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
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: SC1915
Collector: Stanford University, Libraries, Department of Special Collections and University Archives

Narrow Your Results

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.

Enter search terms here  Search  Clear

Page 1 of 17 (1,693 Total Results)

Sort By: Title (A-Z) | Title (Z-A) | URL (A-Z) | URL (Z-A)
### Stanford News Service Stanford News: Stanford University Communications

**Overview**

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

**Details**

<table>
<thead>
<tr>
<th>Field</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Object type</td>
<td>Item</td>
</tr>
<tr>
<td>Content type</td>
<td>webarchive-seed</td>
</tr>
<tr>
<td>Project</td>
<td></td>
</tr>
<tr>
<td>Created</td>
<td>March 09, 2021</td>
</tr>
<tr>
<td>Released to</td>
<td>Searchworks</td>
</tr>
<tr>
<td>Preservation size</td>
<td>0 Bytes</td>
</tr>
<tr>
<td>Catkey</td>
<td>None assigned</td>
</tr>
<tr>
<td>Barcode</td>
<td>Not recorded</td>
</tr>
<tr>
<td>Tags</td>
<td>webarchive : seed, Registered By : pchan3, and Process :</td>
</tr>
</tbody>
</table>
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.
SUL Web Archives Storage Growth
1,025,365,726 snapshots
Why switch to pywb?
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 harassment. The questions are part of the ongoing research I’m doing at Don’t Die: www.dontdiedie.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@dontdiedie.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.
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 harassment. 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, 28, office worker.
7/1/2015 6:20 AM

Daniel, 38, teacher
7/5/2015 6:07 AM

Dan, 31, game tool programmer
7/4/2015 2:55 PM
This branch is 82 commits ahead, 643 commits behind iipc:master.
Commits on Jun 25, 2021

Fix capitalization
- ldko committed on Jun 25, 2021

Suggest using pywb over OpenWayback
- ldko committed on Jun 25, 2021

Commits on Oct 13, 2020

Merge pull request #435 from iipc/dependabot/maven/junit-junit-4.13.1
- ldko committed on Oct 13, 2020

Bump junit from 3.8.1 to 4.13.1
- dependabot[bot] committed on Oct 13, 2020
How did we switch?
Merge pull request #176 from sul-dlss

- .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 present (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)

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
Reindexing the WARC\s

- 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
collections:
  was:
  index_paths: /web-archiving-stacks/data/indexes/cdxj/
  archive_paths:
    - /web-archiving-stacks/data/collections/
  acl_paths: /web-archiving-stacks/data/access.aclj
Chinese Railroad Workers in North America Project at Stanford University

Announcement
August 31, 2023

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:// 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
Memento API

https://swap.stanford.edu/timemap/http:/news.stanford.edu/
How did we upgrade to 2.7?
Update dependencies #176

Merged 1wrubel merged 2 commits into main from update-dependencies 4 days ago

sul-devops-team commented 4 days ago

No description provided.

sul-devops-team added 2 commits 4 days ago

- Update Python dependencies
e34d171

- Update Ruby dependencies

lwrubel approved these changes 4 days ago

Reviewers
- lwrubel

Still in progress? Learn about draft PRs

Assignees
- No one—assign yourself

Labels
- None yet

View changes
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.
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

```yaml
ui:
  logo: images/swaplogo.png
  navbar_background_hex: 343a40
  navbar_color_hex: ffffff
  navbar_lightButtons: ffffff
```

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

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

anjackson commented on Mar 30, 2022

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": "20201226"}
```

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>