“What You See No One Saw”

Mat Kelly, Alexander H. Poole, Michele C. Weigle, Michael L. Nelson,
Travis Reid, Christopher Rauch, and Hyung Wook Choi

Drexel University CCI & **Old Dominion University WS-DL**
mkelly@drexel.edu – @machawk1

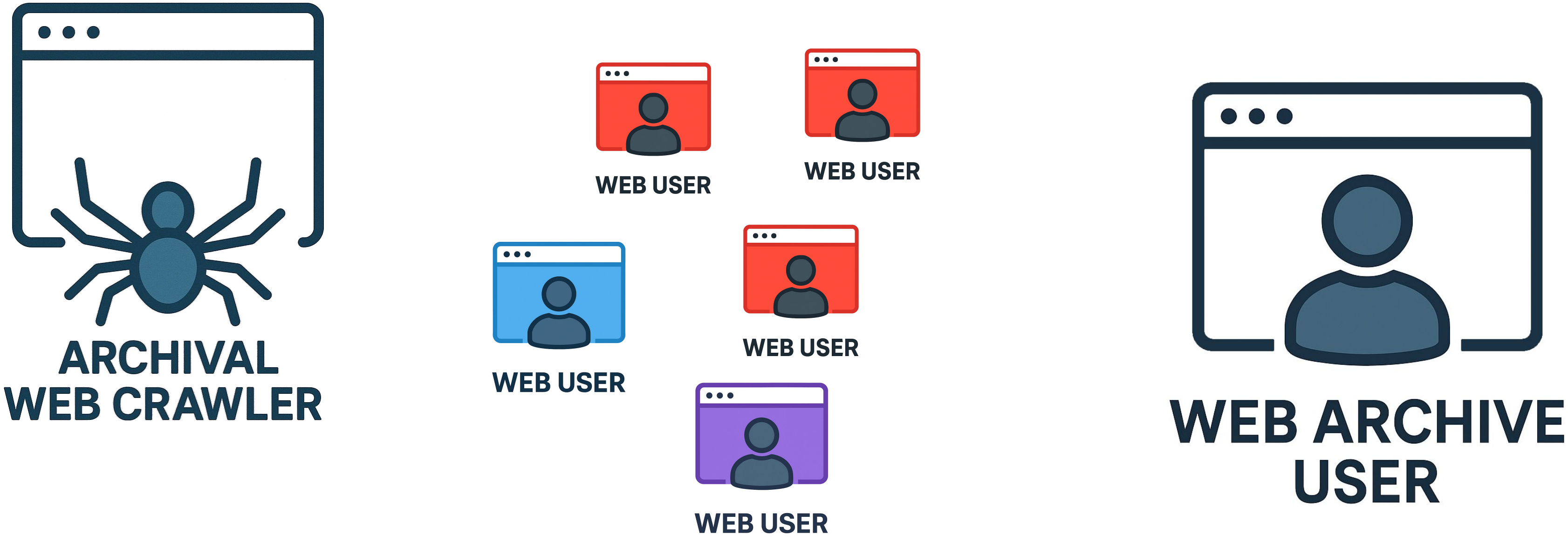
IIPC Web Archiving Conference (WAC)
Oslo, Norway
April 10, 2025

Supported by:
INSTITUTE of Museum and Library SERVICES
#LG-256695-OLS-24 &
#LG-252362-OLS-22

slides:
bit.ly/iipcwac2025

National Library of Norway
IIPC GENERAL ASSEMBLY & WEB ARCHIVING CONFERENCE
OSLO, 8–10 APRIL 2025 | NETPRESERVE.ORG/GA2025
IIPC
netpreserve.org

