



# One Site Among Many: Stanford and Collaborative Technical Development for Web Archiving

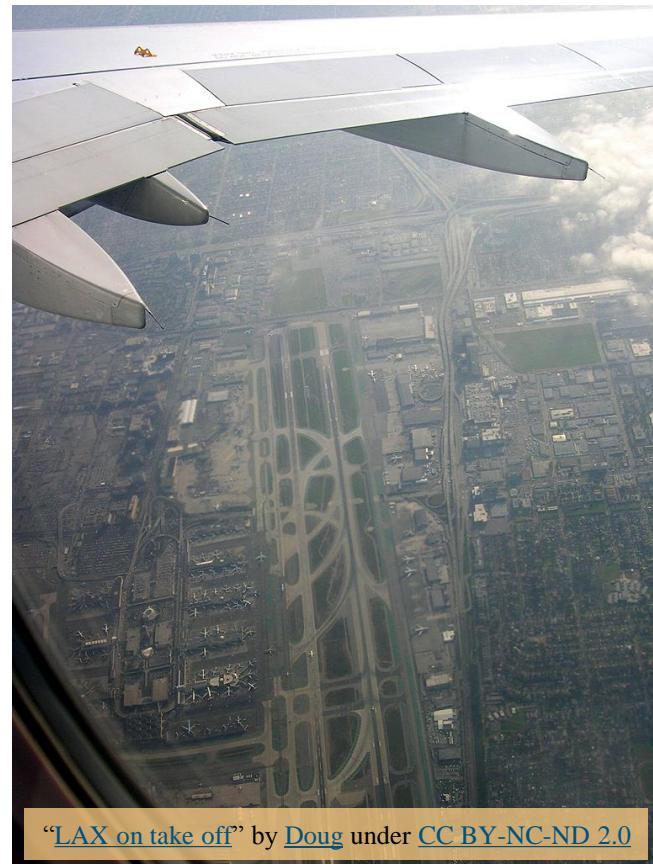
Nicholas Taylor  
[Web Archiving Service Manager](#)  
[Stanford University Libraries](#)

[PASIG 2016](#)  
March 11, 2016



# overview

- web archiving  
opportunity gaps
- situation of SUL web  
archiving
- APIs + community  
(technical)  
development





A photograph of a railway track that splits into two paths, symbolizing opportunity gaps. The tracks are made of wood and metal, set into a bed of gravel. They lead into a misty distance where a small figure of a person and a dog can be seen walking away. The surrounding area is covered in tall, dry grass and bushes. A large, semi-transparent yellow rectangular box covers the lower third of the image, containing the text "OPPORTUNITY GAPS".

# OPPORTUNITY GAPS

"Mind The Gap" by [R~P~M](#) under [CC BY-NC-ND 2.0](#)



