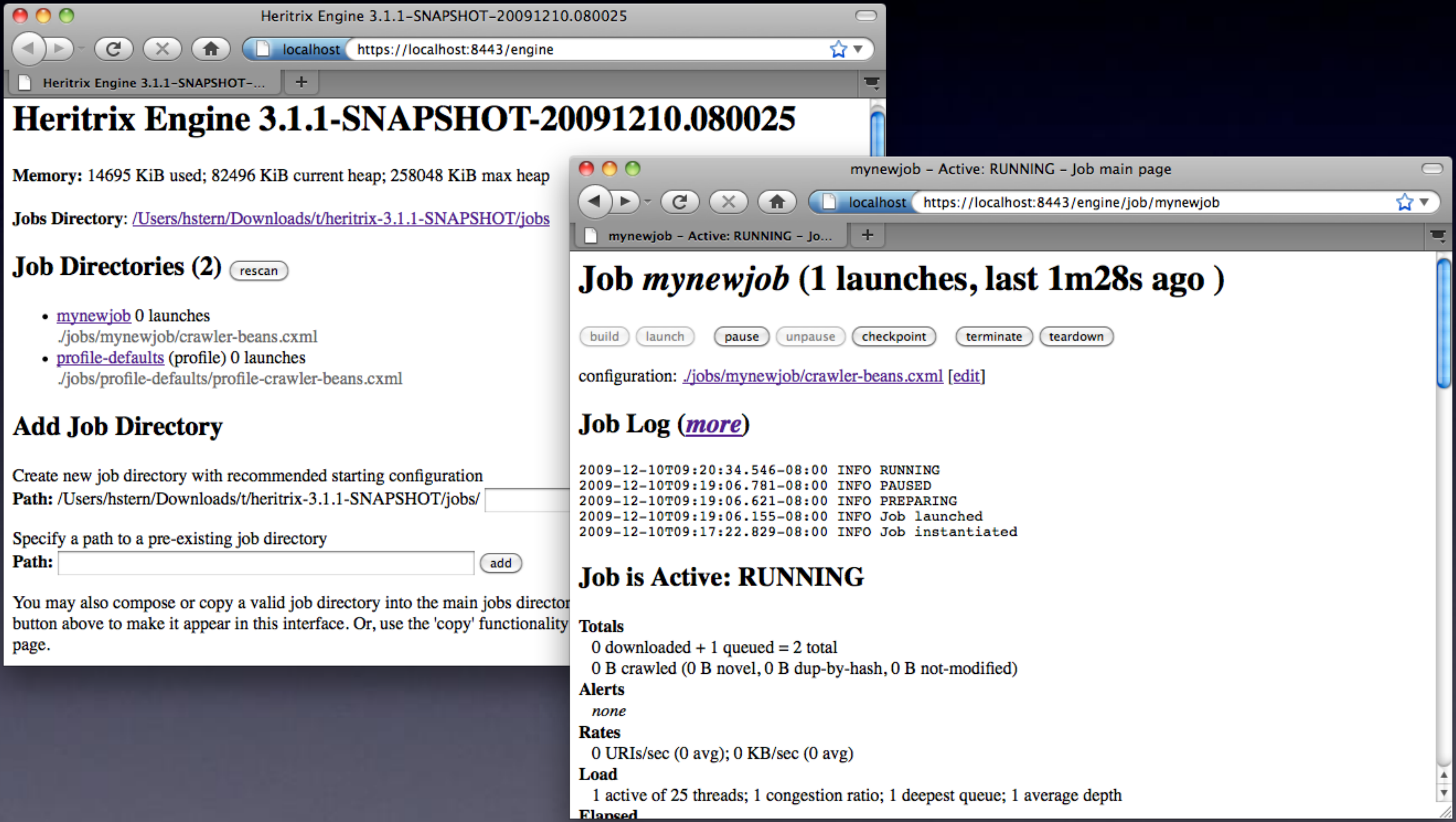
the purpose of scoping

“More god?” by Flickr user one two one three under CC BY-NC-SA 2.0

Data Capture



Heritrix


Heritrix Engine 3.1.1-SNAPSHOT-20091210.080025
localhost
https://localhost:8443/engine
Heritrix Engine 3.1.1-SNAPSHOT-...
rescan
add
mynewjob - Active: RUNNING - Job main page
localhost
https://localhost:8443/engine/job/mynewjob
mynewjob - Active: RUNNING - Jo...
build
launch
pause
unpause
checkpoint
terminate
teardown

