

pywb @Stanford

Laura Wrubel
lwrubel@stanford.edu

Ed Summers
edsu@stanford.edu

March 22, 2023
IIPC Technical Seminar Series

Overview

1. Web archives at Stanford
2. Why switch to pywb?
3. How did we switch?
4. How did we upgrade?
5. What's next?

Stanford Web Archives

- Started in 2012 with a grant from the university
- Housed in University Archives (Peter Chan)
- Supported by Digital Library Systems and Services (DLSS)

<https://library.stanford.edu/projects/web-archiving>



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.

[Search](#) [Clear](#)

[Sites](#) [Search Page Text](#)

Page 1 of 17 (1,693 Total Results)

[Next Page](#) ▶

Sort By: [Title \(A-Z\)](#) | [Title \(Z-A\)](#) | [URL \(A-Z\)](#) | [URL \(Z-A\)](#)



Search...

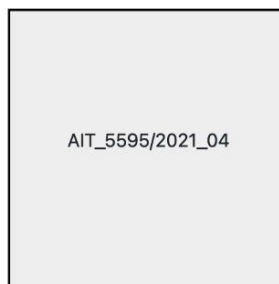
Search 

« Previous | 1 of 79 | Next »

[Back to search](#)

 **AIT_5595/2021_04**

ITEM



View in new window

- [MODS](#)
- [PURL](#)
- [SearchWorks](#)
- [Cocina model](#)
- [Solr document](#)
- [Dublin Core](#)

Actions

- [Reindex](#)
- [Manage release](#)
- [Manage PURL ▾](#)
- [Add workflow](#)
- [Manage description ▾](#)
- [Purge](#)
- [Apply APO defaults](#)
- [Create embargo](#)

Overview	
DRUID	druid:bf032vf4803
Admin policy	Web Archive Crawl Object Public APO (All objects with this APO)
Collection	Stanford News Service website collection, 2015- (All objects in this collection)
Status	v1 Accessioned
Access rights	View: Citation-only, Download: None
Copyright	Not entered
License	No license
Thumbnail	Not entered

Details	
Object type	item
Content type	file
Project	
Source IDs	sul:ait-5595-2021_04
Created	May 02, 2021
Released to	Not released
Preservation size	5.47 GB
Catkey	None assigned
Barcode	Not recorded

<https://purl.stanford.edu/bf032vf4803>



[Back to search](#)

Stanford News Service *Stanford News: Stanford University Communications*

ITEM



View in new window

- [MODS](#)
- [PURL](#)
- [SearchWorks](#)
- [Cocina model](#)
- [Solr document](#)
- [Dublin Core](#)

Actions

- [Reindex](#)
- [Manage release](#)
- [Manage PURL ▾](#)
- [Add workflow](#)
- [Manage description ▾](#)
- [Purge](#)
- [Apply APO defaults](#)
- [Create embargo](#)

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 :





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





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

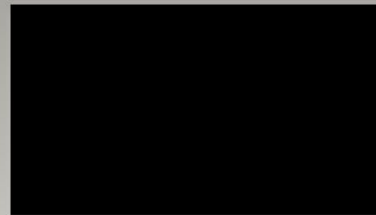
Researchers examined 14,000 genetic differences between modern humans and our most recent ancestors at a new level of detail. They found that differences in gene activation – not just genetic code – could underlie evolution of the brain and vocal tract.