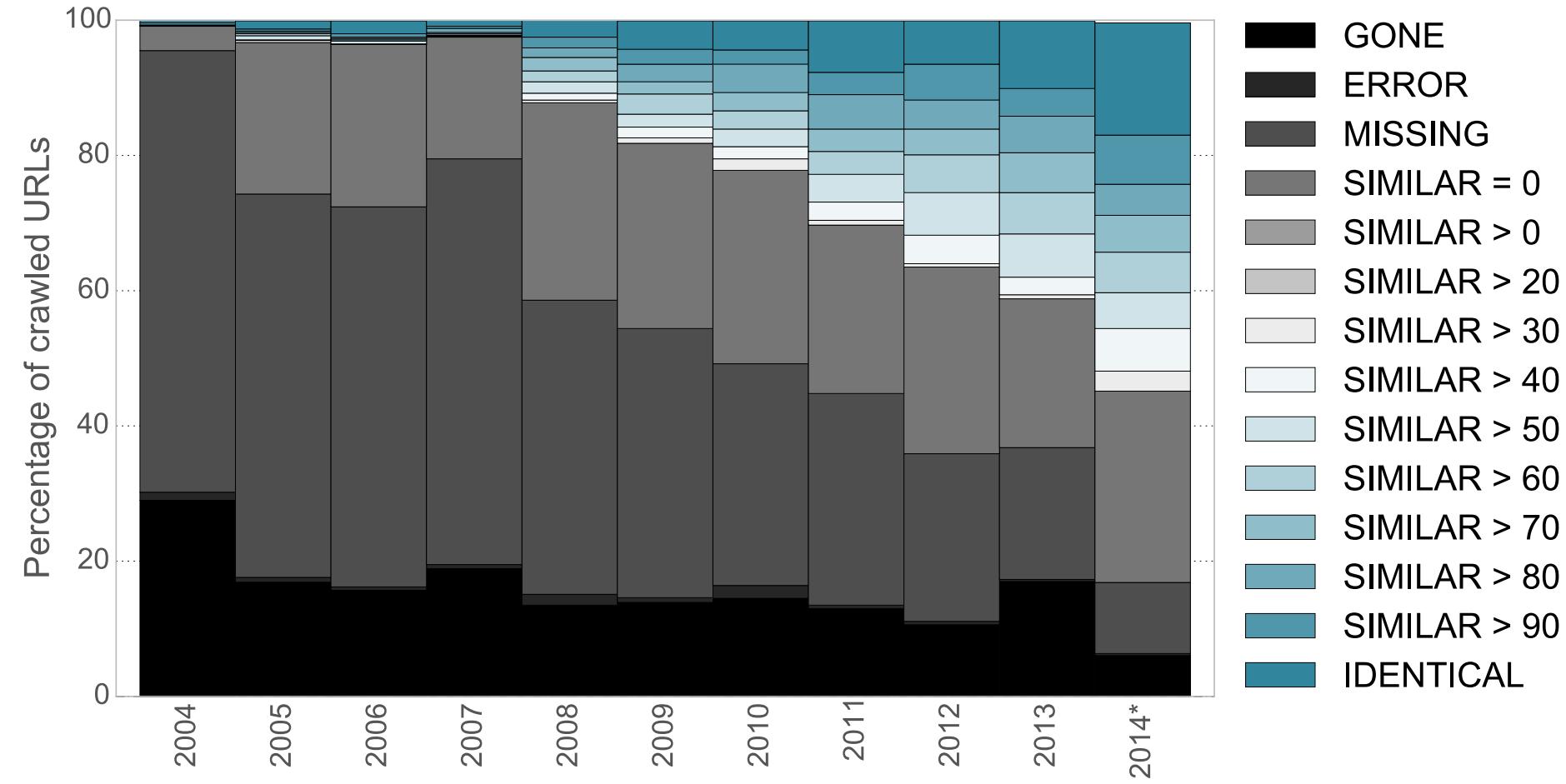
# web content > preserved web content



“[The Seeker](#)” by [C MB 166](#) under [CC BY-ND 2.0](#)