The Past Web Saved Is Not The Web That Was



ARCHIVAL WEB CRAWLER

WEB USER

WEB USER

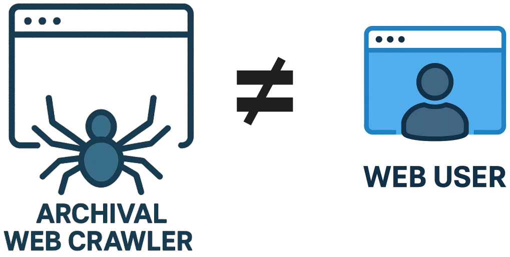
WEB USER

WEB USER

WEB USER

WEB ARCHIVE USER

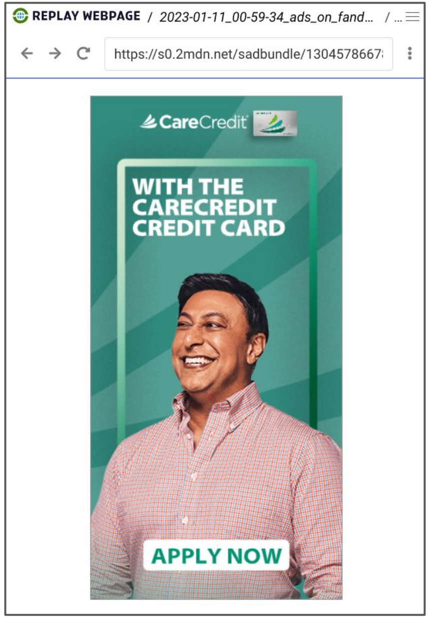
Crawler Preserve A False Web (Premise)



- Archival crawlers preserve a version of the web inconsistent with web users' experience, a web that actually wasn't
- Customization, personalization based on user history is not *canonical*
- Crawlers (rightfully) see a clean/agnostic version of web sites, devoid of any individuals' experience, PII
- Ergo, what crawlers preserve is a version of the web inconsistent with what a user would have seen at that time
- False history? Nature of experience

A Valid Perspective, Just Not A Web Users’

- Crawler’s perspective ***is*** a valid representation
- There is no *one true* representation of a personalized web
- Web Ads?
  - ○ Annoying but useful for study
  - ○ Hyper-personalized, distinguishing factor from generic representation
  - ○ Google’s Manifest V3 / Ad blocker drama* means Chrome users are returning to an ad-ridden web
  - ○ As with analog advertisements, web ads represent zeitgeist in retrospect


REPLAY WEBPAGE / 2023-01-11_00-59-34_ads_on_fand... / ...

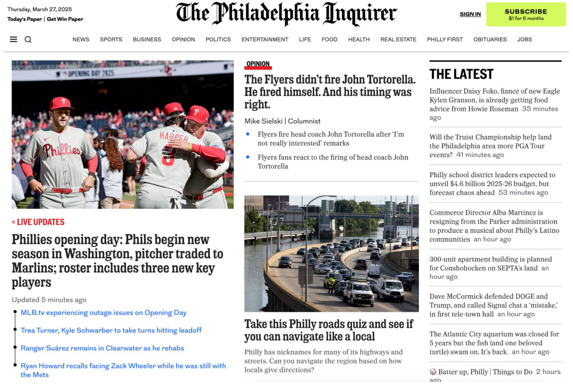
https://s0.2mdn.net/sadbundle/1304578667

INSTITUTE of Museum and Library SERVICES

* {
https://www.eff.org/deeplinks/2021/12/chrome-users-beware-manifest-v3-deceitful-and-threatening
https://developer.chrome.com/docs/extensions/develop/migrate/mv2-deprecation-timeline
https://blog.mozilla.org/addons/2024/03/13/manifest-v3-manifest-v2-march-2024-update/
https://developer.chrome.com/blog/resuming-the-transition-to-mv3/

Can We Save The Web We See From Our Perspective?

- Repurpose user’s daily driver profile as crawler basis
- Permutate attributes of a user to represent a “persona”, producing a web experience closer to that of an actual web user cf. crawler
- Avoid clean slate crawling and delegation to a user-agnostic crawler
- Scale?



