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

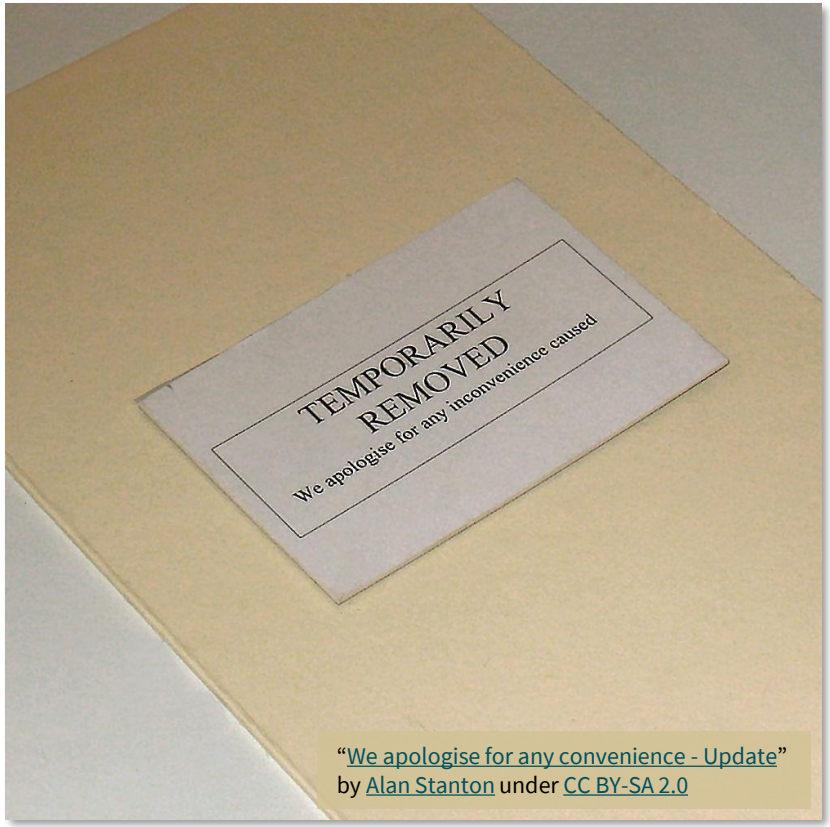
- selective
  - self-archiving
  - 3rd party content
- 7 Archive-It accounts
- Heritrix, Webrecorder
- local preservation, discovery, access
- program manager, curators, tech services staff, assistants
- tens of collections
- thousands of seeds

Internet Archive: "Stanford University Homepage"

Policy

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

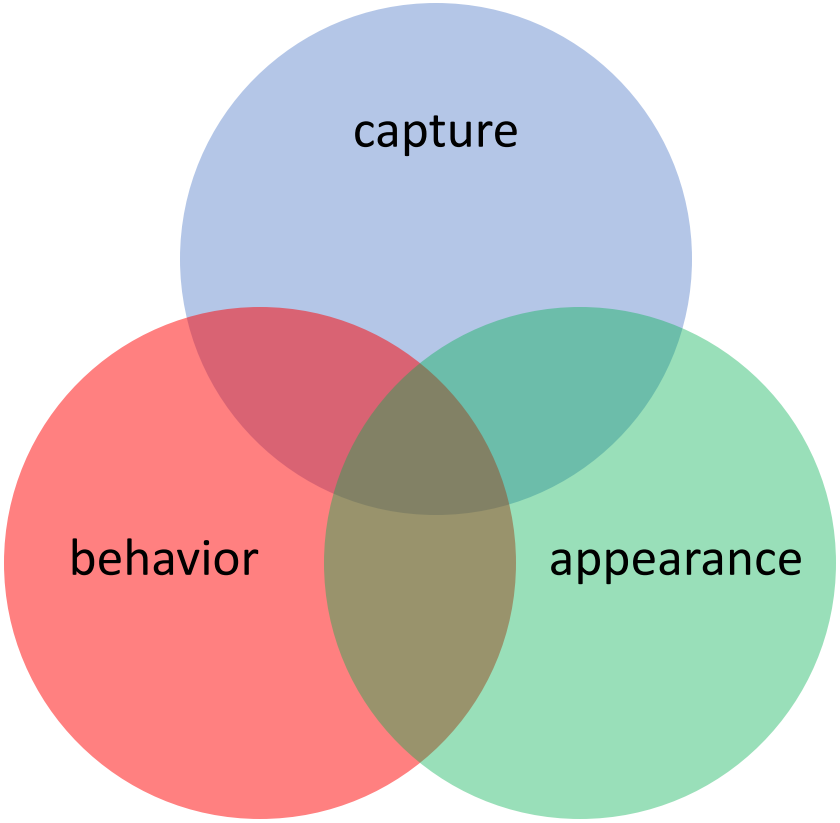
- inform content owners of:
  - inclusion in SU collections
  - preservation in SDR
  - access via SWAP after embargo
  - right of opt-out
- affirmative permission needed to override robots.txt or narrative directives preventing archiving
- translate when possible

