



Stanford | LIBRARIES

SUL/EAL Web Archiving Programmatic and Technical Concerns

Nicholas Taylor ([@nullhandle](#))

Program Manager, [LOCKSS](#) and [Web Archiving
Stanford Libraries](#)

[Collaborative, Selective, Contemporary
IIPC Web Archiving Conference](#)

13 November 2018



Stanford web archiving

- selective
 - self-archiving
 - 3rd party content
- 7 Archive-It accounts
- Heritrix, Webrecorder
- local preservation, discovery, access
- program manager, curators, tech services staff, assistants
- tens of collections
- thousands of seeds





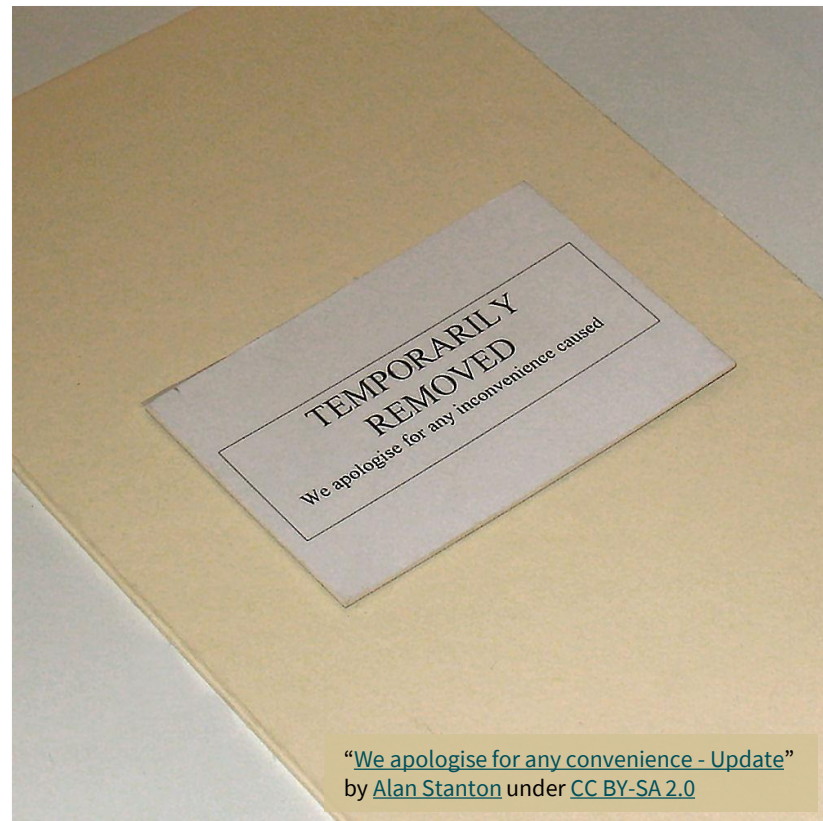
Policy

DO NOT
DUPLICATE



policy overview

- **obey** robots.txt + narrative policy directives
- **notify** of archiving + allow opt-out for most third-party content
- six-month public access **embargo** on SU-hosted platforms
- can **skip embargo + notice** for SU, open license (e.g., CC), U.S. FedGov content





notification / permission

- inform **content owners** of:
 - **inclusion** in SU collections
 - **preservation** in SDR
 - **access** via SWAP after embargo
 - right of **opt-out**
- affirmative **permission** needed to override robots.txt or narrative directives preventing archiving
- **translate** when possible