Thursday, March 27, 2025
Today's Paper | Get Win Paper

The Philadelphia Inquirer

SIGN IN | SUBSCRIBE $1 for 6 months

NEWS SPORTS BUSINESS OPINION POLITICS ENTERTAINMENT LIFE FOOD HEALTH REAL ESTATE PHILLY FIRST OBITUARIES JOBS

● LIVE UPDATES

Phillies opening day: Phils begin new season in Washington, pitcher traded to Marlins; roster includes three new key players

Updated 5 minutes ago

- MLB.tv experiencing outage issues on Opening Day
- Trea Turner, Kyle Schwarber to take turns hitting leadoff
- Ranger Suárez remains in Clearwater as he rehabs
- Ryan Howard recalls facing Zack Wheeler while he was still with the Mets

OPINION

The Flyers didn’t fire John Tortorella. He fired himself. And his timing was right.

Mike Sielski | Columnist

- Flyers fire head coach John Tortorella after ‘I’m not really interested’ remarks
- Flyers fans react to the firing of head coach John Tortorella

Take this Philly roads quiz and see if you can navigate like a local

Philly has nicknames for many of its highways and streets. Can you navigate the region based on how locals give directions?

THE LATEST

Influencer Daisy Foko, fiancé of new Eagle Kylen Granson, is already getting food advice from Howie Roseman 35 minutes ago

Will the Truist Championship help land the Philadelphia area more PGA Tour events? 41 minutes ago

Philly school district leaders expected to unveil $4.6 billion 2025-26 budget, but forecast chaos ahead 53 minutes ago

Commerce Director Alba Martinez is resigning from the Parker administration to produce a musical about Philly’s Latino communities an hour ago

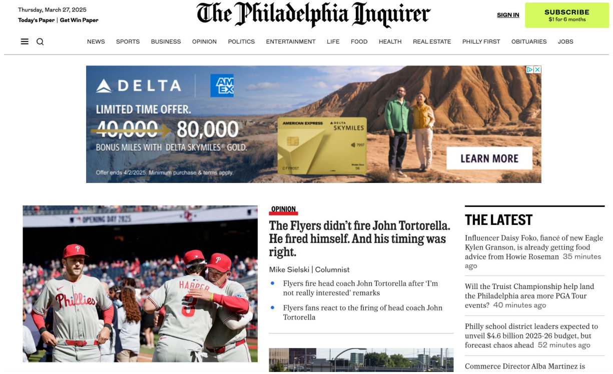
300-unit apartment building is planned for Conshohocken on SEPTA’s land an hour ago

Dave McCormick defended DOGE and Trump, and called Signal chat a ‘mistake,’ in first tele-town hall an hour ago

The Atlantic City aquarium was closed for 5 years but the fish (and one beloved turtle) swam on. It’s back. an hour ago

Batter up, Philly | Things to Do 2 hours ago

vs



Thursday, March 27, 2025
Today's Paper | Get Win Paper

The Philadelphia Inquirer

SIGN IN | SUBSCRIBE $1 for 6 months

NEWS SPORTS BUSINESS OPINION POLITICS ENTERTAINMENT LIFE FOOD HEALTH REAL ESTATE PHILLY FIRST OBITUARIES JOBS

DELTA | AMEX
LIMITED TIME OFFER.
~~40,000~~ 80,000
BONUS MILES WITH DELTA SKYMILES® GOLD.
Offer ends 4/2/2025. Minimum purchase & terms apply.
LEARN MORE

OPINION

The Flyers didn’t fire John Tortorella. He fired himself. And his timing was right.