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

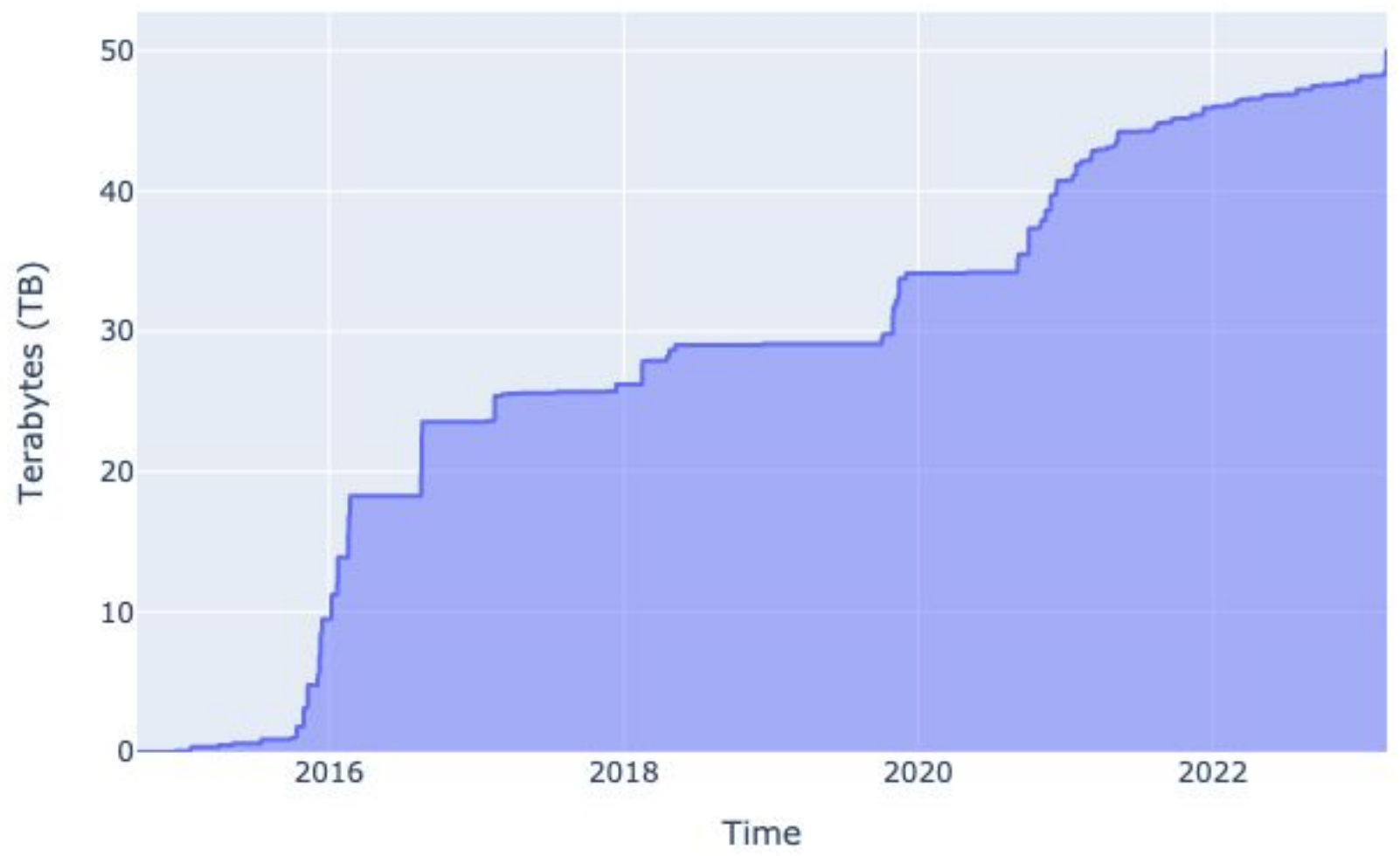


SCIENCE & TECHNOLOGY

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




SUL Web Archives Storage Growth



1,025,365,726 snapshots


Why switch to pywb?

← → ↻ 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 COPY



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 harrasment. 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

QUESTION SUMMARIES

INDIVIDUAL RESPONSES

Q1

What is your full name, age, and occupation?

Answered: 40 Skipped: 0



Add a comment X

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 programmer

7/4/2015 2:55 PM

Share Link

<https://www.surveymonkey.com/re>

COPY

Tweet

Share

40 responses

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



Search or jump to...

[Pull requests](#) [Issues](#) [Codespaces](#) [Marketplace](#) [Explore](#)



[iipc / openwayback](#) Public

[Watch](#) 64

[Fork](#) 272

[Starred](#) 440

[Code](#) [Issues](#) 100 [Pull requests](#) 5 [Actions](#) [Projects](#) [Wiki](#) [Security](#) [Insights](#)

[master](#)

Commits on Jun 25, 2021

Fix capitalization

Idko committed on Jun 25, 2021

Verified



98bbc1a



Suggest using pywb over OpenWayback

Idko committed on Jun 25, 2021

Verified



688ec80



Commits on Oct 13, 2020

Merge pull request #435 from iipc/dependabot/maven/junit-junit-4.13.1 ...

Idko committed on Oct 13, 2020 ✓

Verified



c7fac11



Bump junit from 3.8.1 to 4.13.1 ...

dependabot[bot] committed on Oct 13, 2020 ✓

Verified



a165558



Commits on Sep 28, 2020



How did we switch?



Search or jump to...