Heritrix Engine 3.1.1-SNAPSHOT-20091210.080025

Memory: 14695 KiB used; 82496 KiB current heap; 258048 KiB max heap

Jobs Directory: /Users/hstern/Downloads/t/heritrix-3.1.1-SNAPSHOT/jobs

Job Directories (2)

- mynewjob 0 launches
  ./jobs/mynewjob/crawler-beans.cxml
- profile-defaults (profile) 0 launches
  ./jobs/profile-defaults/profile-crawler-beans.cxml

Add Job Directory

Create new job directory with recommended starting configuration
Path: /Users/hstern/Downloads/t/heritrix-3.1.1-SNAPSHOT/jobs/

Specify a path to a pre-existing job directory
Path:

You may also compose or copy a valid job directory into the main jobs director
button above to make it appear in this interface. Or, use the 'copy' functionality
page.

Job *mynewjob* (1 launches, last 1m28s ago)

configuration: ./jobs/mynewjob/crawler-beans.cxml [edit]

Job Log (*more*)

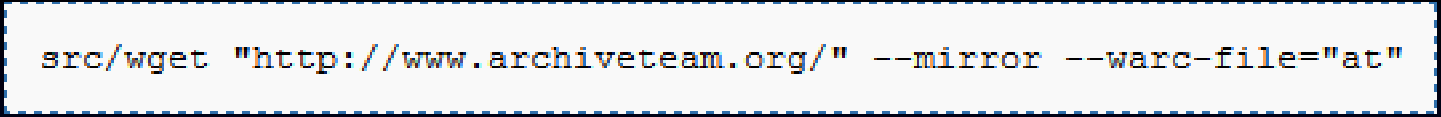
```
2009-12-10T09:20:34.546-08:00 INFO RUNNING
2009-12-10T09:19:06.781-08:00 INFO PAUSED
2009-12-10T09:19:06.621-08:00 INFO PREPARING
2009-12-10T09:19:06.155-08:00 INFO Job launched
2009-12-10T09:17:22.829-08:00 INFO Job instantiated
```

Job is Active: RUNNING

Totals
0 downloaded + 1 queued = 2 total
0 B crawled (0 B novel, 0 B dup-by-hash, 0 B not-modified)
Alerts
none
Rates
0 URIs/sec (0 avg); 0 KB/sec (0 avg)
Load
1 active of 25 threads; 1 congestion ratio; 1 deepest queue; 1 average depth
Elapsed