Quality Assurance

quality assurance goals

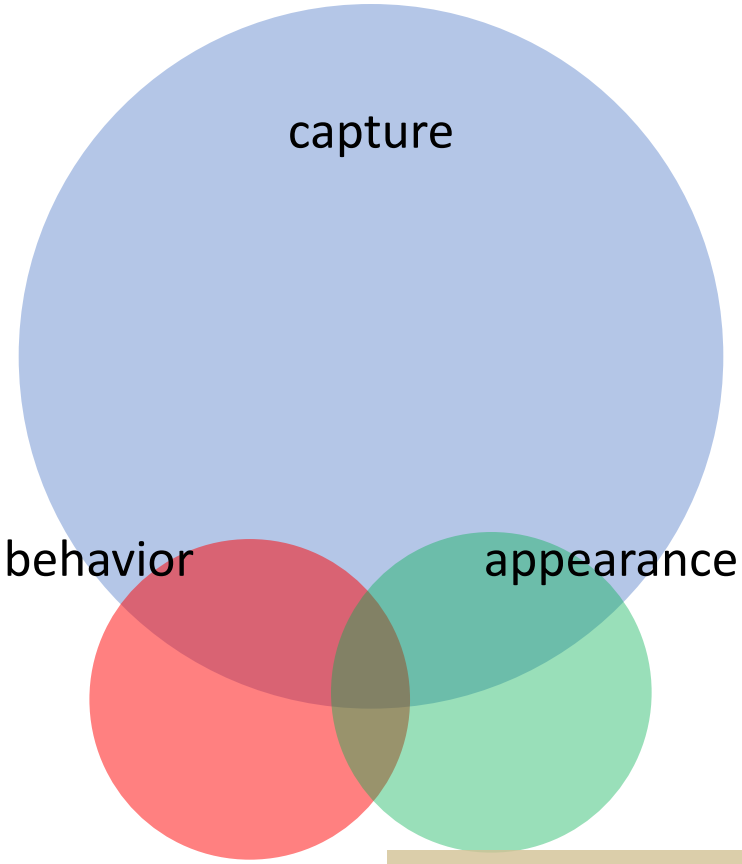
- maximize impact + efficiency of QA efforts
- enable diverse, distributed, + approachable contributions
- calibrate investments in quality based on tool capabilities

capture, behavior, appearance



NYARC: "I. Introduction - NYARC Documentation"

capture, behavior, appearance



NYARC: "I. Introduction - NYARC Documentation"

in practice

care more about…

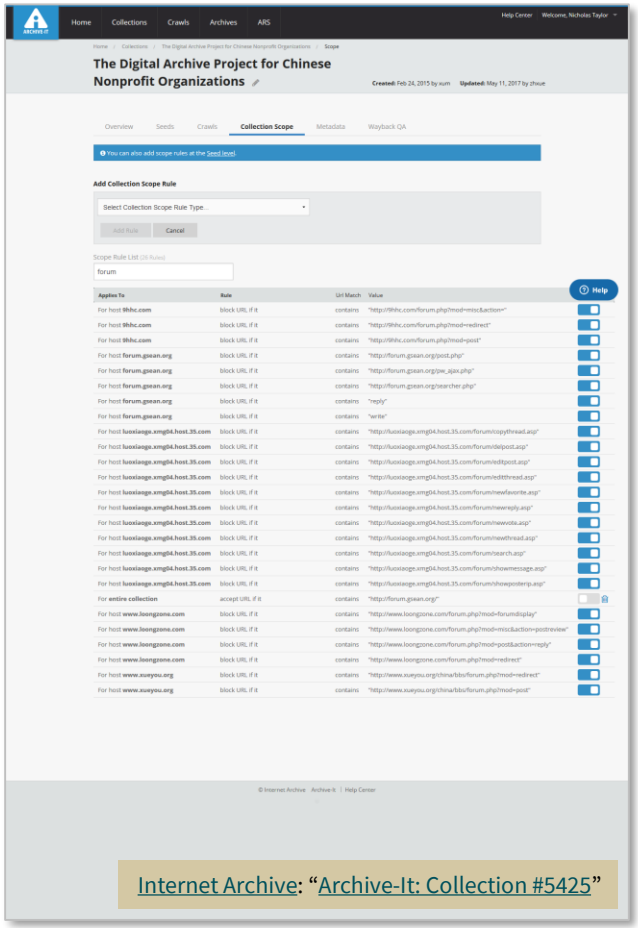
- report data
- crawl finishing
- 4xx, 5xx, complete robots.txt block
- plausible duration
- plausible object counts
- scoping out extraneous content
- new seeds