Pull requests Issues Codespaces Marketplace Explore



Public sul-dlss / was-pywb

Edit Pins

Unwatch 12

Fork 0

Star 2

Code

Issues 18

Pull requests

Discussions

Actions

Wiki

Security

Insights

Settings

main

1 branch

44 tags

Go to file

Add file

Code

About



Configuration Stanford's pywb instance

swap.stanford.edu

Readme

View license

2 stars

12 watching

0 forks

Releases

44 tags

Create a new release



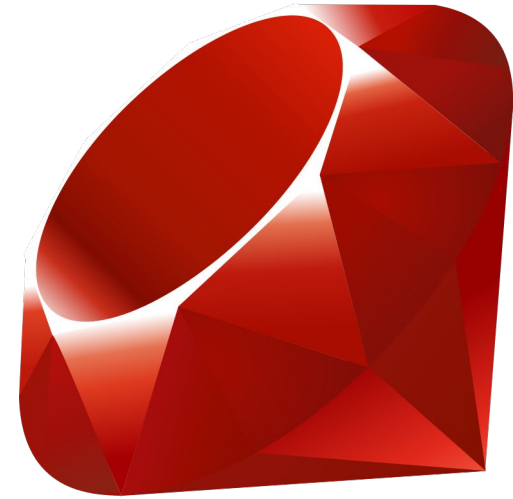
lwrubel Merge pull request #176 from sul-dl... ad601e3 4 days ago 216 commits

.autoupdate	Bring parity to was-pywb CI hook	5 months ago
.circleci	Update CircleCI orb	last month
.github	Adding PR template	8 months ago
bin	Set up RSpec and rubocop	9 months ago
config	ensure pip upgrades poetry in deploy	6 months ago
lib	Fix test behavior when cdxj temp file is not pres...	8 months ago
pywb	Update Python dependencies	4 days ago
spec	Set up RSpec and rubocop	9 months ago
test-data	More test data	9 months ago





Poetry



Capistrano



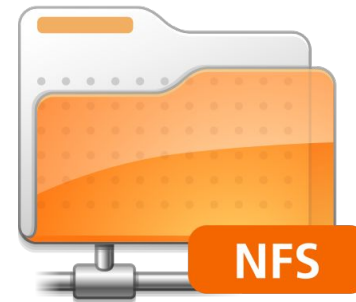
vmware®



docker

ubuntu®

**4 Intel Xeon
2.9 GHz CPUs
8 GB RAM**



NetApp®



Reindexing the WARC

- 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 “Rollups”

```
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
```

config.yaml

```
1   collections:
2     was:
3       index_paths: /web-archiving-stacks/data/indexes/cdxj/
4       archive_paths:
5         - /web-archiving-stacks/data/collections/
6       acl_paths: /web-archiving-stacks/data/access.aclj
7
```

ChineseRailroadWorkers
in North America Project

北美鐵路華工研究工程
宗旨

HOME

KEY QUESTIONS

PUBLICATIONS

RESOURCES

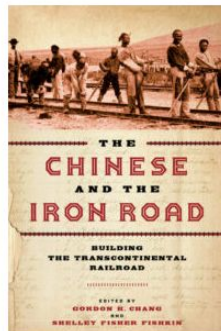
UPDATES

PRESS

PROJECT TEAM

中文

Search ...



©2019 Chinese Railroad Workers in North America Project at Stanford University. All Rights Reserved.

Chinese Railroad Workers in North America Project at Stanford University



Announcement

August 31, 2020

The [Chinese Railroad Workers in North America Project](#) at Stanford University has successfully completed its study, research, and writing. In 2012 when we began our work, we knew we faced formidable research challenges, but the dedicated and determined work of hundreds of scholars, students, and volunteers from around the world helped us recover Chinese railroad worker history unprecedented in richness and comprehensiveness. Our interpretations of culture and scholarship provided original insights into that central experience in Chinese American history and our efforts continue to attract international attention and inspire further efforts to see that the workers receive their due recognition. We are enormously gratified by the warm reception we have received.

Stanford Web Archive Portal

Locate archived sites by entering URL:

https://

Open results in new window

Date Range (YYYYMMDD) - optional

From: To:

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

Memento API













The screenshot shows the Stanford SearchWorks catalog interface. At the top, there's a search bar with "All fields" and "books & media" selected. Below the search bar, there's a navigation bar with "Help", "Advanced search", "Course reserves", and "Select". The main content area displays a record for "Stanford News".

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://swap.stanford.edu/timemap/http://news.stanford.edu/>

How did we upgrade to 2.7?



Search or jump to...

