Maintenance Practices for Web Archives

Stanford University

Laura Wrubel
lwrubel@stanford.edu

Ed Summers
edsu@stanford.edu

IIPC Web Archiving Conference
May 12, 2023

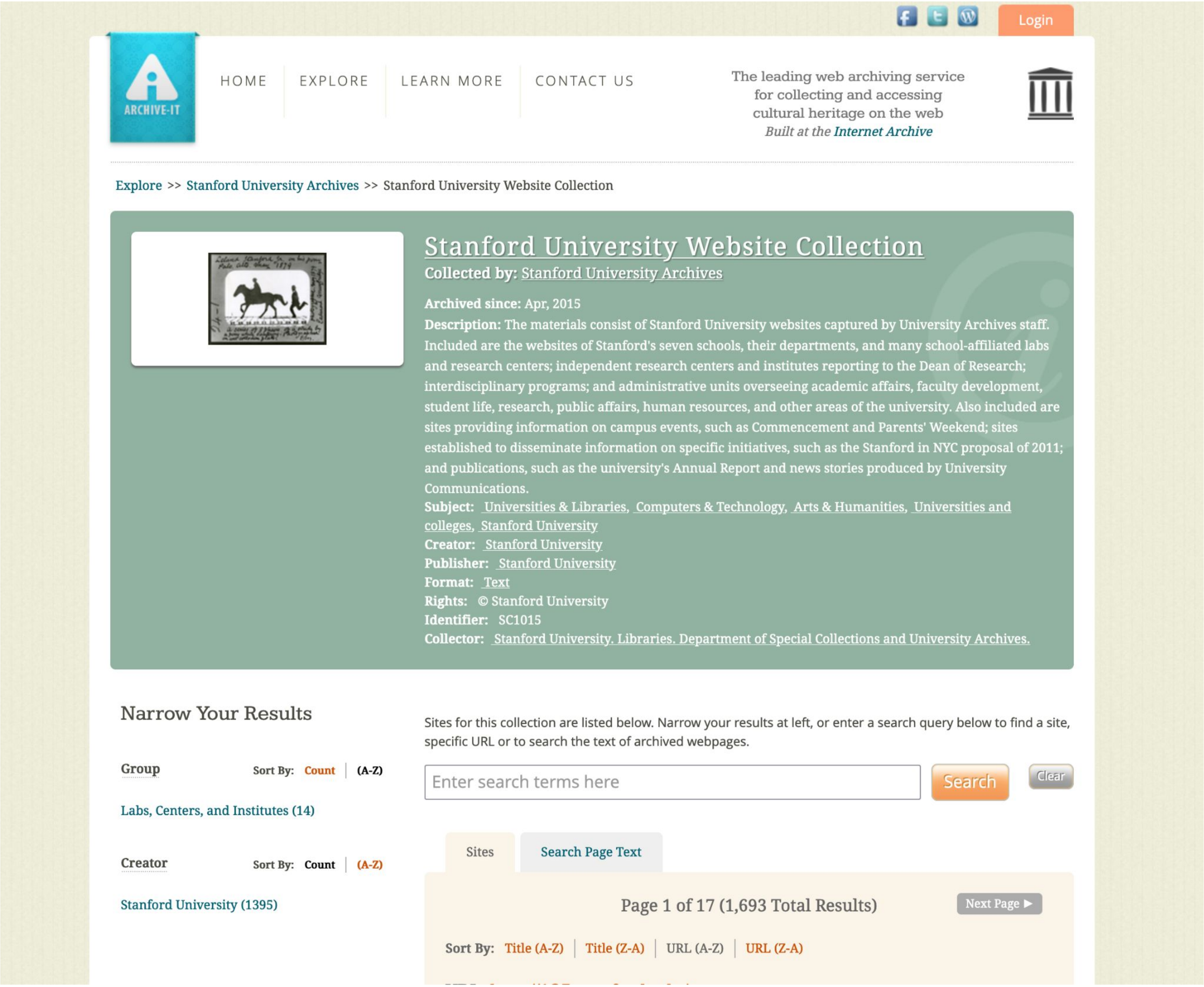
Alan Light. Tulips. 2009 Flickr.under CC BY 2.0

Hefin Owen. Daffodil bulbs and tulips. 2021. Flickr under CC BY-SA 2.0

Stanford Web Archives

- Started in 2012 with a grant from the university
- Crawl and capture:
    - California Digital Library Web Archiving Service (WAS)
    - IA's Archive-It
    - Heritrix
    - HTTrack
    - wget

The Miriam and Ira D. Wallach Division of Art, Prints and Photographs: Print Collection, The New York Public Library. "A group of tulips." *The New York Public Library Digital Collections*. 1807.



The leading web archiving service
for collecting and accessing
cultural heritage on the web
Built at the Internet Archive

Explore >> Stanford University Archives >> Stanford University Website Collection

Stanford University Website Collection

Collected by: Stanford University Archives

Archived since: Apr, 2015

Description: The materials consist of Stanford University websites captured by University Archives staff. Included are the websites of Stanford's seven schools, their departments, and many school-affiliated labs and research centers; independent research centers and institutes reporting to the Dean of Research; interdisciplinary programs; and administrative units overseeing academic affairs, faculty development, student life, research, public affairs, human resources, and other areas of the university. Also included are sites providing information on campus events, such as Commencement and Parents' Weekend; sites established to disseminate information on specific initiatives, such as the Stanford in NYC proposal of 2011; and publications, such as the university's Annual Report and news stories produced by University Communications.

Subject: Universities & Libraries, Computers & Technology, Arts & Humanities, Universities and colleges, Stanford University

Creator: Stanford University

Publisher: Stanford University

Format: Text

Rights: © Stanford University

Identifier: SC1015

Collector: Stanford University. Libraries. Department of Special Collections and University Archives.

Narrow Your Results

Group Sort By: Count | (A-Z)

Labs, Centers, and Institutes (14)

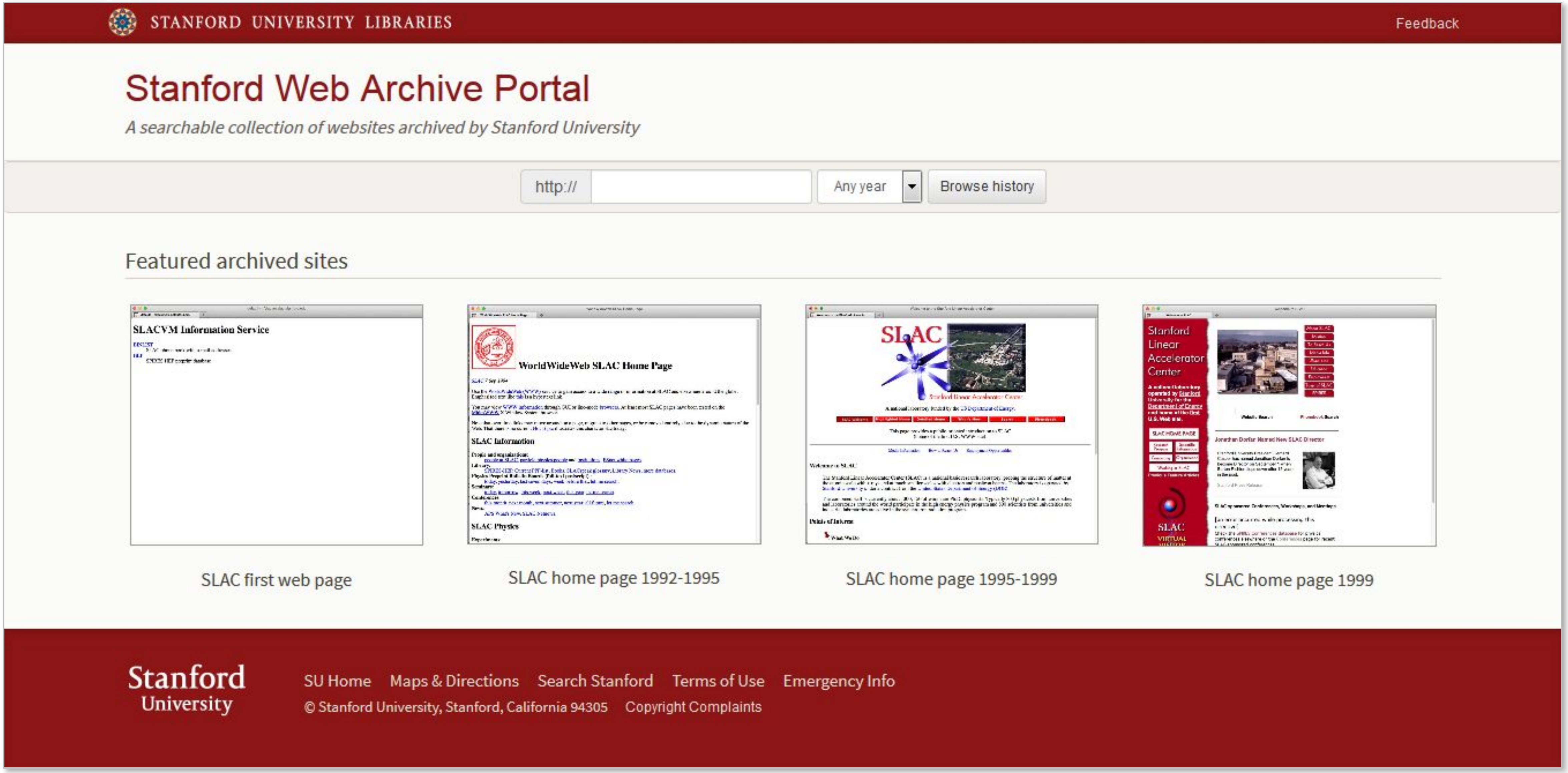
Creator Sort By: Count | (A-Z)

Stanford University (1395)

Sites for this collection are listed below. Narrow your results at left, or enter a search query below to find a site, specific URL or to search the text of archived webpages.