Mike Sielski | Columnist

- Flyers fire head coach John Tortorella after ‘I’m not really interested’ remarks
- Flyers fans react to the firing of head coach John Tortorella

THE LATEST

Influencer Daisy Foko, fiancé of new Eagle Kylen Granson, is already getting food advice from Howie Roseman 35 minutes ago

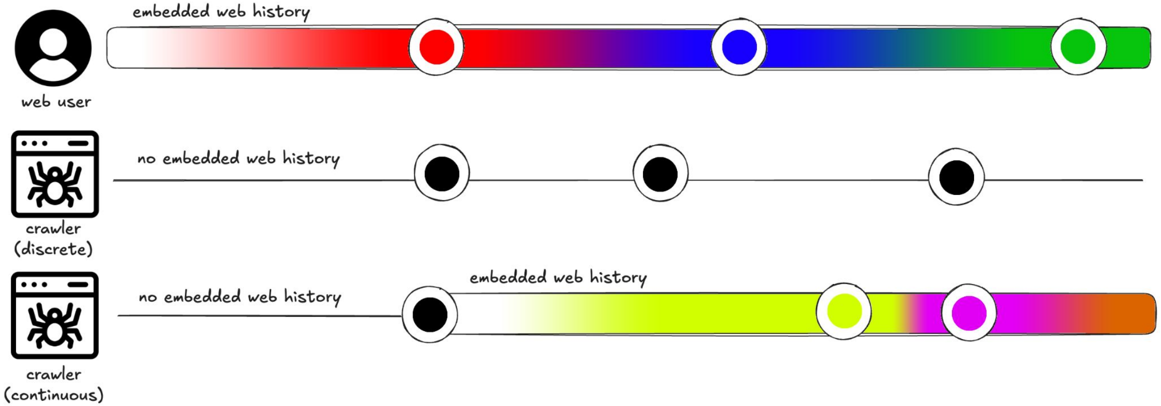
Will the Truist Championship help land the Philadelphia area more PGA Tour events? 40 minutes ago

Philly school district leaders expected to unveil $4.6 billion 2025-26 budget, but forecast chaos ahead 52 minutes ago

Commerce Director Alba Martinez is

Crawling w/ a Web History + Discrete vs. Continuous

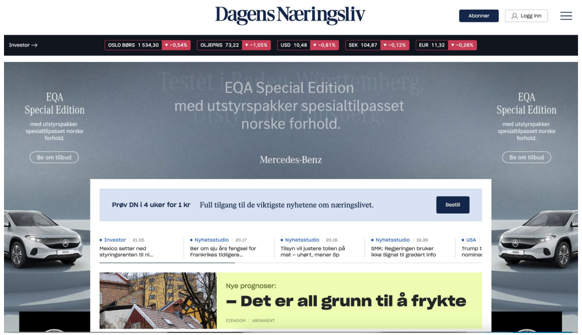
- Want to either reuse browser user profile or extract feature (e.g., cookies) to be used as the basis for what is served at archive time
  - What else is contained in this profile?
  - Is reuse possible/feasible for web archiving? What are Selenium's capabilities? Other headless crawlers



Prior Technical Work

Leveraging Perspective-Based Crawling

- WARCreate - browser extension that archives by-value (cf. URI as basis)
  - Manifest V3 caveat (webRequest)
- Warcprox - save representations as they come over the wire
- Ad Blockers
  - Are users seeing the ad-ridden, true representation of the web?



Prøv DN i 4 uker for 1 kr Full tilgang til de viktigste nyhetene om næringslivet.

Mexico setter ned styringsrenten til ni...

Ber om sju års fengsel for Frankrikes tidligere...

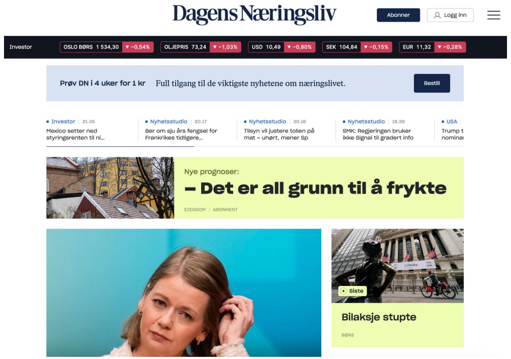
Tilsyn vil justere tollen på mat – uhørt, mener Sp

SMK: Regjeringen bruker ikke Signal til gradert info

Trump t nomina

Nye prognoser:

– Det er all grunn til å frykte



Prøv DN i 4 uker for 1 kr Full tilgang til de viktigste nyhetene om næringslivet.