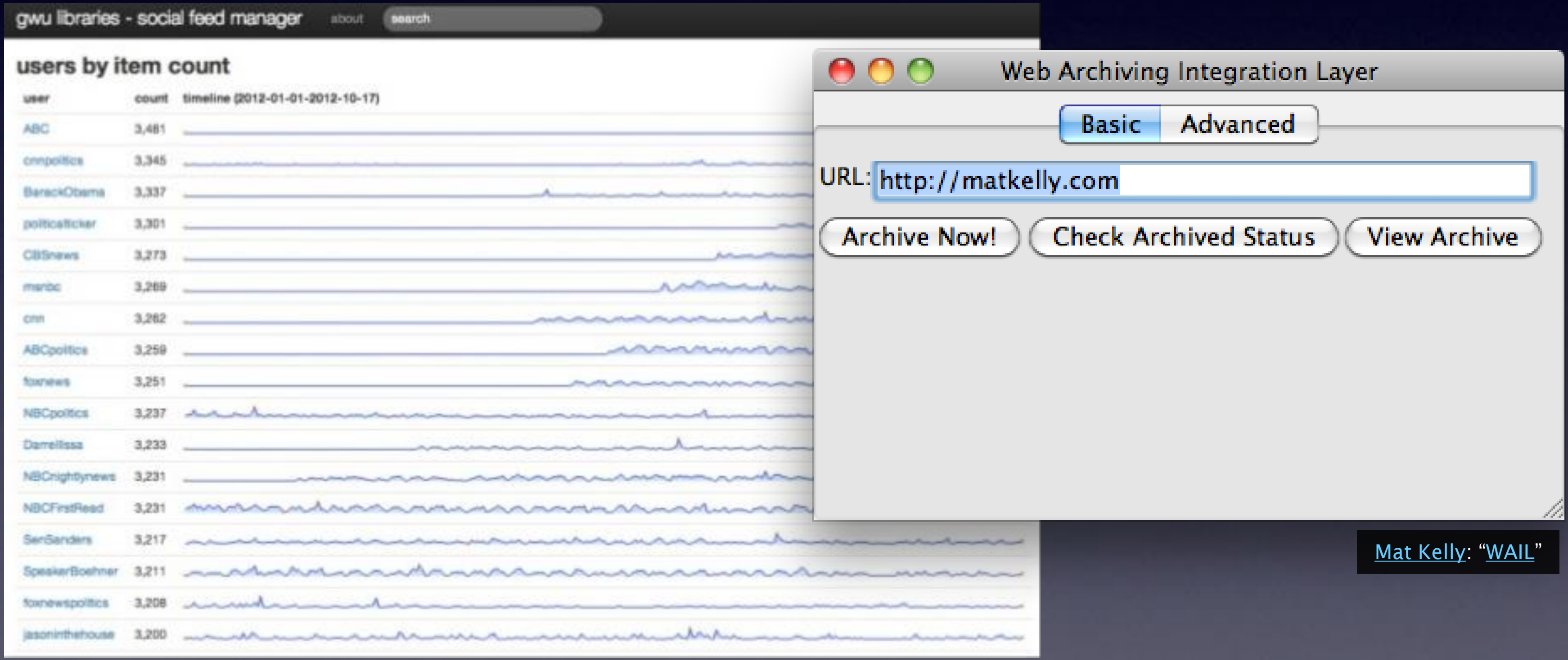
Internet Archive: “A Quick Guide to Running Your First Crawl Job”

other data capture tools



```
src/wget "http://www.archiveteam.org/" --mirror --warc-file="at"
```

Archive Team: “Wget with WARC output”



users by item count

user	count	timeline (2012-01-01-2012-10-17)
ABC	3,481	
cnnpolitics	3,345	
BarackObama	3,337	
politicalticker	3,301	
CBSnews	3,273	
msnbc	3,269	
cnn	3,262	
ABCpolitics	3,259	
foxnews	3,251	
NBCpolitics	3,237	
Darrellissa	3,233	
NBCnightlynews	3,231	
NBCFirstRead	3,231	
SenSanders	3,217	
SpeakerBoehner	3,211	
foxnewspolitics	3,208	
jasoninthehouse	3,200	

URL: http://matkelly.com

Mat Kelly: “WAIL”

Dan Chudnov and Laura Wrubel: “social feed manager”