Page 1 of 17 (1,693 Total Results)

Sort By: Title (A-Z) | Title (Z-A) | URL (A-Z) | URL (Z-A)



STANFORD UNIVERSITY LIBRARIES

Feedback

Stanford Web Archive Portal

A searchable collection of websites archived by Stanford University

http://
Any year
Browse history

Featured archived sites

SLACVM Information Service
WorldWideWeb SLAC Home Page
SLAC Information
SLAC Physics
SLAC
Stanford Linear Accelerator Center

SLAC first web page

SLAC home page 1992-1995

SLAC home page 1995-1999

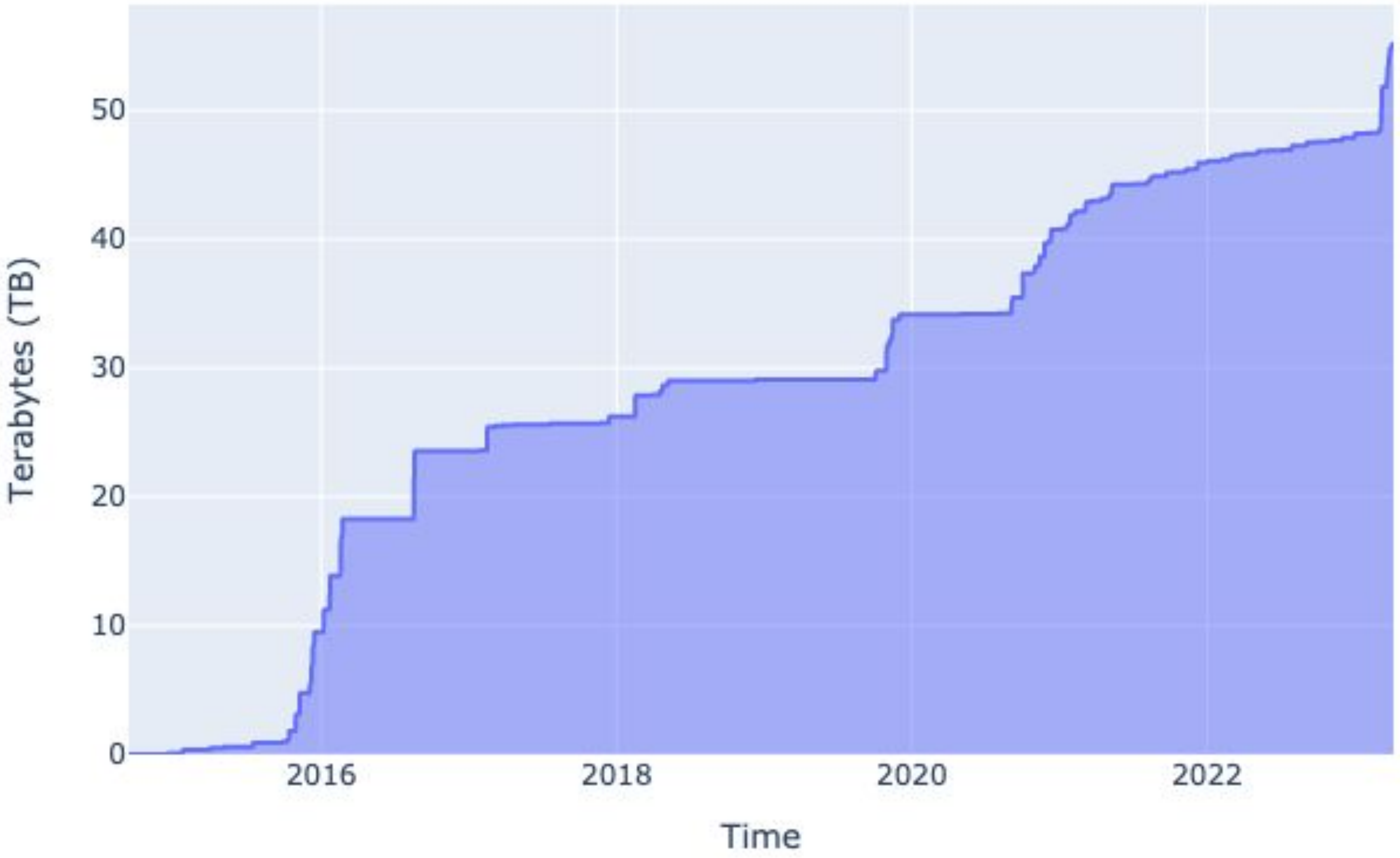
SLAC home page 1999

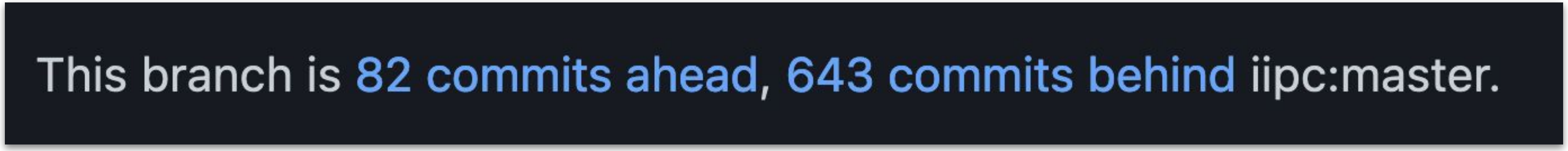
Stanford University

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

© Stanford University, Stanford, California 94305 Copyright Complaints

SUL Web Archives Storage Growth

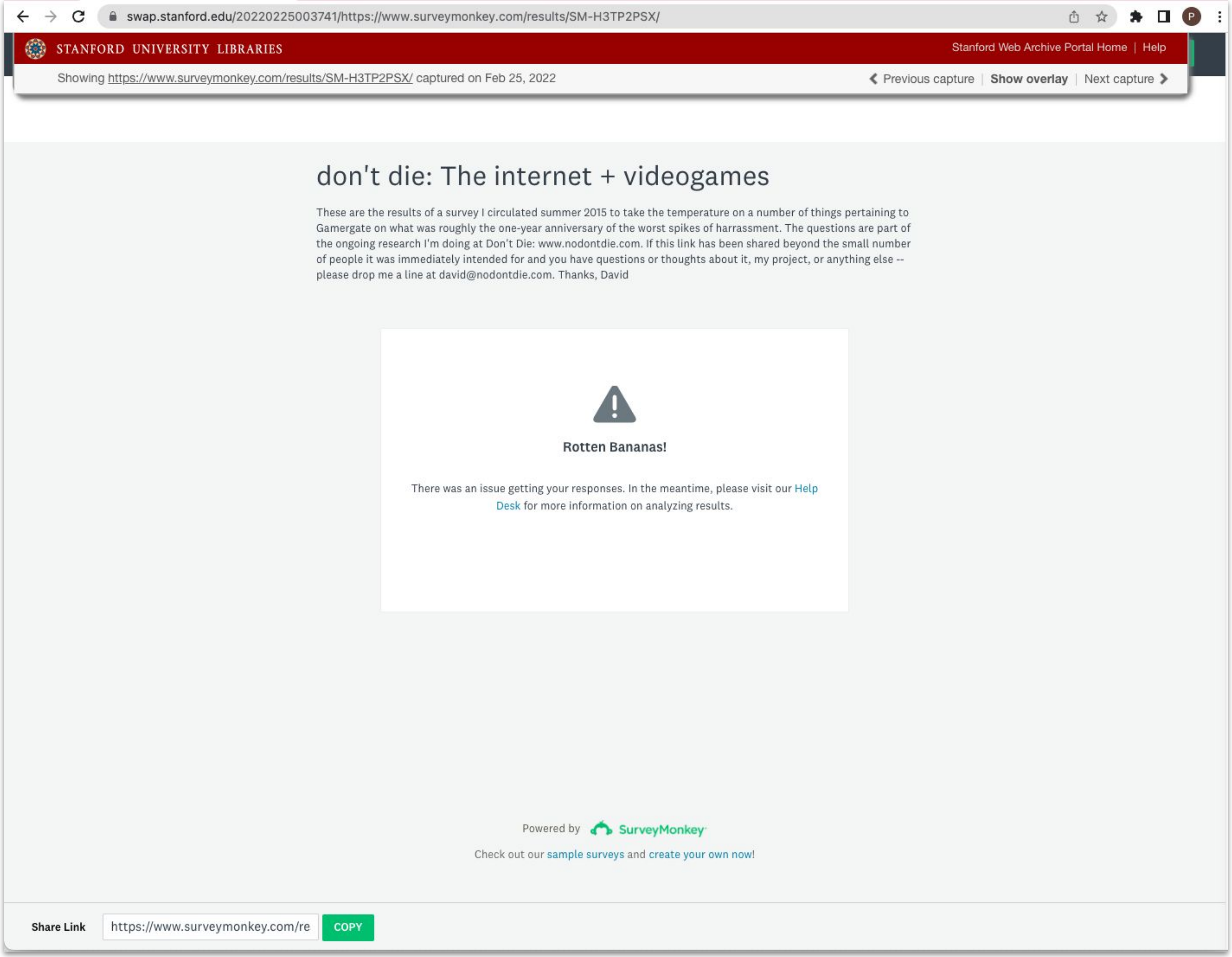




This branch is 82 commits ahead, 643 commits behind iipc:master.