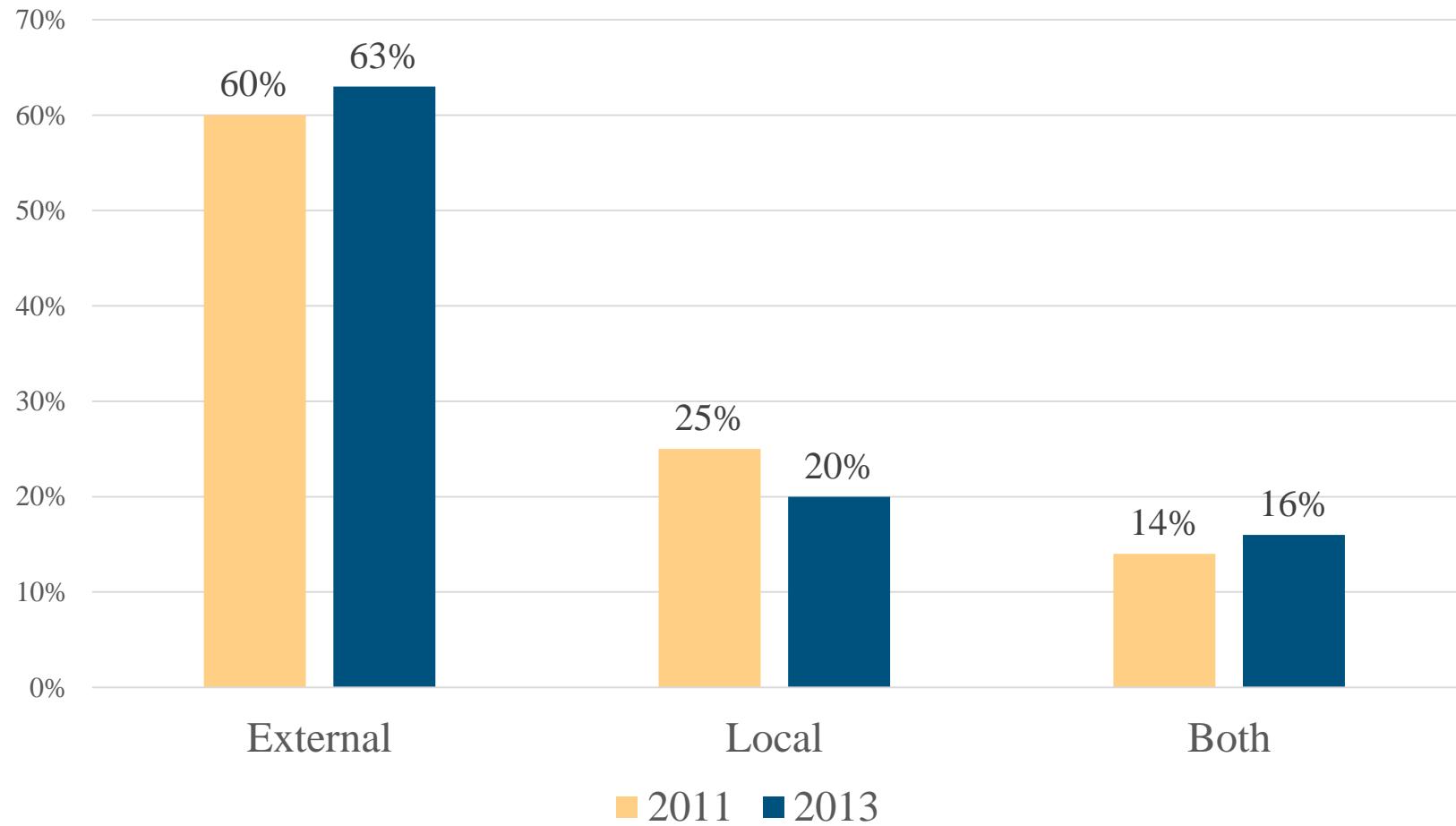
# link rot + content drift



Andrew Jackson: "[Ten years of the UK Web Archive](#)"



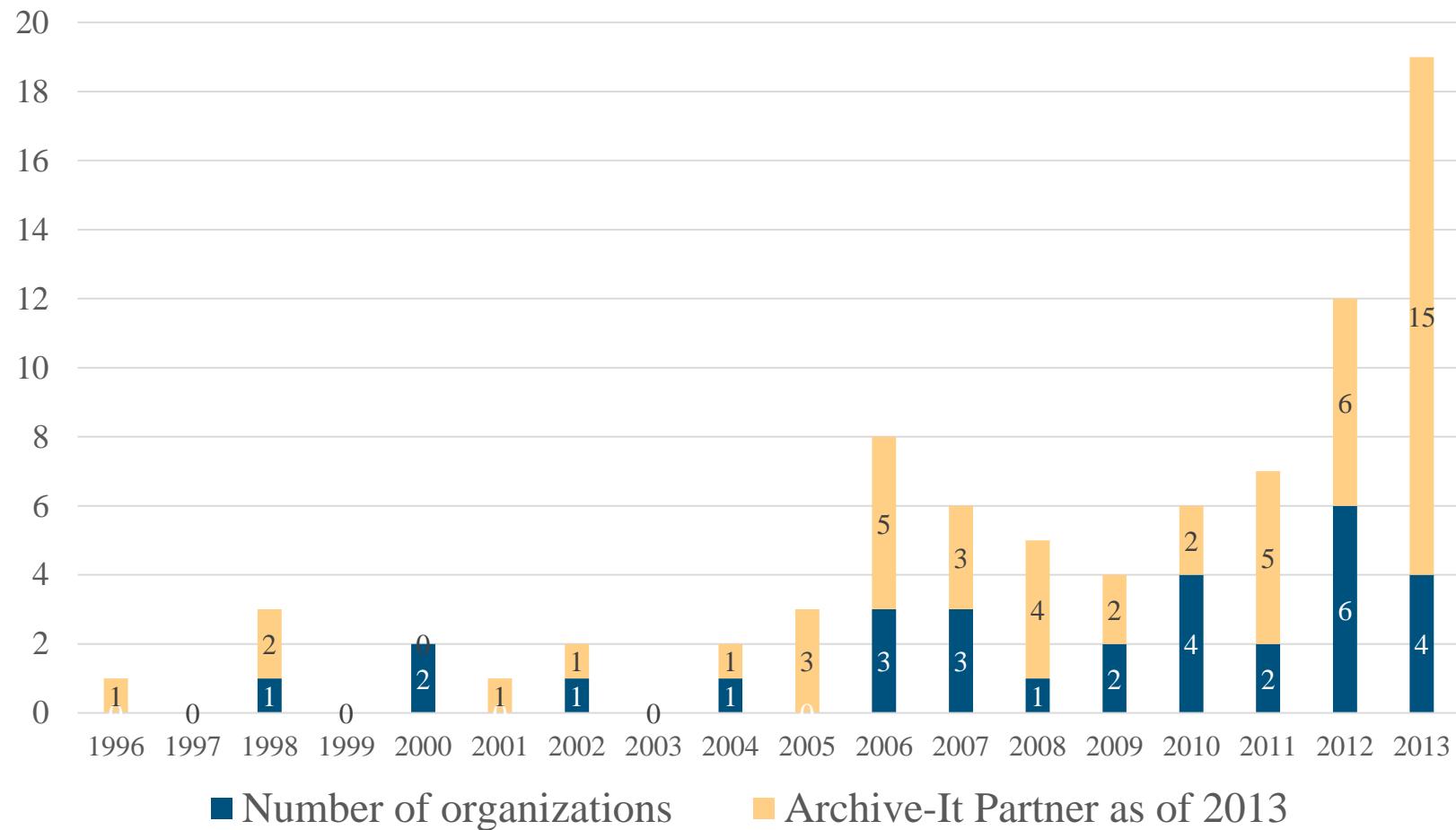
# a centralized enterprise



[NDSA: “Web Archiving in the U.S.: A 2013 Survey”](#)