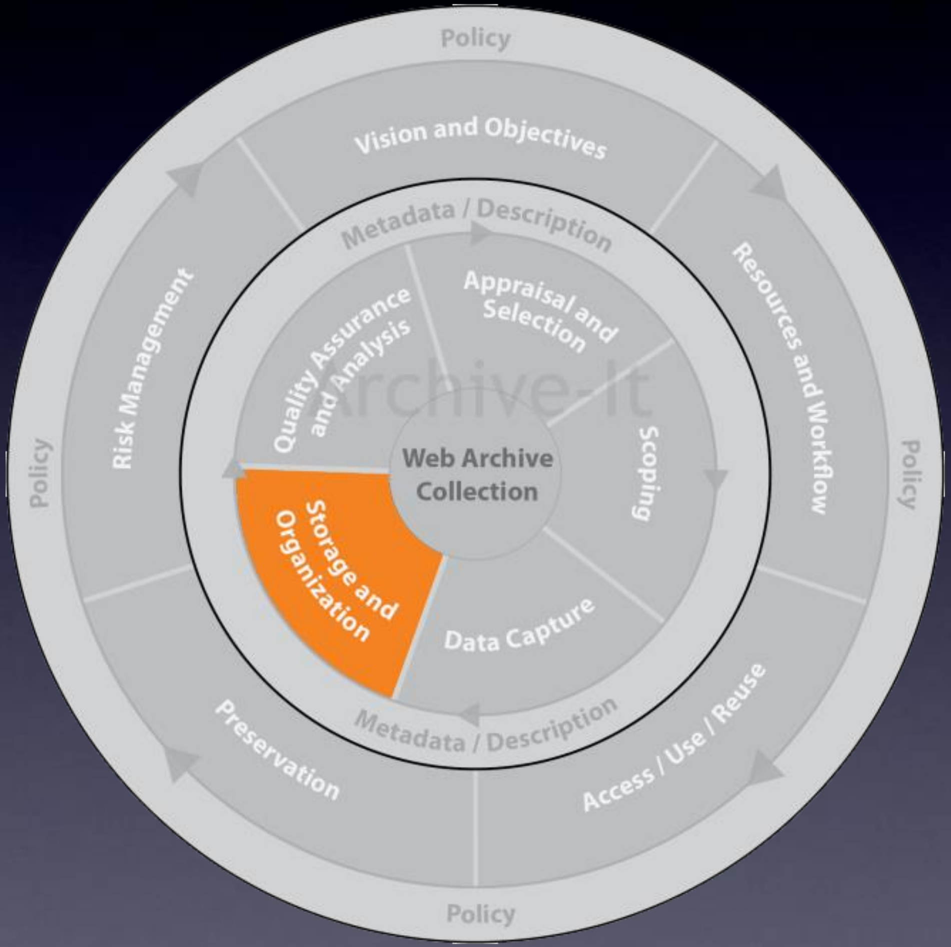
the elusive web

"Light Writing - Spider Web" by Flickr user forcefeed:swede under CC BY-ND 2.0

scale

“chutes and ladders” by Flickr user reallyboring under CC BY-NC-SA 2.0

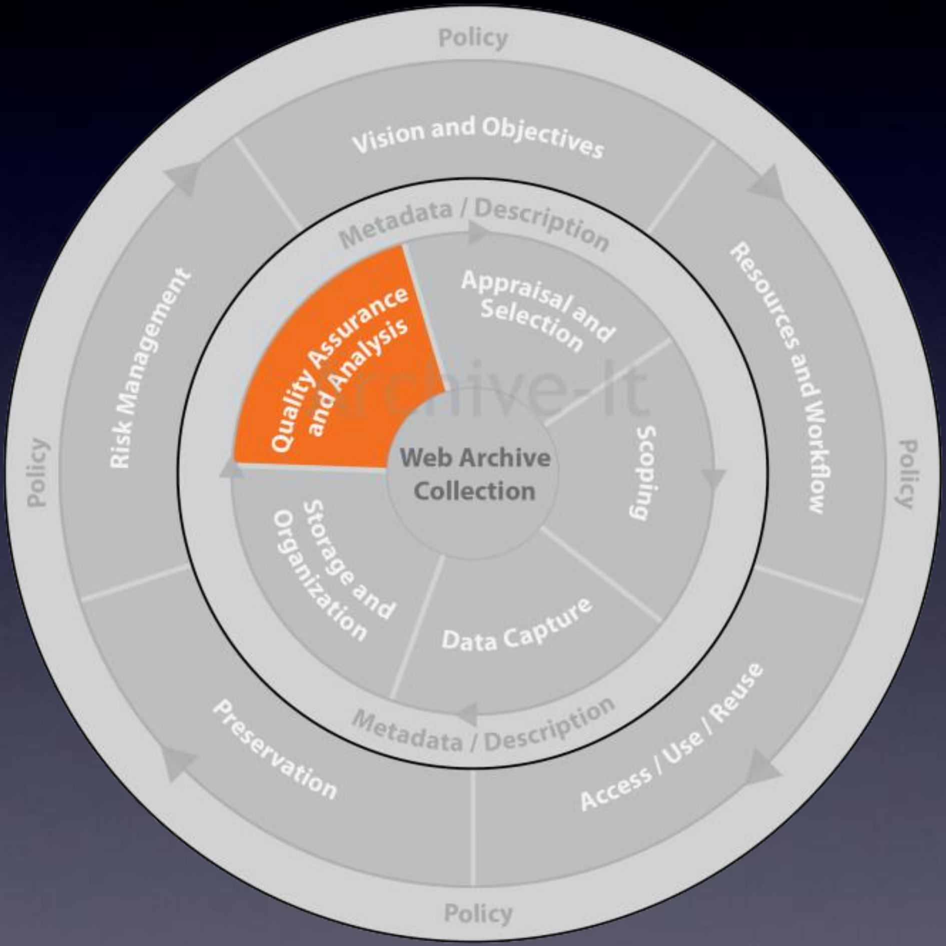
Storage and Organization



packages and their contents

“lots and lots and lots of boxes” by Flickr user Toastwife under CC BY-NC-SA 2.0