Why switch to pywb?

Herford, Oliver. Overheard in a Garden. New York: Scribner and Sons, 1900. Wisconsin Digital Library



swap.stanford.edu/20220225003741/https://www.surveymonkey.com/results/SM-H3TP2PSX/

STANFORD UNIVERSITY LIBRARIES

Stanford Web Archive Portal Home | Help

Showing https://www.surveymonkey.com/results/SM-H3TP2PSX/ captured on Feb 25, 2022

‹ Previous capture | Show overlay | Next capture ›

don't die: The internet + videogames

These are the results of a survey I circulated summer 2015 to take the temperature on a number of things pertaining to Gamergate on what was roughly the one-year anniversary of the worst spikes of harrassment. The questions are part of the ongoing research I'm doing at Don't Die: www.nodontdie.com. If this link has been shared beyond the small number of people it was immediately intended for and you have questions or thoughts about it, my project, or anything else -- please drop me a line at david@nodontdie.com. Thanks, David

⚠

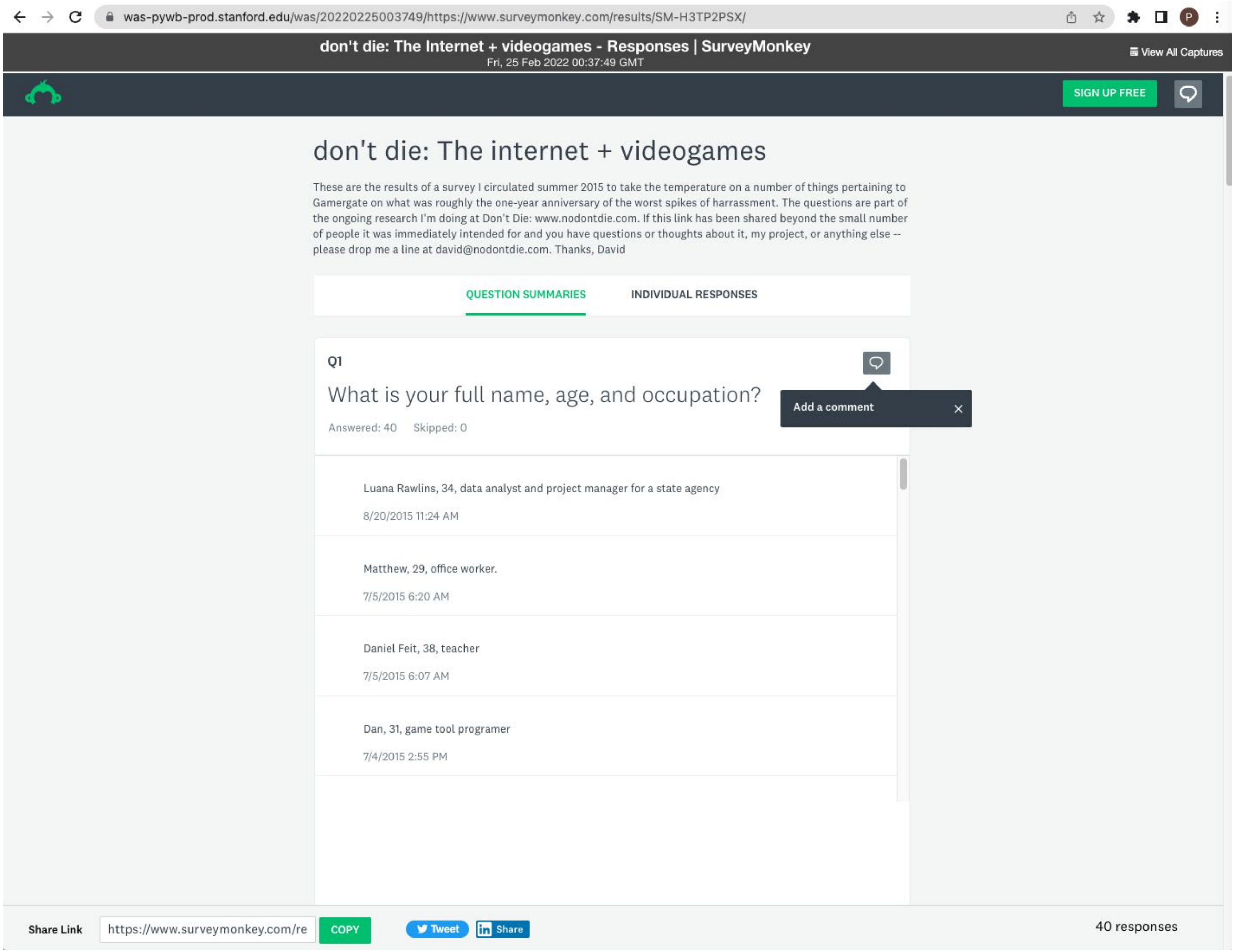
Rotten Bananas!

There was an issue getting your responses. In the meantime, please visit our Help Desk for more information on analyzing results.

Powered by SurveyMonkey

Check out our sample surveys and create your own now!

Share Link https://www.surveymonkey.com/re COPY



was-pywb-prod.stanford.edu/was/20220225003749/https://www.surveymonkey.com/results/SM-H3TP2PSX/
don't die: The Internet + videogames - Responses | SurveyMonkey
Fri, 25 Feb 2022 00:37:49 GMT
View All Captures
SIGN UP FREE

don't die: The internet + videogames

These are the results of a survey I circulated summer 2015 to take the temperature on a number of things pertaining to Gamergate on what was roughly the one-year anniversary of the worst spikes of harrassment. The questions are part of the ongoing research I'm doing at Don't Die: www.nodontdie.com. If this link has been shared beyond the small number of people it was immediately intended for and you have questions or thoughts about it, my project, or anything else -- please drop me a line at david@nodontdie.com. Thanks, David

Q1

What is your full name, age, and occupation?

Answered: 40 Skipped: 0

Luana Rawlins, 34, data analyst and project manager for a state agency

8/20/2015 11:24 AM

Matthew, 29, office worker.

7/5/2015 6:20 AM

Daniel Feit, 38, teacher

7/5/2015 6:07 AM

Dan, 31, game tool programer

7/4/2015 2:55 PM

40 responses

How did we switch?

Frédérique Voisin-Demery, "Préparer le printemps", 2012. Flickr under CC BY 2.0.