Quality Assurance



quality assurance goals

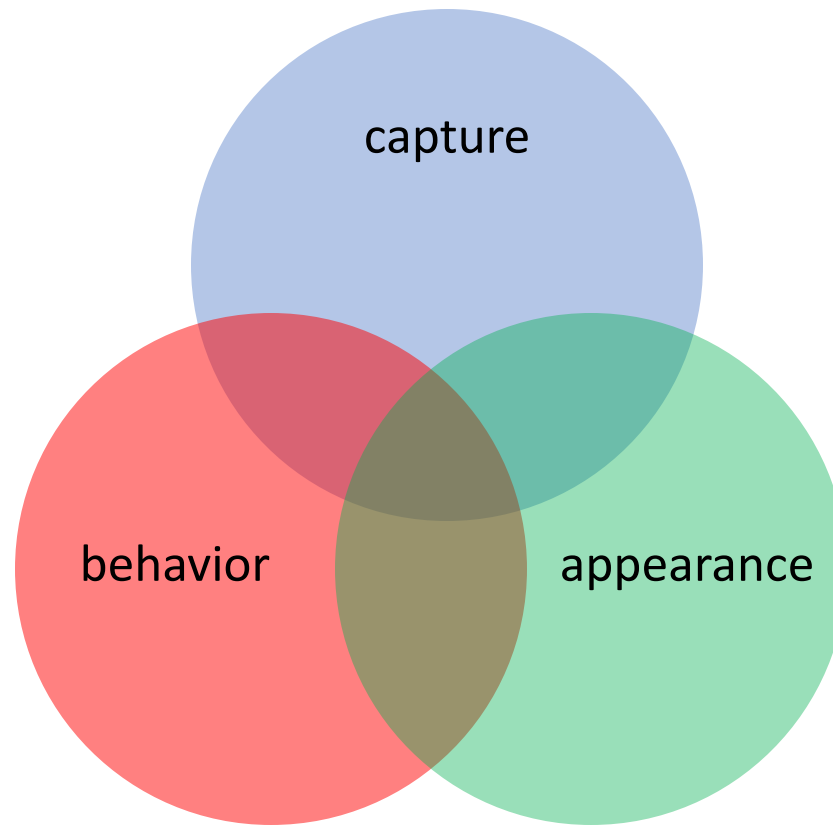
- **maximize** impact + efficiency of QA efforts
- **enable** diverse, distributed, + approachable contributions
- **calibrate** investments in quality based on tool capabilities



“Goals” by [Eric Peacock](#) under [CC BY-NC-SA 2.0](#)



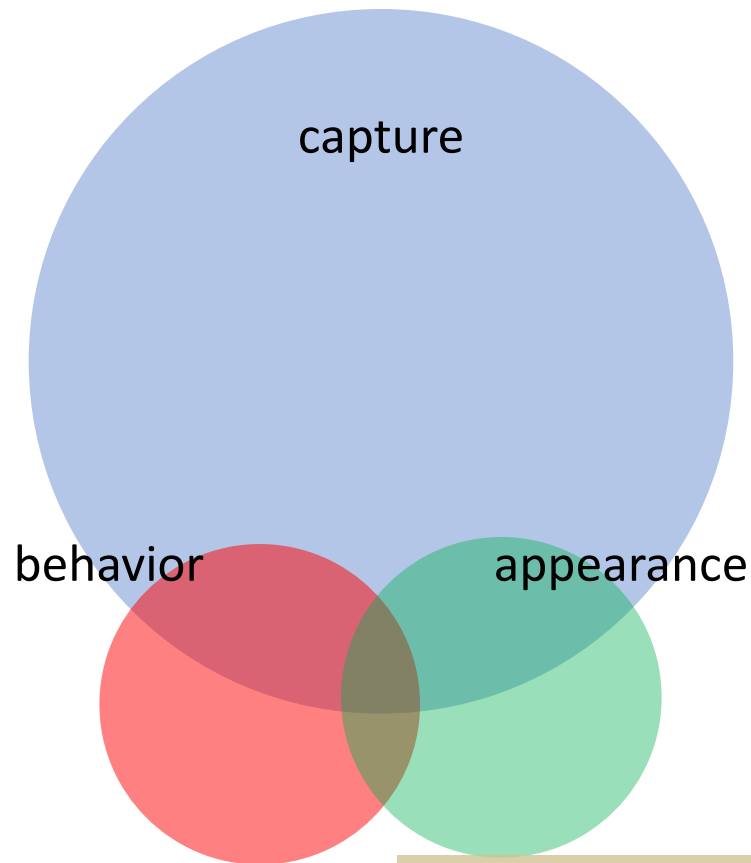
capture, behavior, appearance



[NYARC: "I. Introduction - NYARC Documentation"](#)



capture, behavior, appearance



[NYARC: "I. Introduction - NYARC Documentation"](#)



in practice

care more about...