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

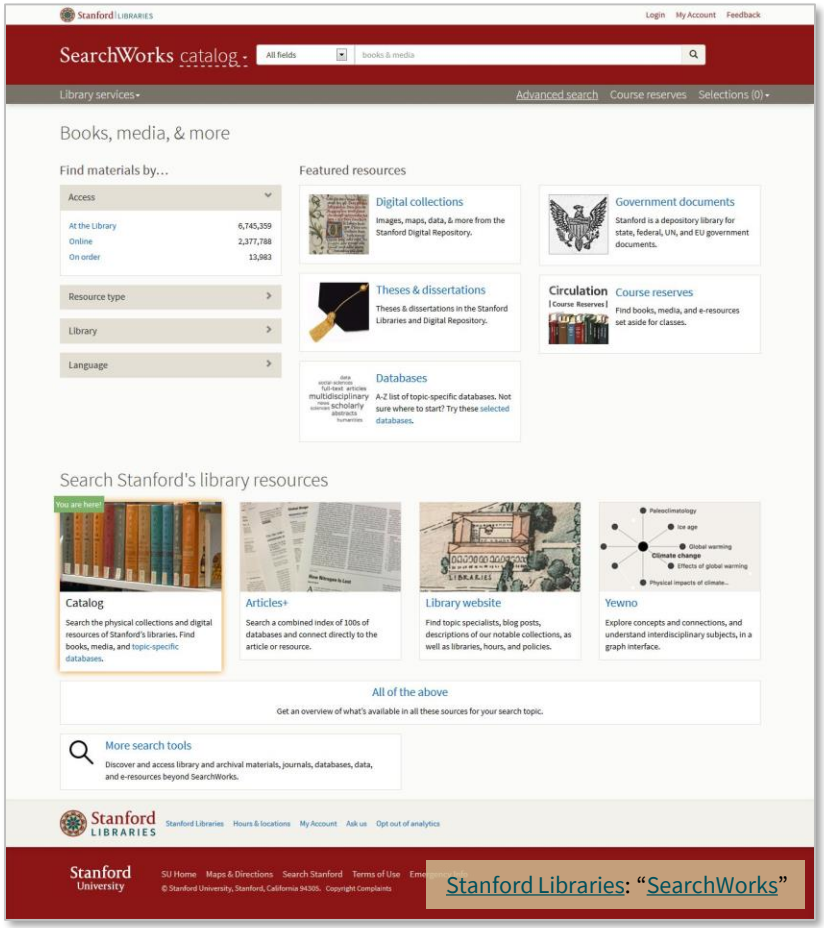
- interpreting foreign language page content
- social media capture
  - authentication
  - JavaScript
  - scoping
- forum websites capture
- sites are ephemeral or change addresses