Poetry

vmware

Capistrano

docker

ubuntu

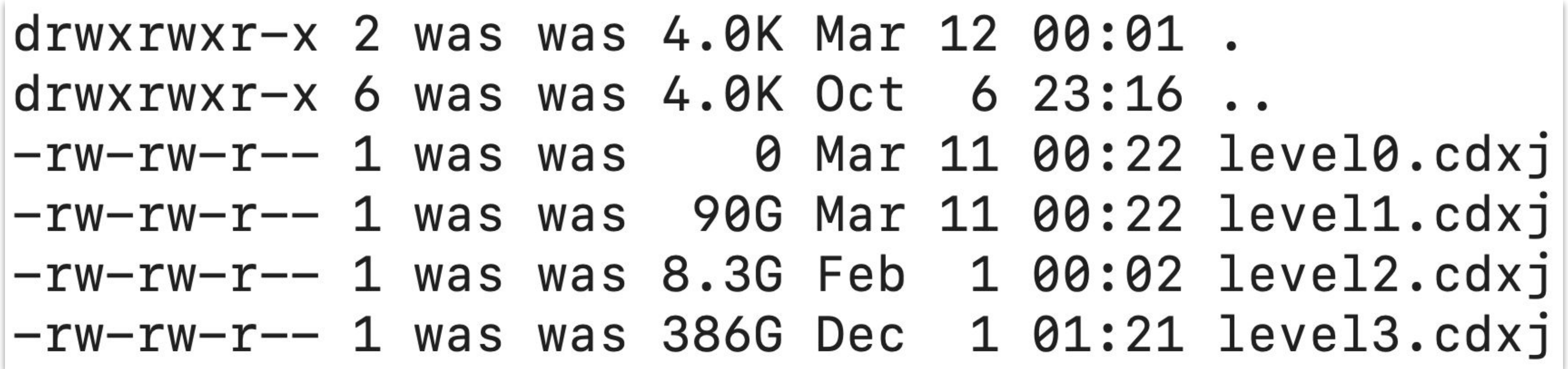
4 Intel Xeon
2.9 GHz CPUs
8 GB RAM

NFS

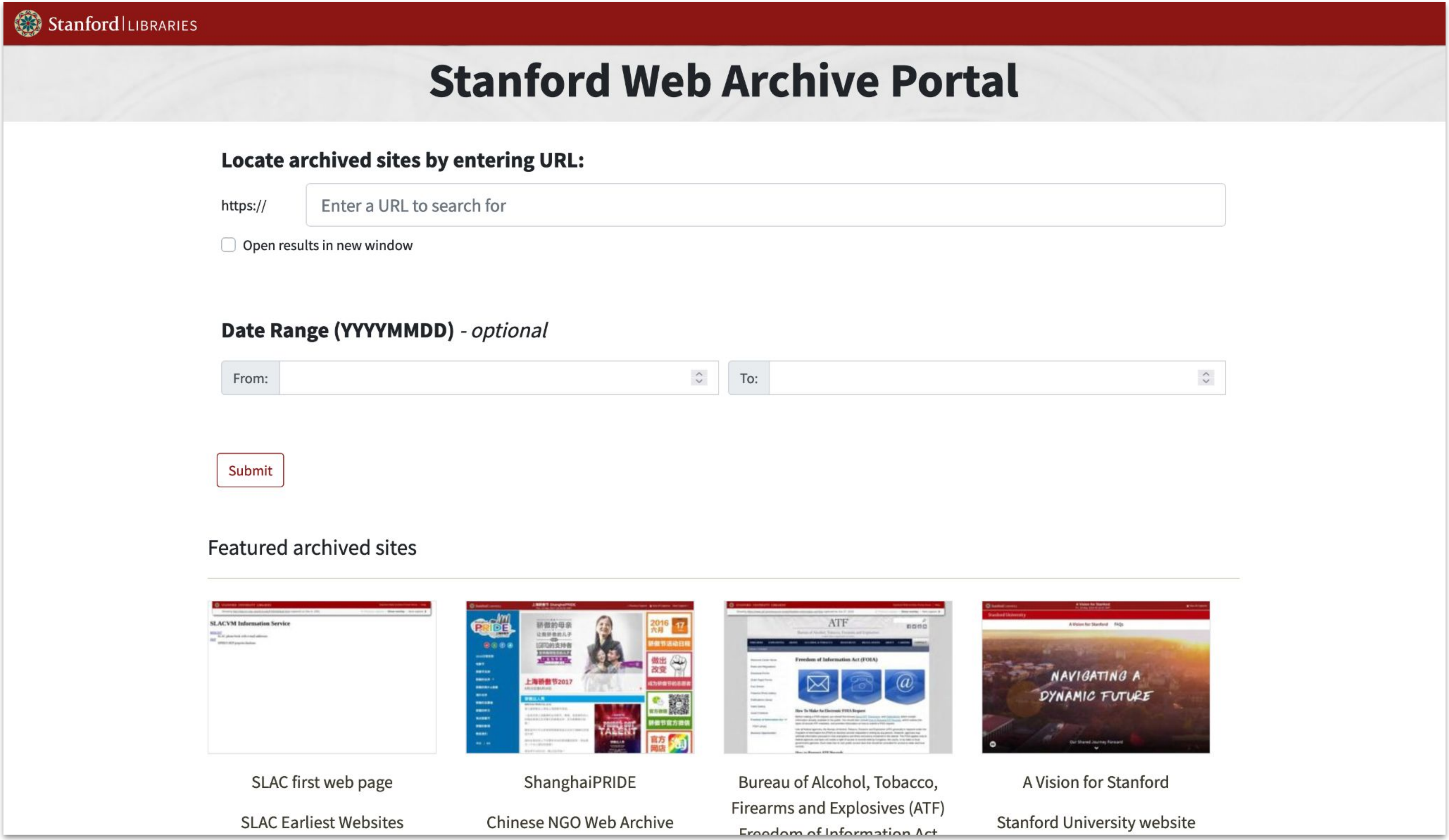
NetApp®

Reindexing the WARCs

- Used **webrecorder/cdxj-indexer** and small coordinating Ruby program to reindex the 50TB of WARC and ARC data (~5 days, 10 CPUs)
- We wanted to take advantage of the latest playback features for dynamic content so we used **--post-append**
- Extra fields in CDXJ were not compatible with **OutbackCDX** so we continued to use the uncompressed CDXJ files instead.
- Index "rollup" strategy:



```
drwxrwxr-x 2 was was 4.0K Mar 12 00:01 .
drwxrwxr-x 6 was was 4.0K Oct  6 23:16 ..
-rw-rw-r-- 1 was was    0 Mar 11 00:22 level0.cdxj
-rw-rw-r-- 1 was was  90G Mar 11 00:22 level1.cdxj
-rw-rw-r-- 1 was was 8.3G Feb  1 00:02 level2.cdxj
-rw-rw-r-- 1 was was 386G Dec  1 01:21 level3.cdxj
```



Stanford | LIBRARIES

Stanford Web Archive Portal

Locate archived sites by entering URL:

https:// Enter a URL to search for

□ Open results in new window

Date Range (YYYYMMDD) *- optional*

From:

To:

Submit

Featured archived sites

SLAC first web page

SLAC Earliest Websites

ShanghaiPRIDE

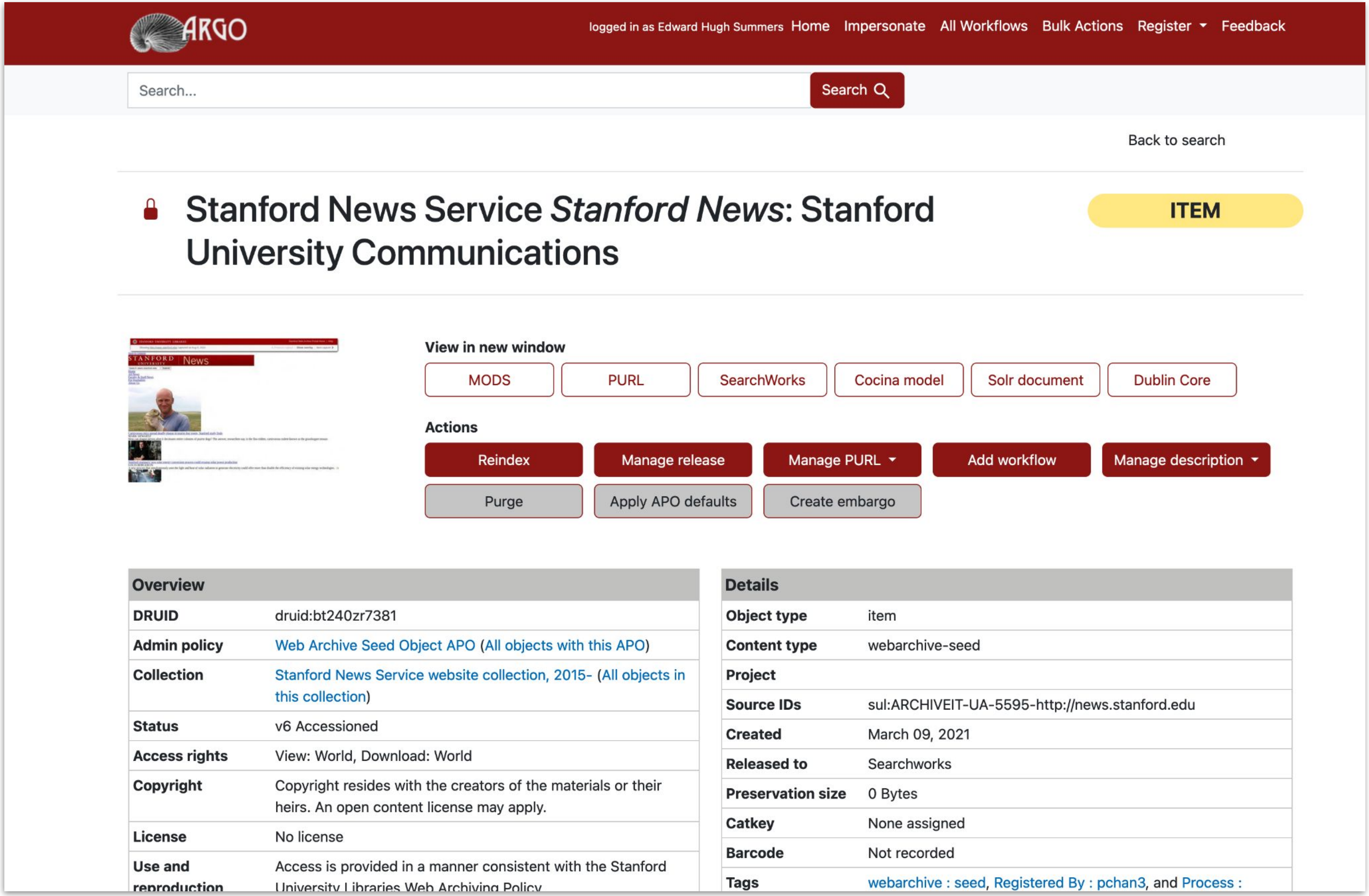
Chinese NGO Web Archive

Bureau of Alcohol, Tobacco, Firearms and Explosives (ATF)