- report data
- crawl finishing
- 4xx, 5xx, complete robots.txt block
- plausible duration
- plausible object counts
- scoping out extraneous content
- new seeds

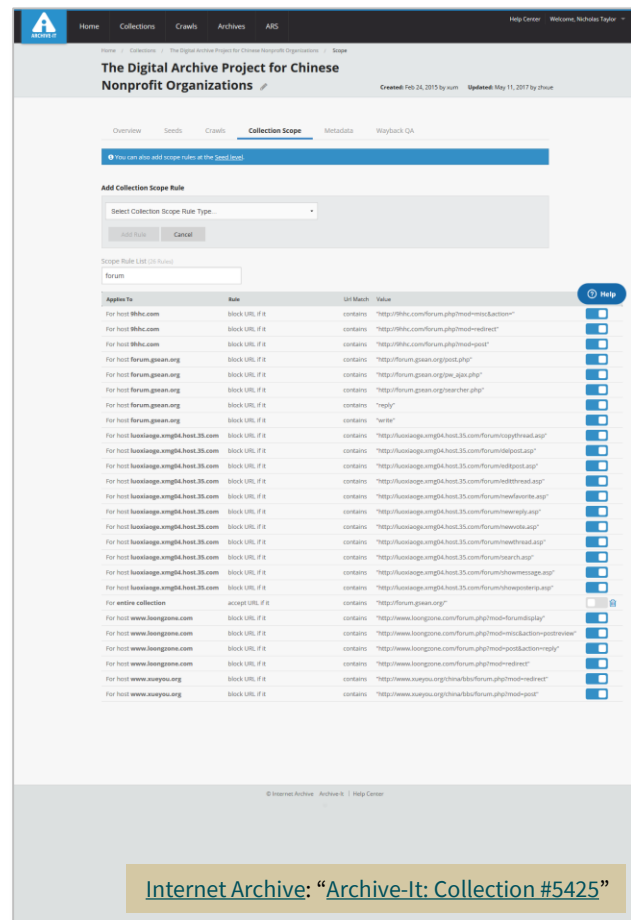
care less about...

- visual inspection
- reviewing every capture
- appearance fidelity
- behavior fidelity
- partial content out of scope
- partial content blocked by robots.txt
- ongoing seeds



QA challenges for EAL collections

- interpreting **foreign language** page content
- **social media** capture
 - authentication
 - JavaScript
 - scoping
- **forum** websites capture
- sites are **ephemeral** or change addresses





Discovery



SearchWorks (online catalog)

The screenshot shows the Stanford SearchWorks catalog homepage. At the top, there's a search bar with "All fields" selected and a search button. Below the search bar, there are links for "Library services", "Advanced search", "Course reserves", and "Selections (0)". The main content area is titled "Books, media, & more" and includes a "Find materials by..." section with filters for Access (At the Library: 6,745,359; Online: 2,377,788; On order: 13,983), Resource type, Library, and Language. There are also "Featured resources" like Digital collections, Government documents, Theses & dissertations, Circulation, Course reserves, and Databases. A "Search Stanford's library resources" section offers options like Catalog, Articles, Library website, and Yewno. At the bottom, there's a "More search tools" section and a footer with Stanford University information.

The screenshot shows the Stanford SearchWorks catalog record for the Carnegie Foundation for the Advancement of Teaching. The record includes a title, a description, and a list of metadata. The metadata includes the type of resource (Text), the date captured (2004/07/12/01743), the language (English), the digital origin (born digital), and the form (electronic). The record also includes a list of creators/contributors, abstracts/contents, subjects, and bibliographic information. The footer of the record shows the Stanford Libraries logo and the text "Stanford Libraries: 'Carnegie Foundation for the Advancement of Teaching'".



metadata overview

- collaboration b/t:
 - digital library group
 - technical services
 - curatorial unit
- collection + seed level records
- original cataloging in spreadsheet template
- crosswalk to MODS + optionally to Archive-It Dublin Core spreadsheet template





metadata fields

- **spreadsheet template:**
 - type of resource, genre, form, digital origin, mime type, “archived by” note, repository
- **digital library group:**
 - druid, sourceId, dateCaptured, collector, site URL, archiving service, SWAP URL
- **curatorial unit:**
 - title, creator (+ type), language, abstract, subject terms (+ type)
- **technical services:**
 - (authorities)





Spotlight (exhibits)

Stanford LIBRARIES

Recording Civic Action in China
Chinese NGO web archiving project

Home Browse Studies on Chinese NGOs About

Everything Search...