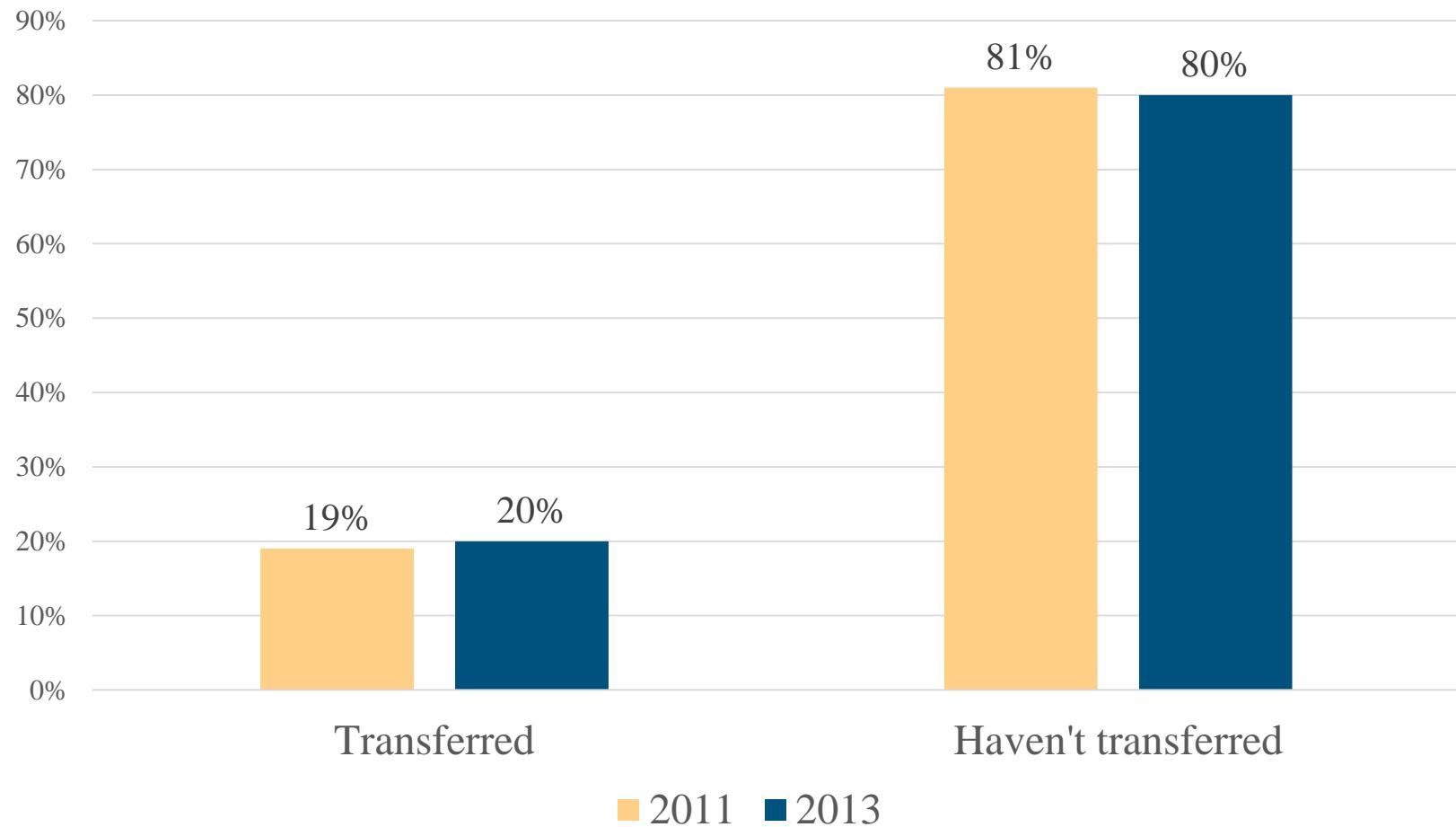
# a centralized enterprise



NDSA: “[Web Archiving in the U.S.: A 2013 Survey](#)”



# minimal local preservation



NDSA: “[Web Archiving in the U.S.: A 2013 Survey](#)”