Freedom of Information Act

A Vision for Stanford

Stanford University website

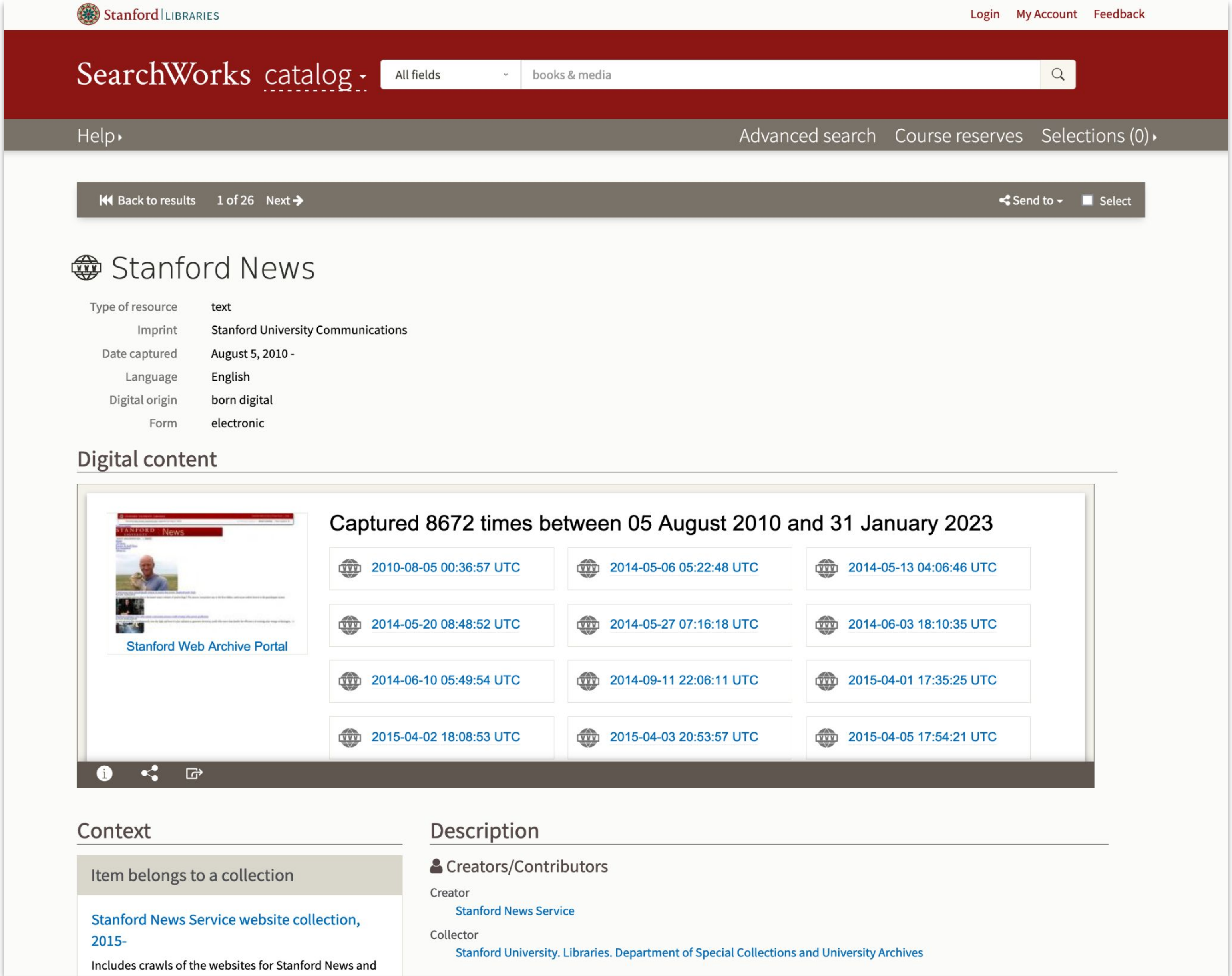


Stanford News Service *Stanford News*: Stanford University Communications

Overview	
DRUID	druid:bt240zr7381
Admin policy	Web Archive Seed Object APO (All objects with this APO)
Collection	Stanford News Service website collection, 2015- (All objects in this collection)
Status	v6 Accessioned
Access rights	View: World, Download: World
Copyright	Copyright resides with the creators of the materials or their heirs. An open content license may apply.
License	No license
Use and reproduction	Access is provided in a manner consistent with the Stanford University Libraries Web Archiving Policy

Details	
Object type	item
Content type	webarchive-seed
Project	
Source IDs	sul:ARCHIVEIT-UA-5595-http://news.stanford.edu
Created	March 09, 2021
Released to	Searchworks
Preservation size	0 Bytes
Catkey	None assigned
Barcode	Not recorded
Tags	webarchive : seed, Registered By : pchan3, and Process :

Memento API



Stanford News

Type of resource	text
Imprint	Stanford University Communications
Date captured	August 5, 2010 -
Language	English
Digital origin	born digital
Form	electronic

Digital content

Captured 8672 times between 05 August 2010 and 31 January 2023

2010-08-05 00:36:57 UTC	2014-05-06 05:22:48 UTC	2014-05-13 04:06:46 UTC
2014-05-20 08:48:52 UTC	2014-05-27 07:16:18 UTC	2014-06-03 18:10:35 UTC
2014-06-10 05:49:54 UTC	2014-09-11 22:06:11 UTC	2015-04-01 17:35:25 UTC
2015-04-02 18:08:53 UTC	2015-04-03 20:53:57 UTC	2015-04-05 17:54:21 UTC

Context

Item belongs to a collection

Stanford News Service website collection, 2015-

Includes crawls of the websites for Stanford News and

Description

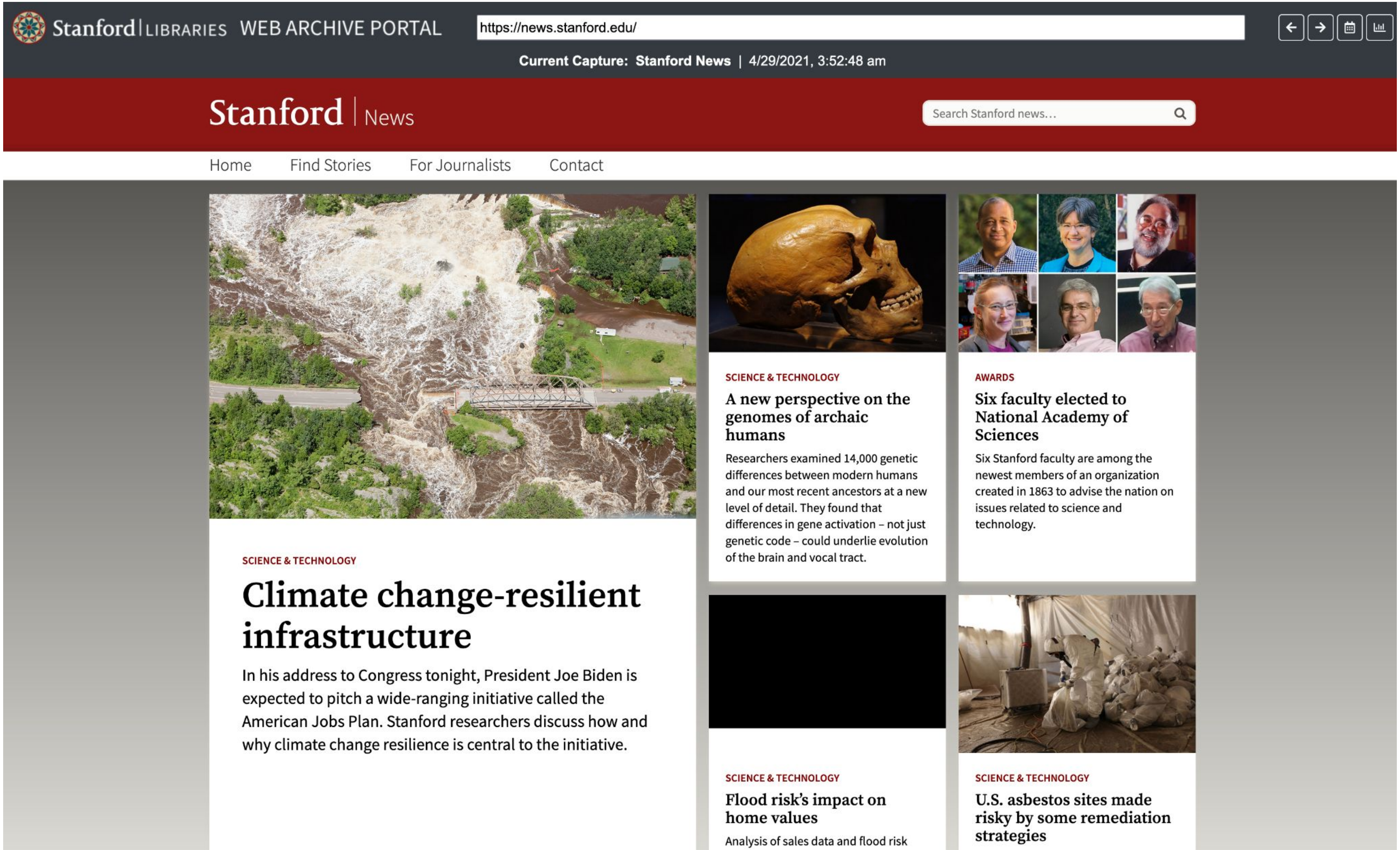
Creators/Contributors

Creator

Stanford News Service

Collector

Stanford University. Libraries. Department of Special Collections and University Archives



https://news.stanford.edu/

Current Capture: Stanford News | 4/29/2021, 3:52:48 am

Stanford | News

Search Stanford news…

Home Find Stories For Journalists Contact

SCIENCE & TECHNOLOGY

Climate change-resilient infrastructure

In his address to Congress tonight, President Joe Biden is expected to pitch a wide-ranging initiative called the American Jobs Plan. Stanford researchers discuss how and why climate change resilience is central to the initiative.

SCIENCE & TECHNOLOGY

A new perspective on the genomes of archaic humans

Researchers examined 14,000 genetic differences between modern humans and our most recent ancestors at a new level of detail. They found that differences in gene activation – not just genetic code – could underlie evolution of the brain and vocal tract.

AWARDS

Six faculty elected to National Academy of Sciences

Six Stanford faculty are among the newest members of an organization created in 1863 to advise the nation on issues related to science and technology.