Home Browse

阿拉善
Environment
25 ITEMS

女权之声
Women
25 ITEMS

工友之家
Migrant Labor
25 ITEMS

梁漱溟乡村建设
Rural Development
25 ITEMS

同城
LGBT
25 ITEMS

益众社区
Social Service
25 ITEMS

chain.net.cn
Health
25 ITEMS

Education
25 ITEMS

NGO China
25 ITEMS

Stanford LIBRARIES

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

Stanford University

Stanford Libraries: "Browse Exhibit | Recording Civic Action in China"

Stanford LIBRARIES

Recording Civic Action in China
Chinese NGO web archiving project

Home Browse Studies on Chinese NGOs About

Everything Search...

Home Browse

阿拉善
Environment
25 ITEMS

女权之声
Women
25 ITEMS

工友之家
Migrant Labor
25 ITEMS

梁漱溟乡村建设
Rural Development
25 ITEMS

同城
LGBT
25 ITEMS

益众社区
Social Service
25 ITEMS

chain.net.cn
Health
25 ITEMS

Education
25 ITEMS

NGO China
25 ITEMS

Stanford LIBRARIES

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

Stanford University

Stanford Libraries: "道和环境与发展研究所。"



SWAP (web archive replay)

STANFORD UNIVERSITY LIBRARIES

Feedback

Stanford Web Archive Portal

A searchable collection of websites archived by Stanford University

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

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

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

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

STANFORD UNIVERSITY LIBRARIES

Stanford Web Archive Portal

Stanford University

About Stanford Admission Academics Research Campus Life

STUDENTS FACULTY/STAFF PARENTS ALUMNI

Rhodes Scholars

Stanford seniors Emily Witts 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 linguist says prejudice toward African American dialect can result in unfair rulings

Expert pilots 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/1VZUol

At Stanford

DEIVERSITY AT STANFORD

Learn about the many programs and initiatives that reflect Stanford's strong commitment to diversity.

Events

DEC 3

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

DEC 4

Concert: Chamber Music Showcase 12:35 p.m.

Film: Sergeant York 7:00 p.m.

EVENT CALENDAR

Athletics

Women's volleyball

No. 1 seed Stanford begins its run for the NCAA Championship at 7 p.m. Friday in Naples Pavilion.

STANFORD.COM

MORE SITES

SCHOOLS

Business

Earth Sciences

Education

Engineering

Humanities & Sciences

Law

Medicine

DEPARTMENTS

Departments A - Z

Interdisciplinary Programs

RESEARCH

Software

Interdisciplinary Institutes

Libraries

HEALTH CARE

Stanford Health Care

Stanford Children's Health

ONLINE LEARNING

Stanford Online

ABOUT STANFORD

Facts

History

Accreditation

ADMISSION

Undergraduate

Graduate

Financial Aid

RESOURCES

A - Z Index

Campus Map

Directory

Stanford Profiles

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

Stanford Libraries: "Stanford Web Archive Portal"

A photograph of a brass telescope mounted on a balcony railing, overlooking a city at night. The city lights are blurred into a bokeh effect against a dark blue sky. The telescope is in the foreground, angled towards the right. A semi-transparent white box with the text "What's Next?" is overlaid on the middle of the image.

What's Next?

“View over Paris” by [Carlos ZGZ](#) under [CC BY 2.0](#)



incremental improvements

- curate + promote collections
- enhance metadata + create records
- address bugs + inefficiencies in workflows
- improve staffing for repository content ingest
- explore Social Feed Manager for social media capture



"Hoses at Burg Eltz" by Isaac Wedin under CC BY 2.0



Questions

“Any Questions?” by Matthias Ripp under [CC BY 2.0](#)