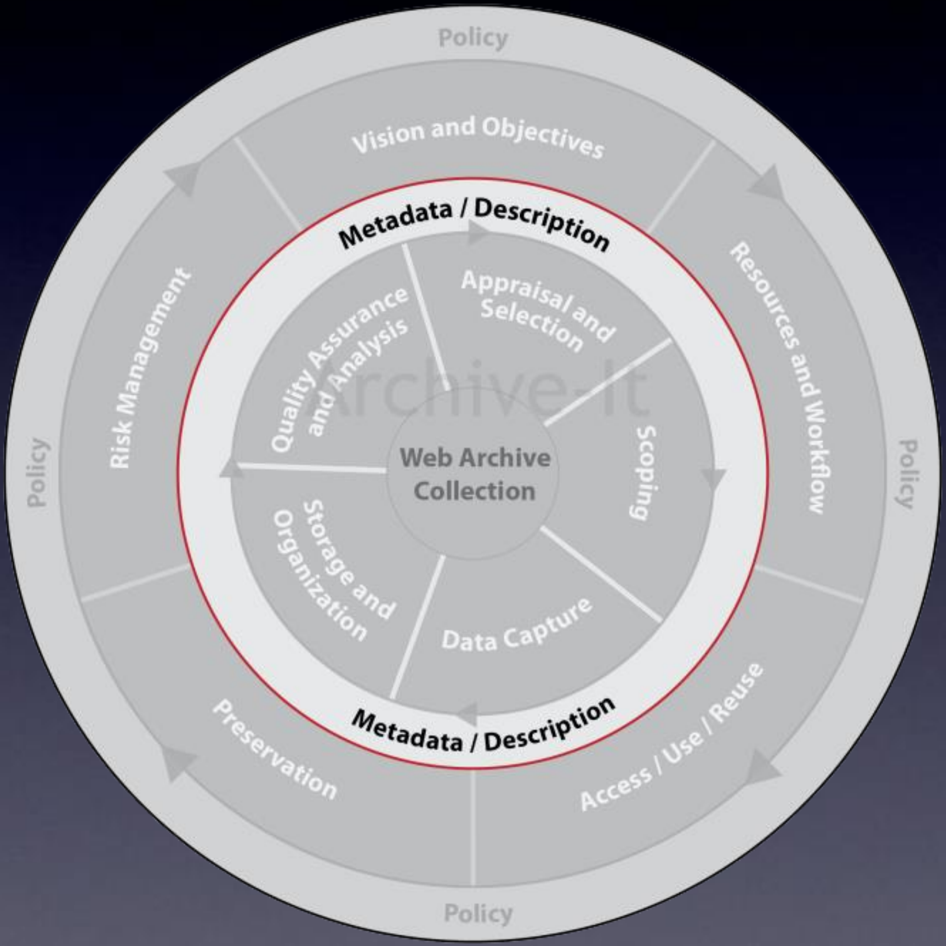
Quality Assurance and Analysis



QA before, after, during

"Check" by Flickr user ex.libris under CC BY-NC-ND 2.0

Metadata / Description



Metadata / Description

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

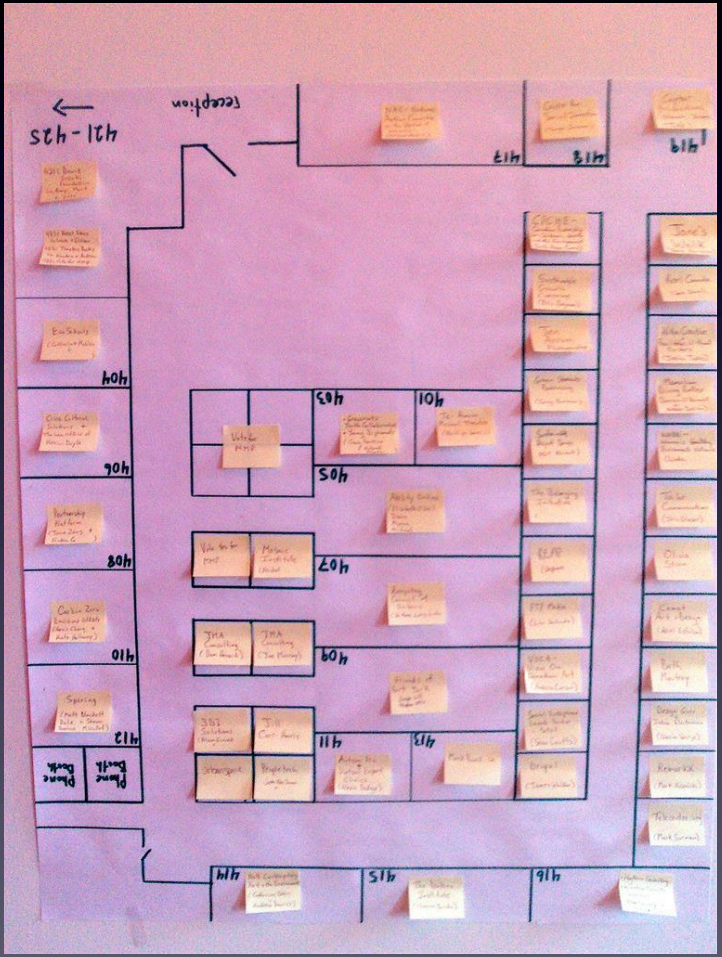
Considerations

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