SCIENCE & TECHNOLOGY

Flood risk's impact on home values

Analysis of sales data and flood risk

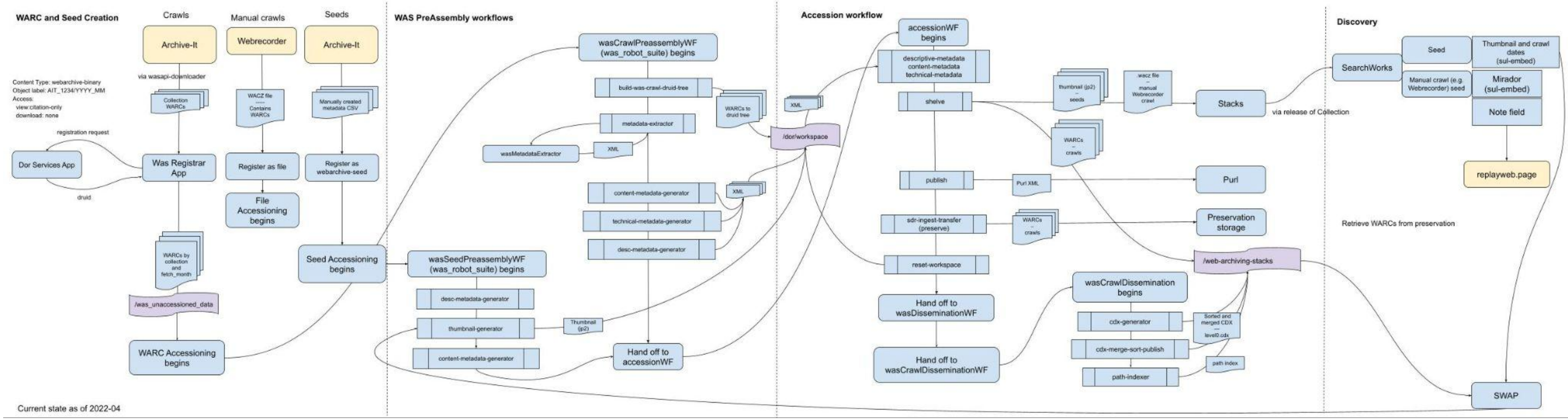
SCIENCE & TECHNOLOGY

U.S. asbestos sites made risky by some remediation strategies

Stanford | LIBRARIES

Repository workflows

Repository workflows



WARC and Seed Creation

Crawls

Manual crawls

Seeds

Archive-It

Webrecorder

Archive-It

via wasapi-downloader

Content Type: webarchive-binary
Object label: AIT_1234/YYYY_MM
Access:
view:citation-only
download: none

Collection WARCs

WACZ file

Contains WARCs

Manually created metadata CSV

registration request

Dor Services App

druid

Was Registrar App

Register as file

Register as webarchive-seed

File Accessioning begins

WARCs by collection and fetch_month

/was_unaccessioned_data

Seed Accessioning begins

WARC Accessioning begins

Current state as of 2022-04

WAS PreAssembly workflows

wasCrawlPreassemblyWF (was_robot_suite) begins

build-was-crawl-druid-tree

WARCs to druid tree

metadata-extractor

wasMetadataExtractor

XML

content-metadata-generator

technical-metadata-generator

desc-metadata-generator

XML

wasSeedPreassemblyWF (was_robot_suite) begins

desc-metadata-generator

thumbnail-generator

Thumbnail (jp2)

content-metadata-generator

Hand off to accessionWF

Accession workflow

XML

/dor/workspace

accessionWF begins

descriptive-metadata
content-metadata
technical-metadata

shelve

thumbnail (jp2)
–
seeds

wacz file
–
manual Webrecorder crawl

Stacks

via release of Collection

WARCs
–
crawls

publish

Purl XML

Purl

sdr-ingest-transfer (preserve)

WARCs
–
crawls

Preservation storage

reset-workspace

/web-archiving-stacks

Hand off to wasDisseminationWF

wasCrawlDissemination begins

cdx-generator

Sorted and merged CDX
—
level0.cdx

cdx-merge-sort-publish

path-indexer

path index

Hand off to wasCrawlDisseminationWF

Discovery

SearchWorks

Seed

Thumbnail and crawl dates (sul-embed)

Manual crawl (e.g. Webrecorder) seed

Mirador (sul-embed)