Pull requests Issues Codespaces Marketplace Explore



Public sul-dlss / was-pywb

Edit Pins

Unwatch 12

Fork 0

Star 2

Code Issues 18 Pull requests Discussions Actions Wiki Security Insights Settings

Update dependencies #176

Edit Code

Merged lwrubel merged 2 commits into main from update-dependencies 4 days ago

Conversation 1 Commits 2 Checks 0 Files changed 2 +116 -122



Member sul-devops-team commented 4 days ago

No description provided.

Member sul-devops-team added 2 commits 4 days ago

- Update Python dependencies e34d171
- Update Ruby dependencies bbac04e



lwrubel approved these changes 4 days ago

View changes

Reviewers

lwrubel ✓

Still in progress? Learn about draft PRs

Assignees

No one—assign yourself

Labels

None yet

Next / Previous Snapshots



Stanford LIBRARIES WEB ARCHIVE PORTAL

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

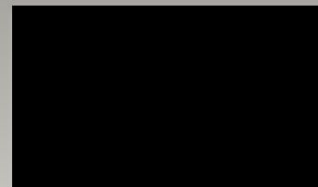
Researchers examined 14,000 genetic differences between modern humans and our most recent ancestors at a new level of detail. They found that differences in gene activation – not just genetic code – could underlie evolution of the brain and vocal tract.



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

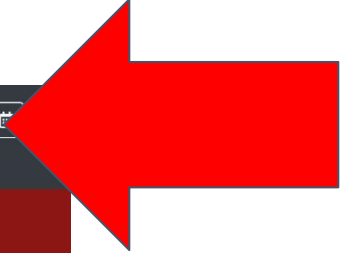
Flood risk's impact on home values

Analysis of sales data and flood risk



SCIENCE & TECHNOLOGY

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



Pywb 2.7 customization and configuration

- Removed our customized banner.html
- Created a new stanford_header.html for the collection search page (top page).
- Removed our customized frame_insert.html
- Removed our custom stylesheets
- New ui config.yaml options for logo and colors

```
8     ui:  
9         logo: images/swaplogo.png  
10        navbar_background_hex: 343a40  
11        navbar_color_hex: fffffff  
12        navbar_light_buttons: fffffff
```

Contributed to community testing of Pywb 2.7 beta

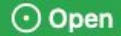
What's next?

Unresolved Issues

- Being able to remove crawls from replay (e.g. bad capture, PII)

Dates in Access Control Lists

Add date ranges to access control lists #703



anjackson opened this issue on Mar 30, 2022 · 1 comment



anjackson commented on Mar 30, 2022

Contributor



Is your feature request related to a problem? Please describe.

We would like to refine our URL blocks, by specifying a time range for the block, we we can limit access to a subset of the snapshots of URLs, rather than blocking the whole URL for all time.

Describe the solution you'd like

Support for a syntax like the `embargo` syntax, e.g. these two statements (which would have similar effects!):

```
org,httpbin)/ - {"access": "block", "url": "httpbin.org/", "before": "20201226"}
org,httpbin)/anything/something - {"access": "allow", "url": "http://httpbin.org/anything/something", "after":'}
```

Describe alternatives you've considered

We could delete CDX records but we don't want to block the same URLs across all access contexts.

Dates in Access Control Lists

ACL support for timestamp & date range #825



VascoRatoFCCN opened this issue 3 minutes ago · 0 comments



VascoRatoFCCN commented 3 minutes ago



We would like to be able to specify timestamps or date ranges on the `.aclj` file. The [documentation](#) already suggests that this feature will eventually be implemented:

The prefix consists of a SURT key and a `-` (currently reserved for a timestamp/date range field to be added later).

We require this feature to answer removal requests, which often only implies blacklisting a SURT for a specific timestamp or a small date range.

Unresolved Issues

- Being able to remove crawls from replay (e.g. bad capture, PII)
- Performance issues related to indexing: index segments with large number of fuzzy match candidates (e.g. googlevideo.com)
- **--post-append** output not supported in OutbackCDX
- Replay differences between Pywb and ReplayWebPage: parity in URL rewriting, fuzzy matching, etc.

More Ongoing Work

- Understanding researcher needs
 - Migrated to Google Analytics 4
 - Configuring custom events
- Browsertrix Cloud pilot through March
 - Hosted by Webrecorder
- Contributing to pywb development and v3 roadmap

Thank you!

Please reach out to us with questions or thoughts via email or in IIPC Slack. Or ask us questions now :-)

Laura Wrubel <lwrubel@stanford.edu>

Ed Summers <edsu@stanford.edu>