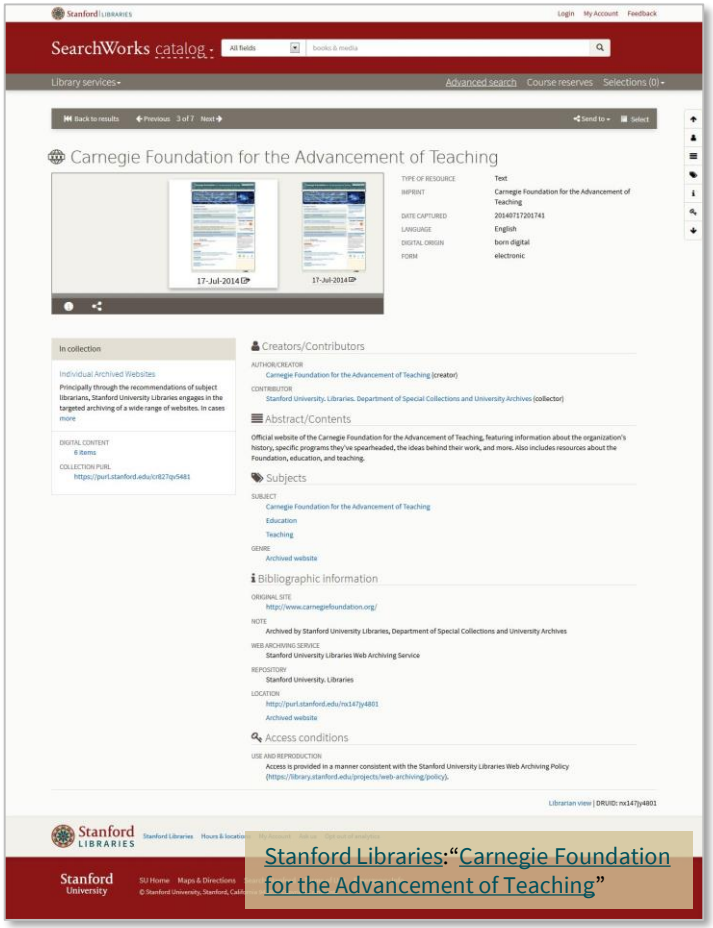
Internet Archive: "Archive-It: Collection #5425"

Discovery

SearchWorks (online catalog)



Stanford Libraries: "SearchWorks"



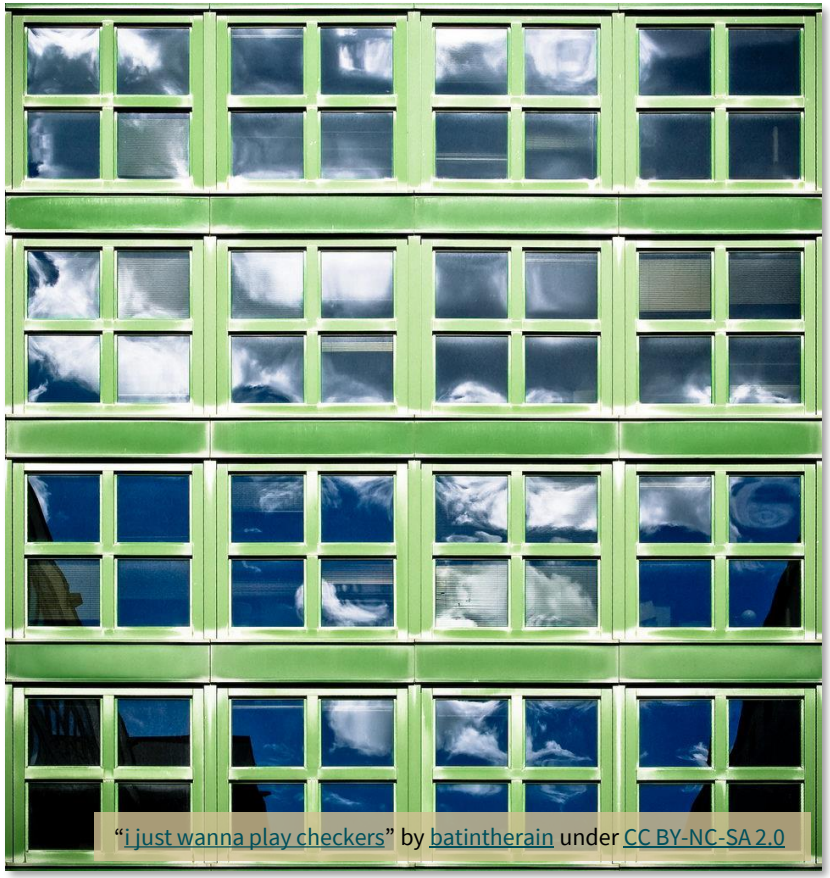
Stanford Libraries:"Carnegie Foundation for the Advancement of Teaching"