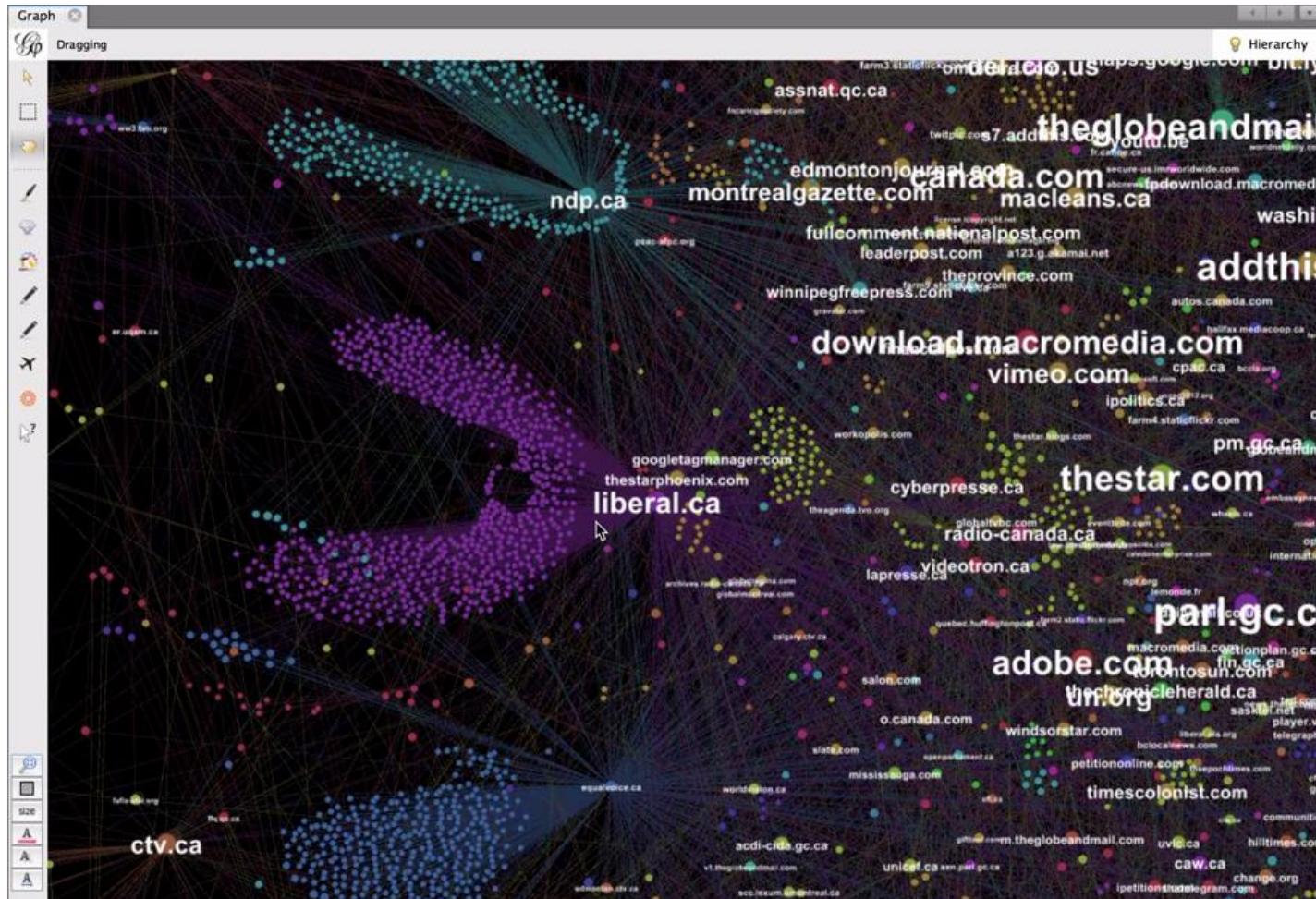
# evolving web



“Light Writing - Spider Web” by [oz dean](#) under [CC BY-ND 2.0](#)



# opportunities for research



"Exploring the Canadian Political Interest Group and Political Parties Web Sphere" by [Ian Milligan](#) under [Standard YouTube License](#)



# WHAT ARE WE DOING?



# Stanford Web Archive Portal

STANFORD UNIVERSITY LIBRARIES

## Stanford Web Archive Portal

A searchable collection of websites archived by Stanford University

Feedback

http://

Any year ▾ Browse history

Captured 2 times between November 10, 2014 and November 10, 2014

1991 1992 1993 1994 1995 1996 1997 1998 1999 2000 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014

JAN FEB 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
---	---	---	---	---	---	---	---	---	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----

MAY JUN 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
---	---	---	---	---	---	---	---	---	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----

SEP OCT 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
---	---	---	---	---	---	---	---	---	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----

Stanford

SU Home Maps & Directions Search Stanford Terms of Use Emergency Info

Stanford University Libraries: "Stanford Web Archive Portal"

STANFORD UNIVERSITY LIBRARIES

Showing 100 links visited since November 10, 2014

Feedback

Stanford University

About Stanford Admission Academics Research Campus Life STUDENTS FACULTY / STAFF PARENTS ALUMNI

Rhodes Scholars

Stanford seniors Emily Witt and Maya Krishnan are among those chosen for the prestigious scholarship.

Top Stories

Ebola response

Stanford provides guidelines for Ebola assessment, response.

Optical link

Stanford engineers take big step toward using light instead of wires inside computers.

Politics in music

Stanford music scholar explores how Indian traditional folk music fuses the devotional with the political.