Note field

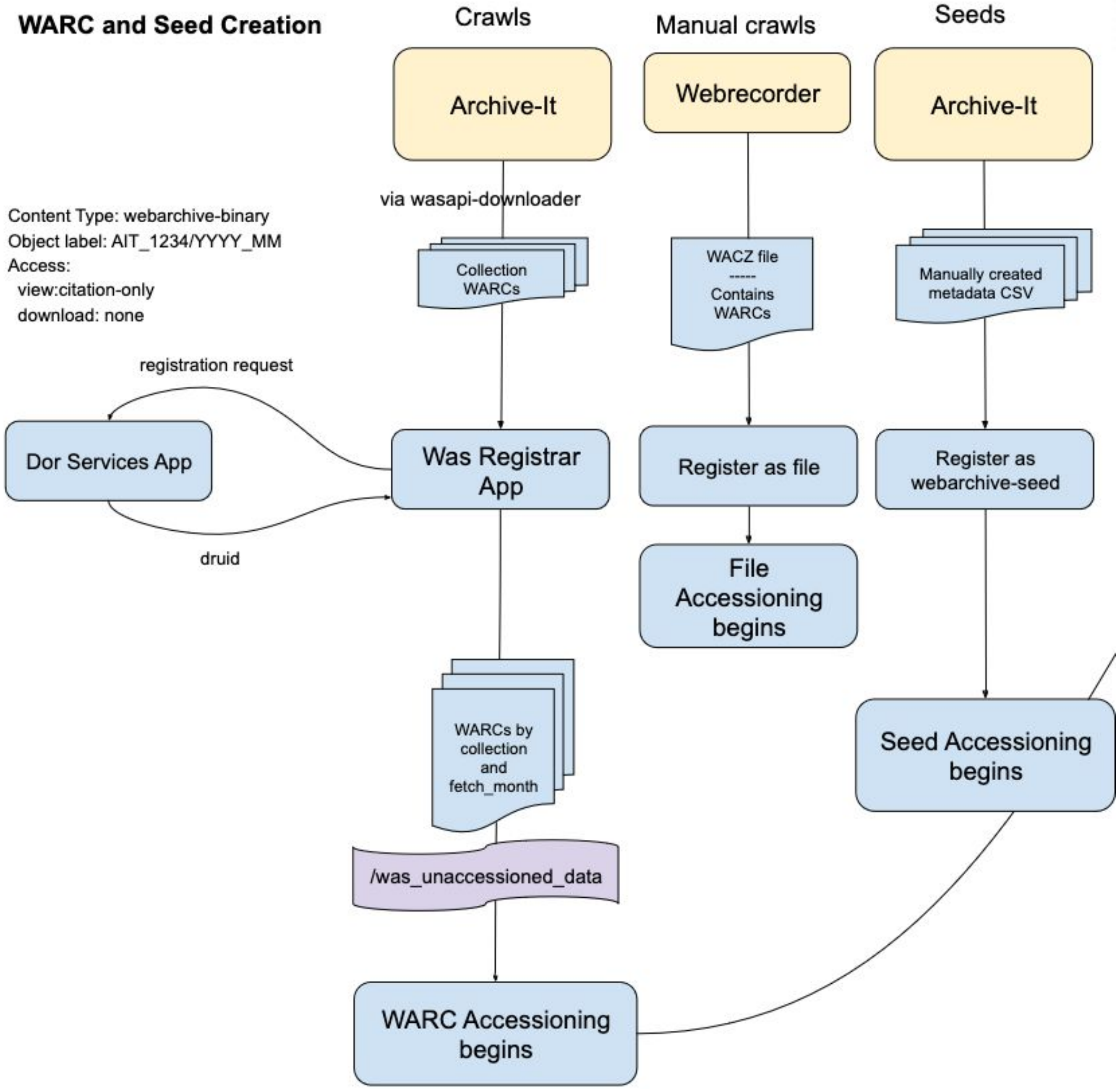
replayweb.page

Retrieve WARCs from preservation

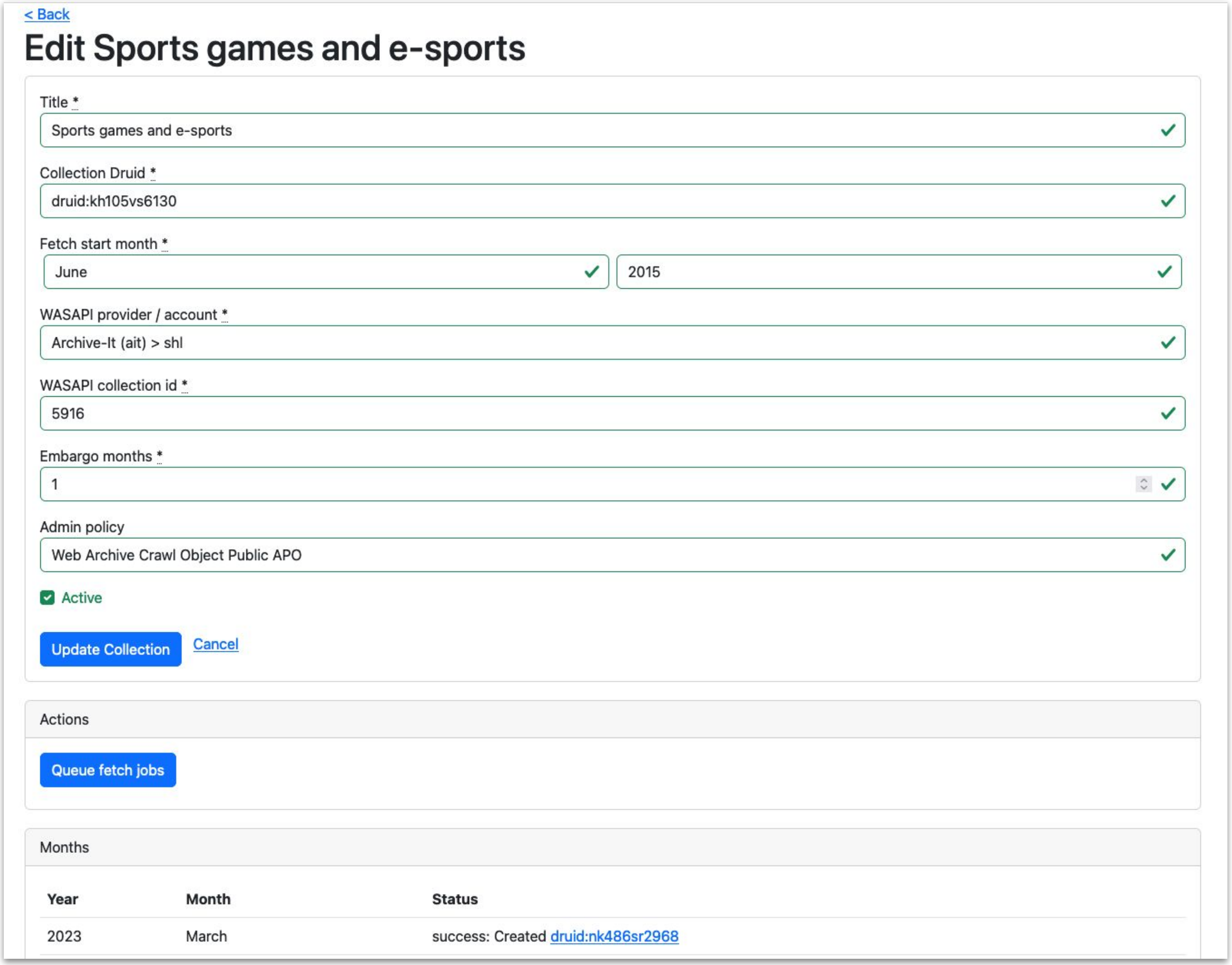
SWAP

WARC and Seed Creation

Content Type: webarchive-binary
Object label: AIT_1234/YYYY_MM
Access:
view:citation-only
download: none