metadata overview

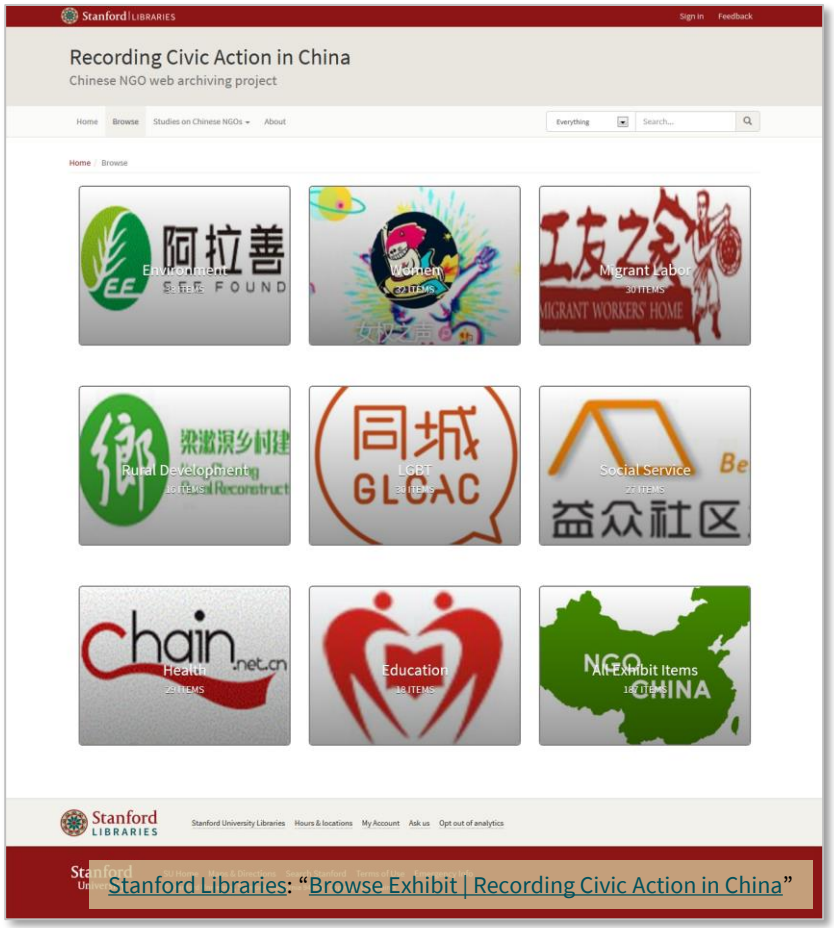
- collaboration b/t:
  - digital library group
  - technical services
  - curatorial unit
- collection + seed level records
- original cataloging in spreadsheet template
- crosswalk to MODS + optionally to Archive-It Dublin Core spreadsheet template

metadata fields

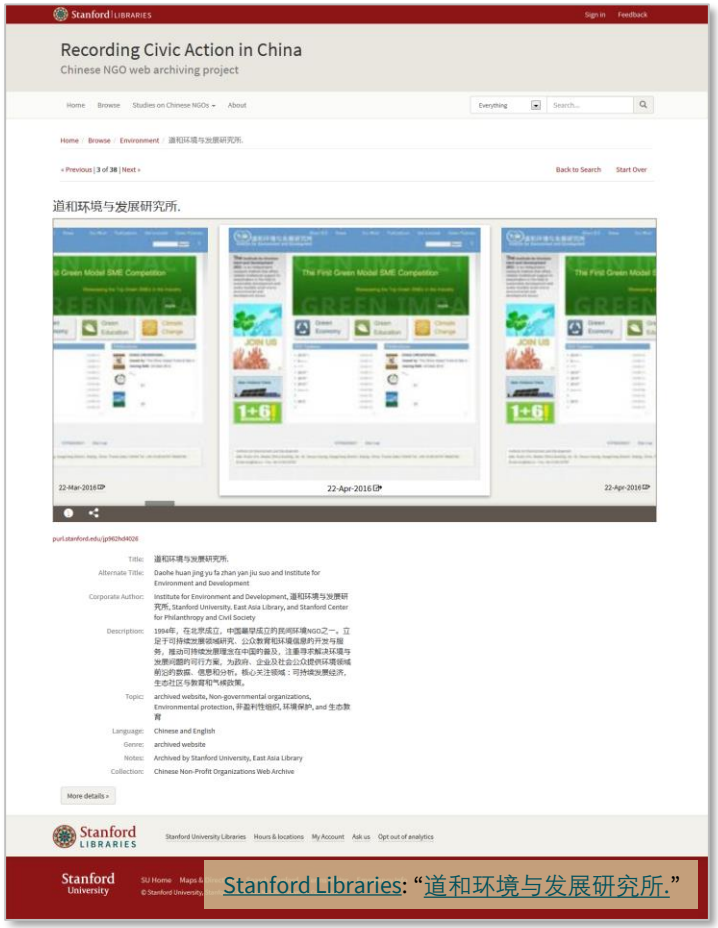
- spreadsheet template:
  - type of resource, genre, form, digital origin, mime type, “archived by” note, repository
- digital library group:
  - druid, sourceId, dateCaptured, collector, site URL, archiving service, SWAP URL
- curatorial unit:
  - title, creator (+ type), language, abstract, subject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