“Geckos” by Flickr user smashz under CC BY-NC-ND 2.0

community engagement

LELAND STANFORD JUNIOR UNIVERSITY

DIE LUFT DER FREIHEIT WEHT

1891

JCDL

NDSA

Web Archiving

SAA

Roundtable

iP

IIPC

netpreserve.org

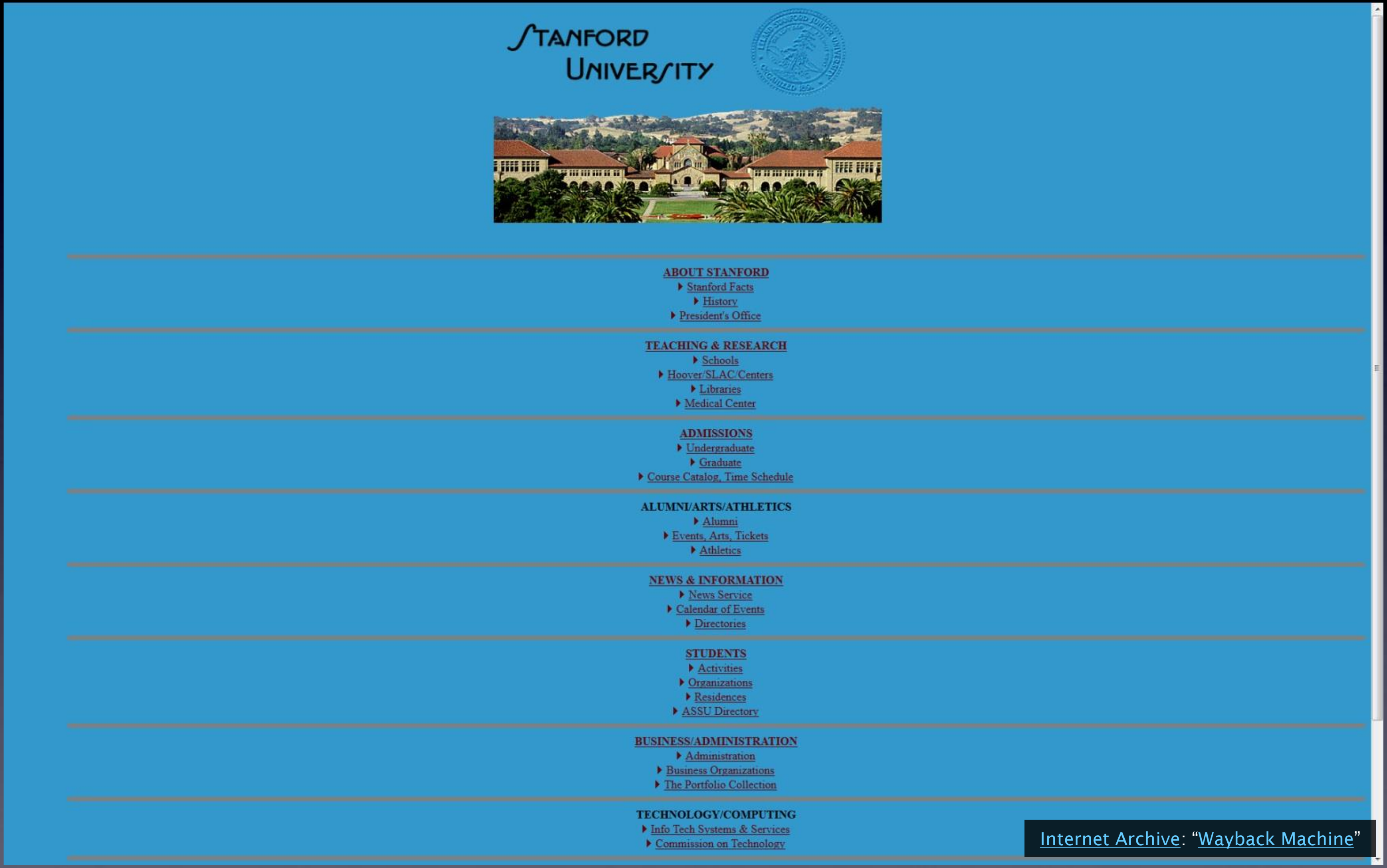
INTERNATIONAL
INTERNET
PRESERVATION
CONSORTIUM

AoIR

ASSOCIATION
OF
INTERNET
RESEARCHERS

sa

the web changes



STANFORD UNIVERSITY