MORE HEADLINES

- Stanford biologist says prejudice toward African American dialect can result in unfair rulings
- Expert places process multiple visual cues more efficiently, Stanford and VA scientists find
- Five Stanford professors named fellows of American Association for the Advancement of Science

MORE NEWS

1 DAYS AGO [@stanford](#) Scientific literature does not support claims that software-based "brain games" improve cognitive performance: stanford.io/1vZUoI

At Stanford

Mae Jemison: Imagining the Universe DEC 3 6:00 p.m.

Concert: Chamber Music Showcase DEC 4 12:00 p.m.

Film: Sergeant York 7:00 p.m.

EVENT CALENDAR

Events

Athletics

Women's volleyball

No. 3 seed Stanford begins its run for the NCAA Championship at 7 p.m. Friday in Maples Pavilion.

GOSTANFORD.COM

SCHOOLS

Business

Earth Sciences

Education

Engineering

Humanities & Sciences

Law

Medicine

DEPARTMENTS

Departments A-Z

Interdisciplinary Programs

HEALTH CARE

Stanford Health Care

Stanford Children's Health

ABOUT STANFORD

Facts

History

Accreditation

RESEARCH

Dofusresearch

Interdisciplinary Institutes

LIBRARIES

ONLINE LEARNING

Stanford Online

ADMISSION

Undergraduate

Graduate

Financial Aid

RESOURCES

A-Z Index

Maps & Directions

Search Stanford

Terms of Use

Emergency Info

Stanford University, Stanford, California 94305 Copyright Complaints Trademark Notice

Apply

Visit Campus

Make a Gift

Find a Job

Contact Us

Stanford University

SU Home Maps & Directions Search Stanford Terms of Use Emergency Info

Stanford University, Stanford, California 94305 Copyright Complaints Trademark Notice



# SearchWorks (online catalog)

STANFORD UNIVERSITY LIBRARIES

## SearchWorks catalog

All fields

Find materials by...

Access

At the Library	6,467,350
Online	1,844,821

Resource type

Archive/Manuscript	28,164
Book	7,004,387
Database	1,837
Dataset	1,095
Equipment	350
Image	6,968
Journal/Periodical	432,863
Map	82,999
Music recording	171,600
Music score	94,943
Newspaper	7,172
Object	69
Software/Multimedia	8,904
Sound recording	4,584
Video	74,758

Library

STANFORD UNIVERSITY LIBRARIES

Stanford University Libraries Hours & locations My Account Ask us Opt out of analytics

SU Home Maps & Directions Search Stanford Terms of Use Emergency Info  
© Stanford University, Stanford, California 94305. Copyright Complaints

Stanford University Libraries: "SearchWorks"

STANFORD UNIVERSITY LIBRARIES  My Account Feedback

## SearchWorks catalog

All fields

Back to results  1 of 6,962,846   Cite   Select

### Consolidated federal funds reports--county areas

AUTHOR/CREATOR U.S. Census Bureau, creator.  
LANGUAGE English.  
IMPRINT U.S. Department of Commerce, United States Census Bureau  
Text  
FORMAT electronic  
DIGITAL ORIGIN born digital  
ORIGINAL WEBSITE <http://www.census.gov/prod/www/abs/cfr.html>

  
12-Apr-2013  12-Oct-2013  12-Apr-2013

In collection Web Archive Collection From Public APO

DIGITAL CONTENT 76 items  
COLLECTION PURL <http://purl.stanford.edu/gb667dy0829>

Contributors CONTRIBUTOR Stanford University, Social Sciences Resource Group, collector.

Contents/Summary DESCRIPTION This is a Census Bureau Web page that contains PDFs of the Consolidated federal funds reports (CFFR). These reports present data on federal government expenditures or obligations in state, county, and subcounty areas of the United States, including the District of Columbia and U.S. Outlying Areas. The reports contain statistics on the geographic distribution of federal program expenditures, using data submitted by federal departments and agencies. (source: Nielsen Book Data)

Subjects SUBJECT Grants-in-aid > United States > Statistics  
Economic assistance, Domestic > United States > Statistics  
Government lending > United States > Statistics  
Public contracts > United States > Statistics  
United States > Appropriations and expenditures > Statistics

Genre GENRE Archived website

Bibliographic information NOTE Archived by Stanford University, Social Sciences Resource Group.