Spotlight (exhibits)

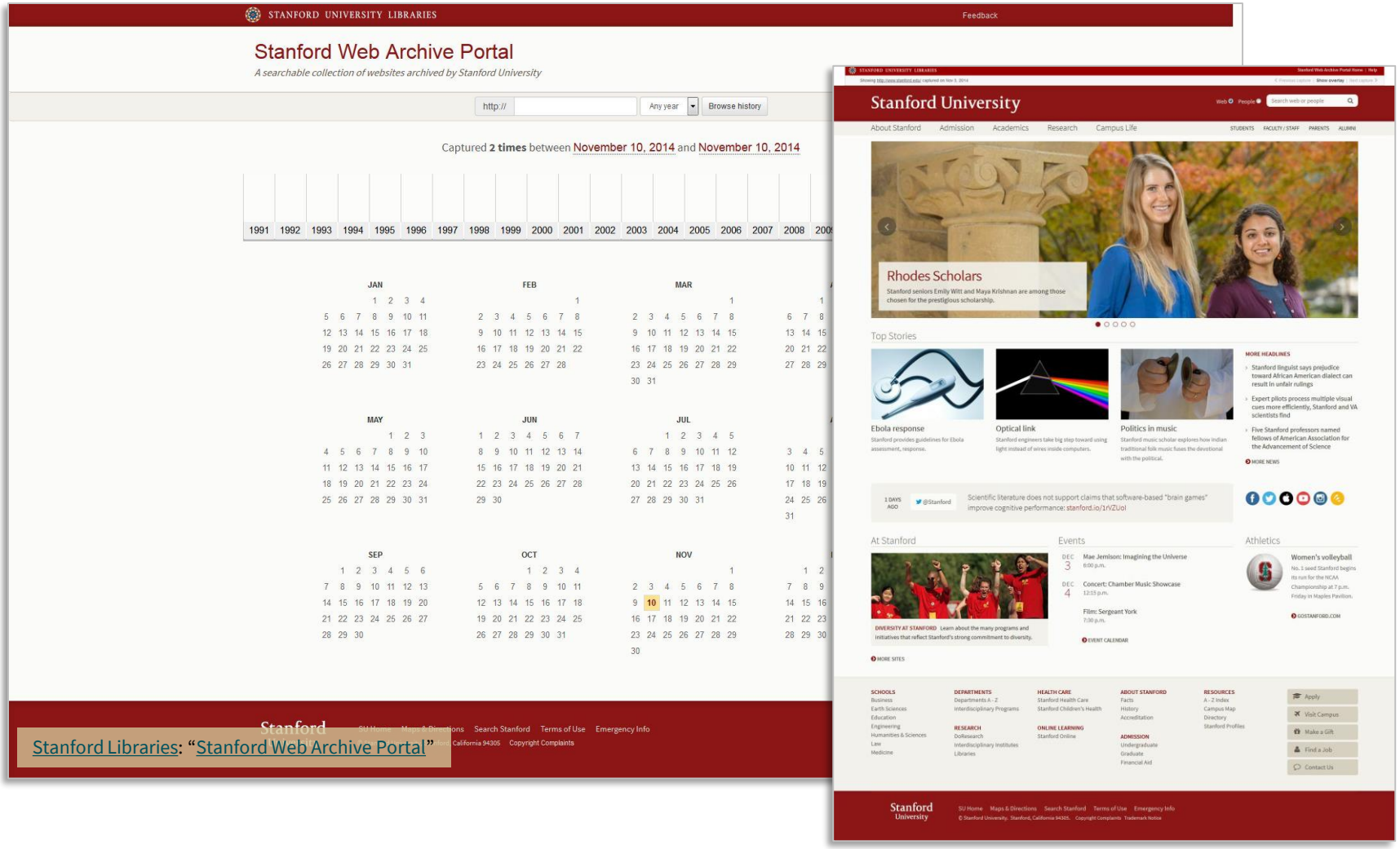


Stanford Libraries: "Browse Exhibit | Recording Civic Action in China"



Stanford Libraries: "道和环境与发展研究所."

SWAP (web archive replay)



Stanford Libraries: "Stanford Web Archive Portal"

What's Next?

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