ABOUT STANFORD
▸ Stanford Facts
▸ History
▸ President's Office

TEACHING & RESEARCH
▸ Schools
▸ Hoover/SLAC/Centers
▸ Libraries
▸ Medical Center

ADMISSIONS
▸ Undergraduate
▸ Graduate
▸ Course Catalog, Time Schedule

ALUMNI/ARTS/ATHLETICS
▸ Alumni
▸ Events, Arts, Tickets
▸ Athletics

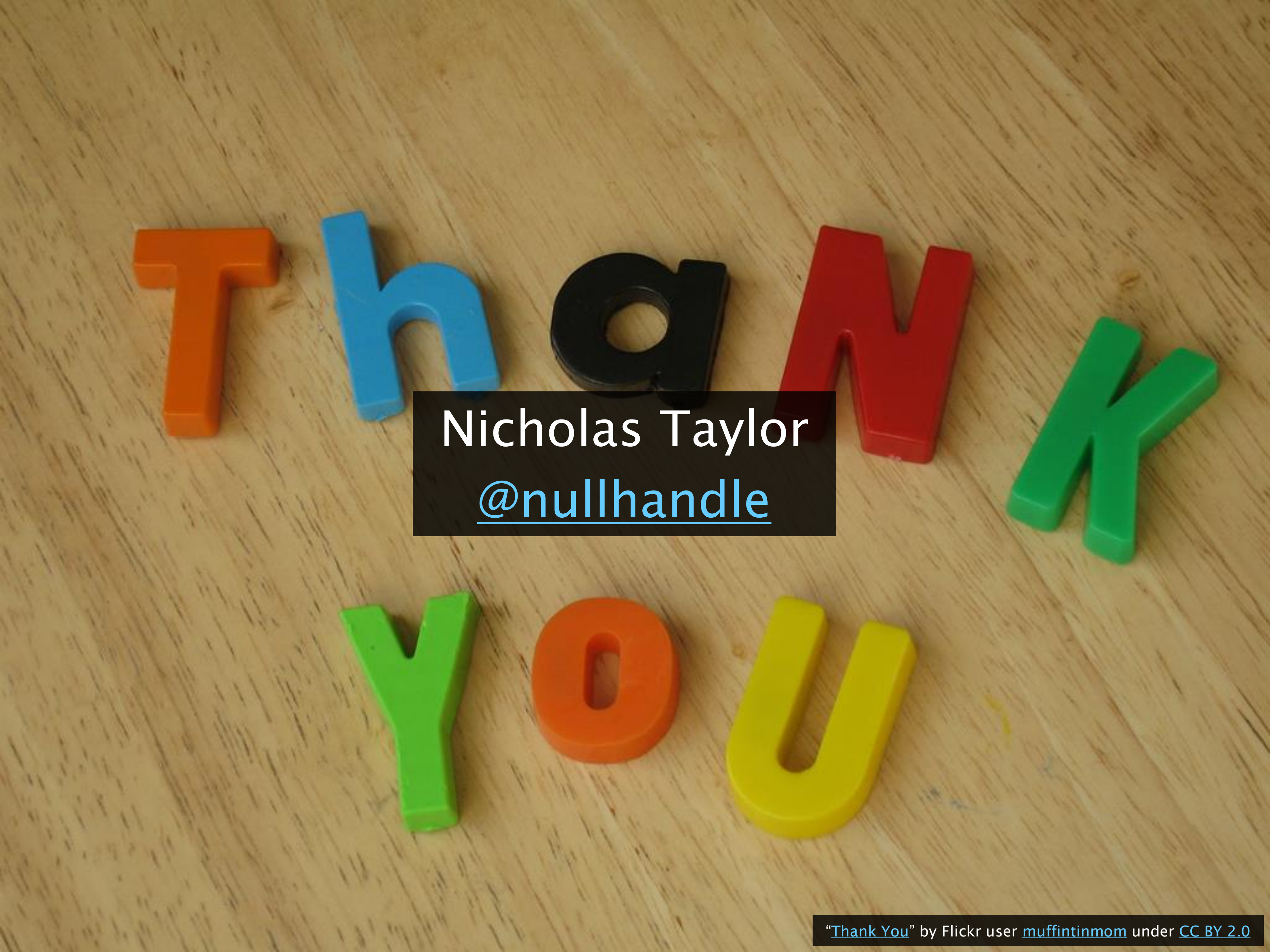
NEWS & INFORMATION
▸ News Service
▸ Calendar of Events
▸ Directories

STUDENTS
▸ Activities
▸ Organizations
▸ Residences
▸ ASSU Directory

BUSINESS/ADMINISTRATION
▸ Administration
▸ Business Organizations
▸ The Portfolio Collection

TECHNOLOGY/COMPUTING
▸ Info Tech Systems & Services
▸ Commission on Technology

Internet Archive: “Wayback Machine”



Nicholas Taylor

@nullhandle

"Thank You" by Flickr user muffintinmom under CC BY 2.0