WEB ARCHIVING SERVICE Archive-It

COLLECTION Web Archive Collection From Public APO

Librarian view | Catkey: 10790679

STANFORD UNIVERSITY LIBRARIES

Stanford University Libraries Hours & locations My Account Ask us Opt out of analytics

SU Home Maps & Directions Search Stanford Terms of Use Emergency Info  
© Stanford University, Stanford, California 94305. Copyright Complaints

Stanford University



# web archaeology (SLAC)

oldweb.today Time Left 09:23

NCSA Mosaic on Linux 2.2

[about this browser](#)

Current Page Archived On:  
1994-01-02 00:00:00

Requested Date/Time:  
1994-01-02 00:00:00

Loaded 2 resources, spanning  
1993-05-02 to 1994-01-02  
00:00:00 00:00:00

from public web archives:  
[- Stanford Web Archive](#)

[Donate to support oldweb today!](#)

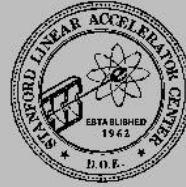
[Source code on github!](#)  
Contact: oldweb.today@rhizome.org

NCSA X Mosaic 2.7b6 [WorldWideWeb SLAC Home Page]

File Options Navigate Annotate News

URL: <http://slacvm.slac.stanford.edu/FIND/slac.html>

None Help

 WorldWideWeb SLAC Home Page

[SLAC](#) 24 Dec 1993

Use the [WorldWideWeb \(WWW\)](#) service to gain access to a wide range of information at SLAC and elsewhere around the globe. Emphasized text like [this](#) is a hypertext link.

You may view [WWW information](#) through GUI or line-mode [browsers](#). At least most SLAC pages have been tested on the [MidasWWW X Window System](#) browser. Note that over time links may move around on a page, migrate to others, or be removed entirely, due to the dynamic nature of the Web.

### SLAC Information

**People and organizations:** [people at SLAC](#), [particle physics people](#) and [institutions](#).

**Library:** [SPIRES-HEP](#), [Current PPF-list](#), [Books](#), [SLACspeak glossary](#), [other databases](#).

**Physics Preprint Bulletin Boards (full-text postscript) :** [today](#), [yesterday](#), [last seven days](#), [week before that](#), [anytime](#).

**Seminars:** [today](#), [tomorrow](#), [this week](#), [next week](#), [anytime](#).

**Conferences:** [this month](#), [next month](#), [next year](#), [next summer](#), [all future](#), [let me search](#).

**News:**

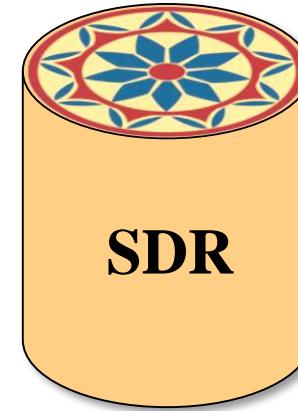
oldweb.today: "WorldWideWeb SLAC Home Page"



# building + integrating infrastructure



capture



preservation

Stanford Web Archive Portal  
A searchable collection of websites archived by Stanford University

http:// [ ] Any year [ ] Browse history

Featured archived sites

- SLAC first web page
- SLAC home page 1992-1995
- SLAC home page 1995-1999
- SLAC home page 1999

Stanford University

SU Home Maps & Directions Search Stanford Terms of Use Emergency Info  
© Stanford University, Stanford, California 94301 Copyright Complaints

access

SearchWorks catalog [ ] Help [ ] Advanced [ ] Google [ ] SUL Catalog [ ]

Looking for articles? Articles are not indexed in SearchWorks. Select a database to search for articles.

Citation Finder [ ] Find an article in one of Stanford's journal subscriptions.

Search [ ] Search for articles in multiple databases simultaneously.

Guides to searching [ ] Guides to searching for articles at Stanford.

Help with SearchWorks [ ] Guide to SearchWorks basics [ ] Overview of SearchWorks features, including basic search strategies, facets, browsing.

Library [ ]

Stanford University Libraries Stanford University Libraries Hours & locations My Account Ask us Opt out of analytics

Stanford University

SU Home Maps & Directions Search Stanford Terms of Use Emergency Info  
© Stanford University, Stanford, California 94301 Copyright Complaints

discovery



# APIS + COMMUNITY DEVELOPMENT





# web archiving lifecycle



Internet Archive: "[The Web Archiving Life Cycle Model](#)"