WAS Registrar App



< Back

Edit Sports games and e-sports

Title *

Sports games and e-sports ✓

Collection Druid *

druid:kh105vs6130 ✓

Fetch start month *

June ✓ 2015 ✓

WASAPI provider / account *

Archive-It (ait) > shl ✓

WASAPI collection id *

5916 ✓

Embargo months *

1 ✓

Admin policy

Web Archive Crawl Object Public APO ✓

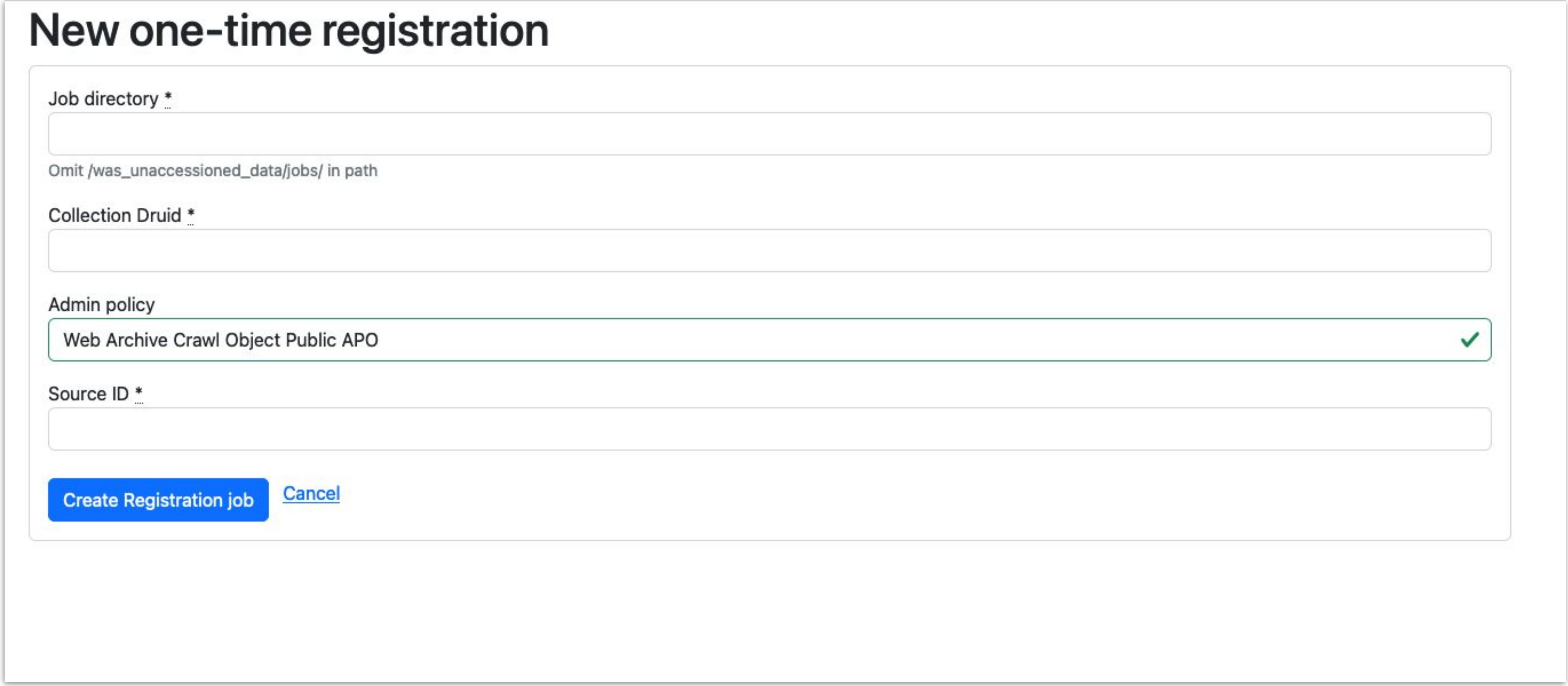
☑ Active

Actions

Months

Year	Month	Status
2023	March	success: Created druid:nk486sr2968

WAS Registrar App: registering a WACZ



New one-time registration

Job directory *

Omit /was_unaccessioned_data/jobs/ in path

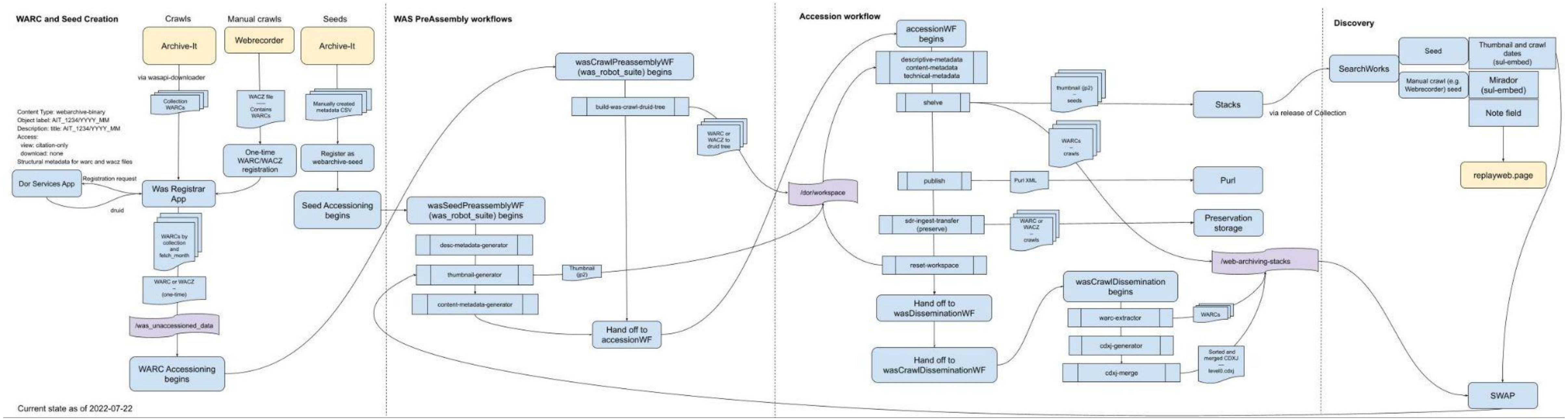
Collection Druid *

Admin policy

Web Archive Crawl Object Public APO ✓

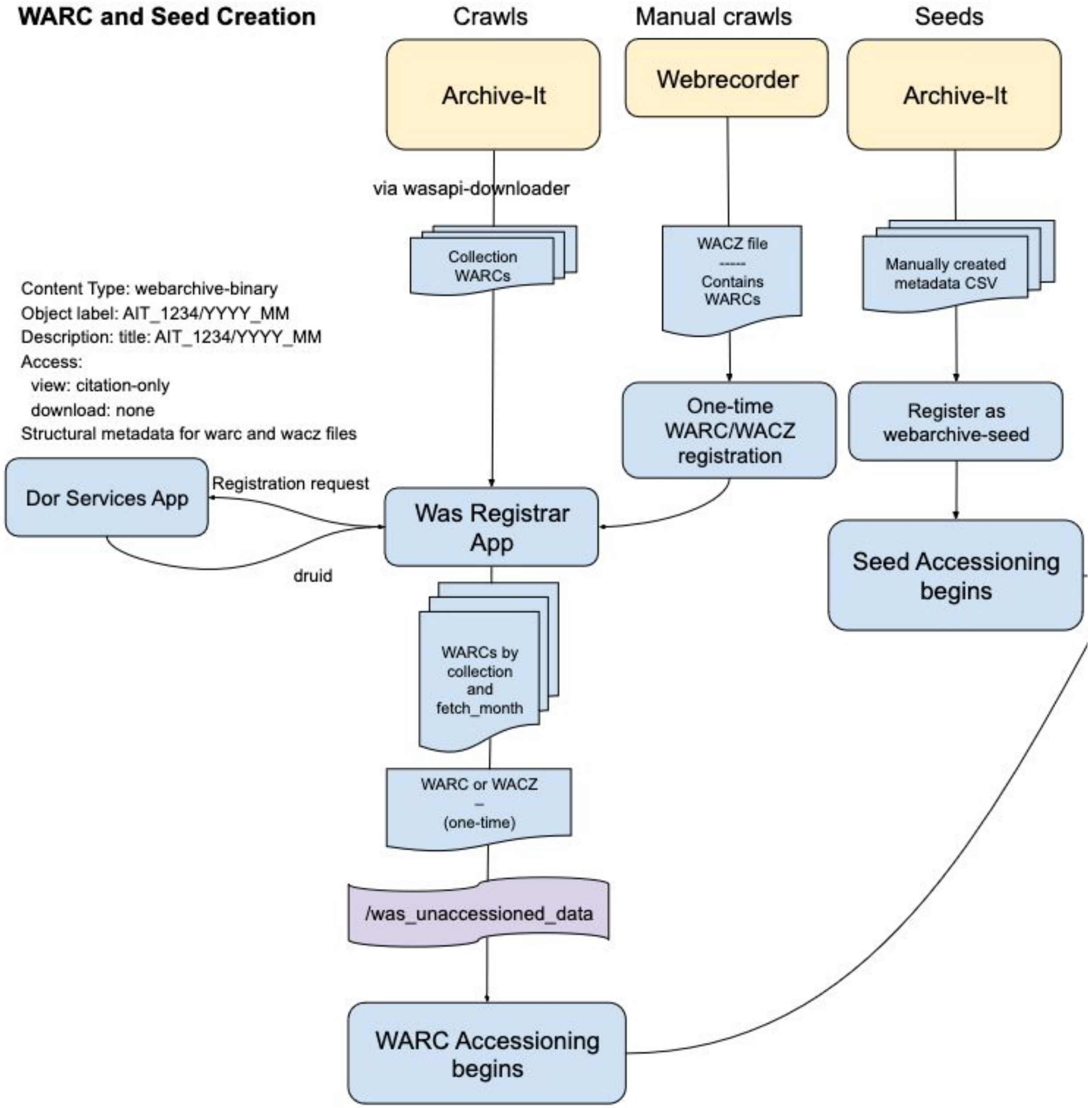
Source ID *

Updated repository workflows


WARC and Seed Creation
Crawls
Manual crawls
Seeds
Archive-It
Webrecorder
Archive-It
via wasapi-downloader
Collection WARCs
WACZ file ---- Contains WARCs
Manually created metadata CSV
Content Type: webarchive-binary
Object label: AIT_1234/YYYY_MM
Description: title: AIT_1234/YYYY_MM
Access:
view: citation-only
download: none
Structural metadata for warc and wacz files
One-time WARC/WACZ registration
Register as webarchive-seed
Dor Services App
Registration request
druid
Was Registrar App
Seed Accessioning begins
WARCs by collection and fetch_month
WARC or WACZ – (one-time)
/was_unaccessioned_data
WARC Accessioning begins
WAS PreAssembly workflows
wasCrawlPreassemblyWF (was_robot_suite) begins
build-was-crawl-druid-tree
WARC or WACZ to druid tree
wasSeedPreassemblyWF (was_robot_suite) begins
desc-metadata-generator
thumbnail-generator
Thumbnail (jp2)
content-metadata-generator
Hand off to accessionWF
Accession workflow
/dor/workspace
accessionWF begins
descriptive-metadata
content-metadata
technical-metadata
shelve
thumbnail (jp2) – seeds
WARCs – crawls
publish
Purl XML
sdr-ingest-transfer (preserve)
WARC or WACZ – crawls
reset-workspace
Hand off to wasDisseminationWF
Hand off to wasCrawlDisseminationWF
wasCrawlDissemination begins
warc-extractor
WARCs
cdxj-generator
cdxj-merge
Sorted and merged CDXJ — level0.cdxj
Stacks
via release of Collection
Purl
Preservation storage
/web-archiving-stacks
Discovery
SearchWorks
Seed
Manual crawl (e.g. Webrecorder) seed
Thumbnail and crawl dates (sul-embed)
Mirador (sul-embed)
Note field
replayweb.page
SWAP

Current state as of 2022-07-22

WARC and Seed Creation



Content Type: webarchive-binary
Object label: AIT_1234/YYYY_MM
Description: title: AIT_1234/YYYY_MM
Access:
  view: citation-only
  download: none
Structural metadata for warc and wacz files

Learning from maintenance

- Shared understanding of terminology, formats, and maintenance practices
- Particular knowledge needed to address quality assurance with replay
- Learning about our collections
  - Capture issues
  - Audit tools using WASAPI
  - SLAC: first U.S. web pages and maintaining access to the reconstruction

Smith, Sandra E. "OK...Who's Been in my Garden?"
"http://hdl.handle.net/11134/140067:219

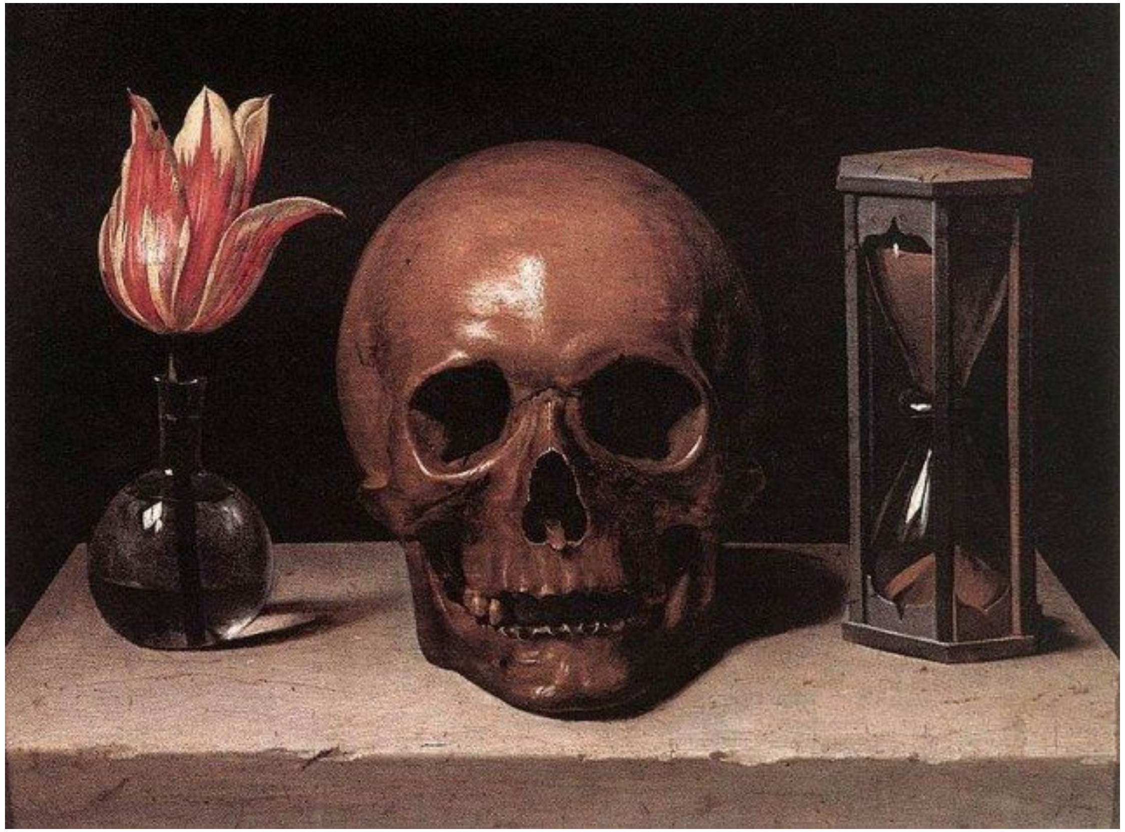
Further maintenance concerns

Oregon Department of Agriculture, "Picking tulips", 2008. Flickr under CC BY-NC-ND 2.0

- Repository policies
- Collections created by researchers
- Sharing WARCs as data
- Organizational approach (people!)

Maintenance and community

- Understanding use cases
- Collectively developing a roadmap
- Funding development
- Contributing directly to maintenance and development



Philippe de Champaigne (1602–1674). Still-Life with a Skull, vanitas painting.
https://commons.wikimedia.org/wiki/File:StillLifeWithASkull.jpg

Thank you!

Please reach out to us with questions or thoughts via email or IIPC Slack.

Laura Wrubel

lwrubel@stanford.edu

Ed Summers

edsu@stanford.edu

Adriano Aurelio Araujo, "Tulips at Keukenhof," 2015. Flickr under CC BY 2.0