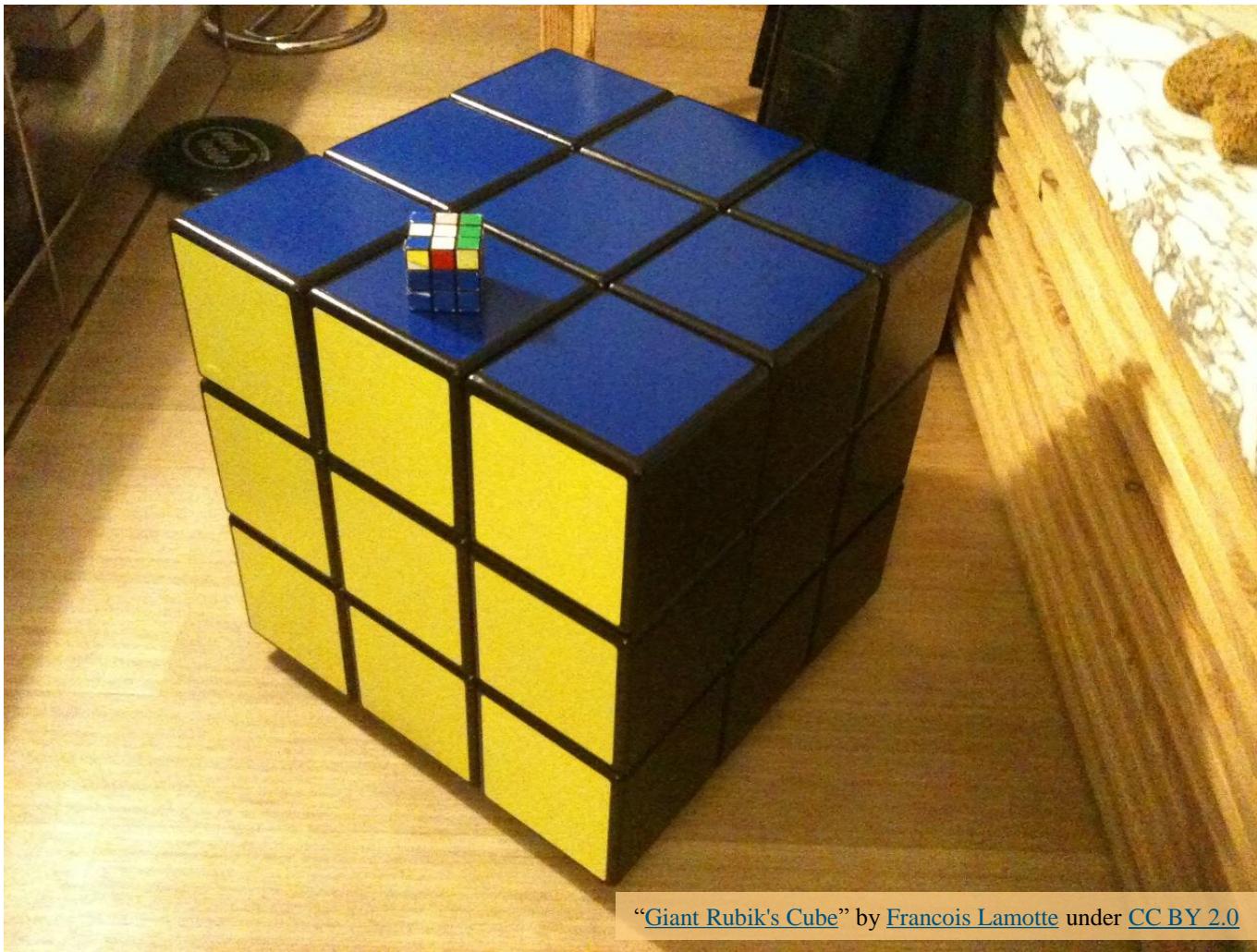


# functional overlap

	Appraisal and Selection	Scoping	Data Capture	Storage and Organization	QA and Analysis	Metadata / Description	Access / Use / Reuse	Preservation	Risk Management
<a href="#">ACT</a>	Yellow				Orange	Grey			Blue
<a href="#">Archive-It</a>	Yellow	Orange	Orange	Orange	Orange	Grey	Blue	Blue	
<a href="#">AtN</a>	Yellow	Orange	Orange	Orange	Orange	Grey	Blue	Blue	
BCWeb	Yellow	Orange	White	Orange	White				
<a href="#">CDL WAS</a>	Yellow	Orange	Orange	Orange	Orange	Grey	Blue	Blue	Blue
<a href="#">DigiBoard</a>	Yellow	Orange	Orange	White	Orange	Grey			Blue
<a href="#">Islandora WARC Solution Pack</a>				Orange	White	Grey	Blue	Blue	
<a href="#">Netarchive Suite</a>	Yellow	Orange	Orange	Orange	Orange	Grey	Blue		
<a href="#">PageFreezer</a>	Yellow	Orange	Orange	Orange	Orange	White	Blue		Blue
<a href="#">UNT Nomination Tool</a>	Yellow	White	White	White	White	Grey			
<a href="#">WCT</a>		Orange	Orange	White	Orange	Grey			Blue



# smaller, modular components



[“Giant Rubik's Cube”](#) by [Francois Lamotte](#) under [CC BY 2.0](#)



# community seed

## National Digital Platform Projects funded in August 2015

### Systems Interoperability and Collaborative Development for Web Archiving

(LG-71-15-0174-15): The Internet Archive, working with partner organizations University of North Texas, Rutgers University, and Stanford University Library will undertake a two-year research project to explore techniques that can expand national web archiving capacity in several areas.





# API candidates

- capture tool/proxy interconnect
- capture tool management
- data import/export
- query + extraction
- integrity audit + repair
- descriptive metadata
- logs + analytics
- renderings/derivative formats
- federated data delivery
- federated replay
- federated full-text search



# let's combine forces



“Stages of flow” by Peter Thoeny under [CC BY-NC-SA 2.0](#)