Mexico setter ned styringsrenten til ni...

Ber om sju års fengsel for Frankrikes tidligere...

Tilsyn vil justere tollen på mat – uhørt, mener Sp

SMK: Regjeringen bruker ikke Signal til gradert info

Trump t nomina

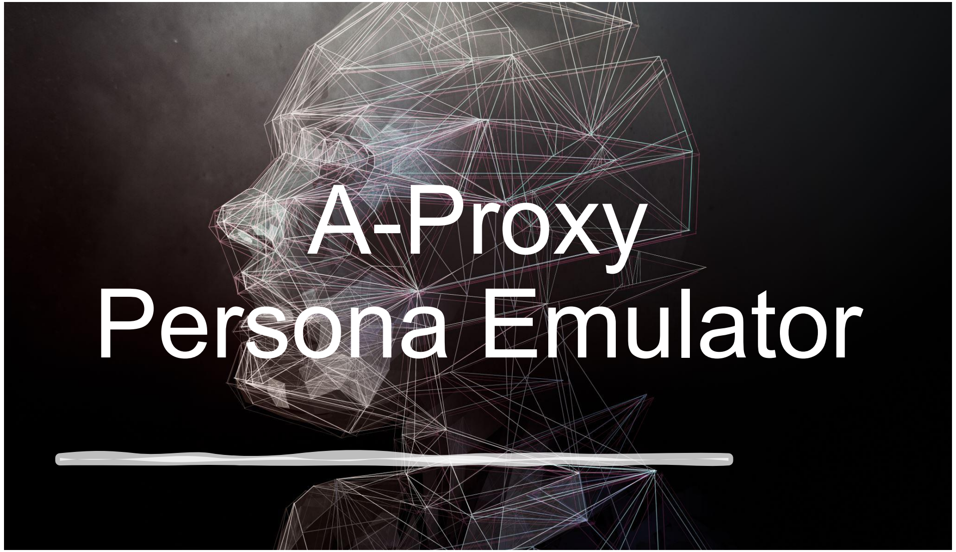
Nye prognoser:

– Det er all grunn til å frykte

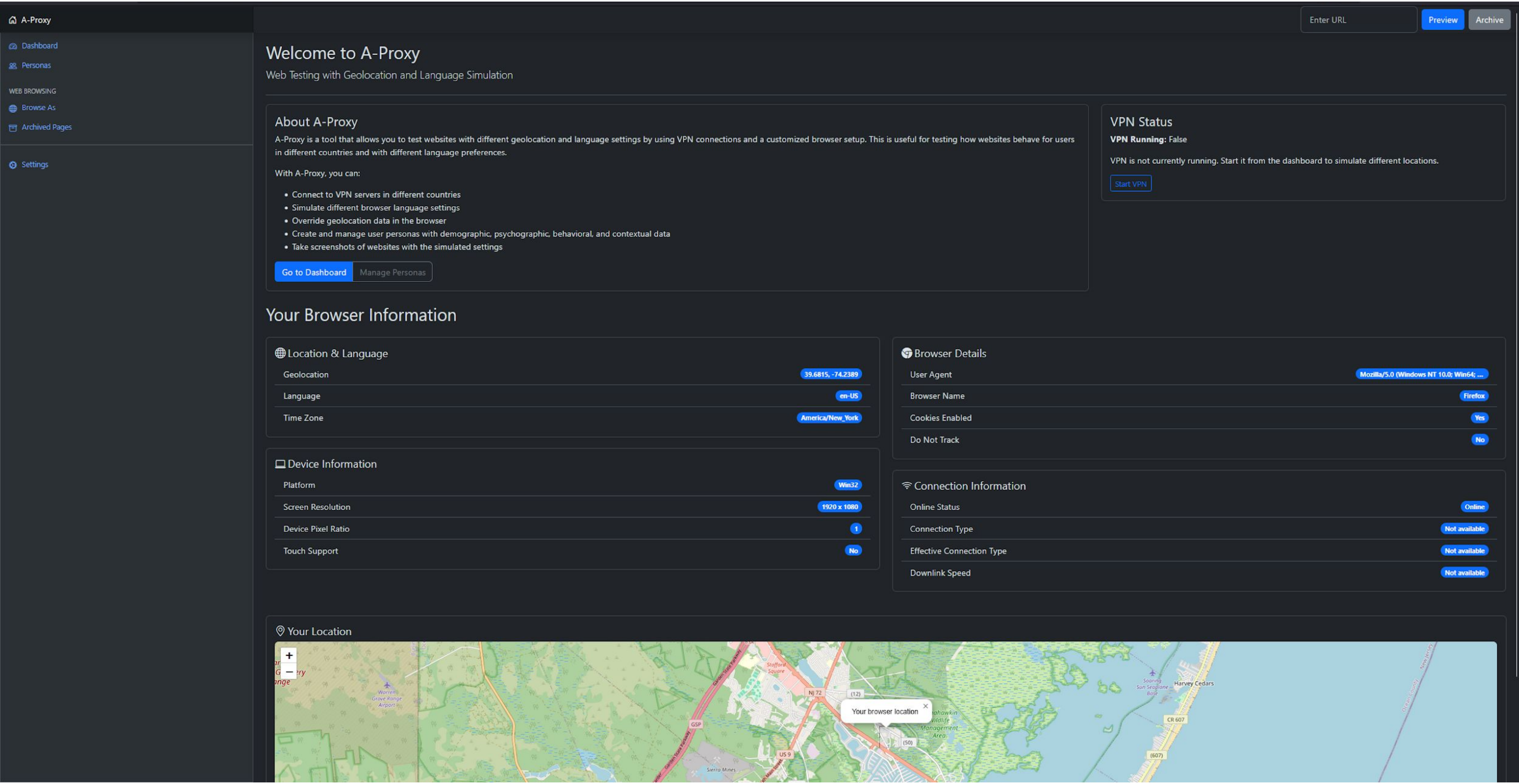
Bilaksje stupte

WIP: Persona-based Web Archiving

- Rapidly prototyped crawl director
- Side-load Chrome profile with Selenium WebDriver
- UI for user to specify crawl profile attributes
- Based on Andy Jackson's Sliver
  - https://github.com/anjackson/sliver



https://github.com/savingads/a-proxy



Welcome to A-Proxy

Web Testing with Geolocation and Language Simulation

About A-Proxy

A-Proxy is a tool that allows you to test websites with different geolocation and language settings by using VPN connections and a customized browser setup. This is useful for testing how websites behave for users in different countries and with different language preferences.

With A-Proxy, you can:

- Connect to VPN servers in different countries
- Simulate different browser language settings
- Override geolocation data in the browser
- Create and manage user personas with demographic, psychographic, behavioral, and contextual data
- Take screenshots of websites with the simulated settings

VPN Status

VPN Running: False

VPN is not currently running. Start it from the dashboard to simulate different locations.

Your Browser Information

Location & Language

Geolocation	39.6815, -74.2389
Language	en-US
Time Zone	America/New_York

Browser Details

User Agent	Mozilla/5.0 (Windows NT 10.0; Win64; ...
Browser Name	Firefox
Cookies Enabled	Yes
Do Not Track	No

Device Information

Platform	Win32
Screen Resolution	1920 x 1080
Device Pixel Ratio	1
Touch Support	No

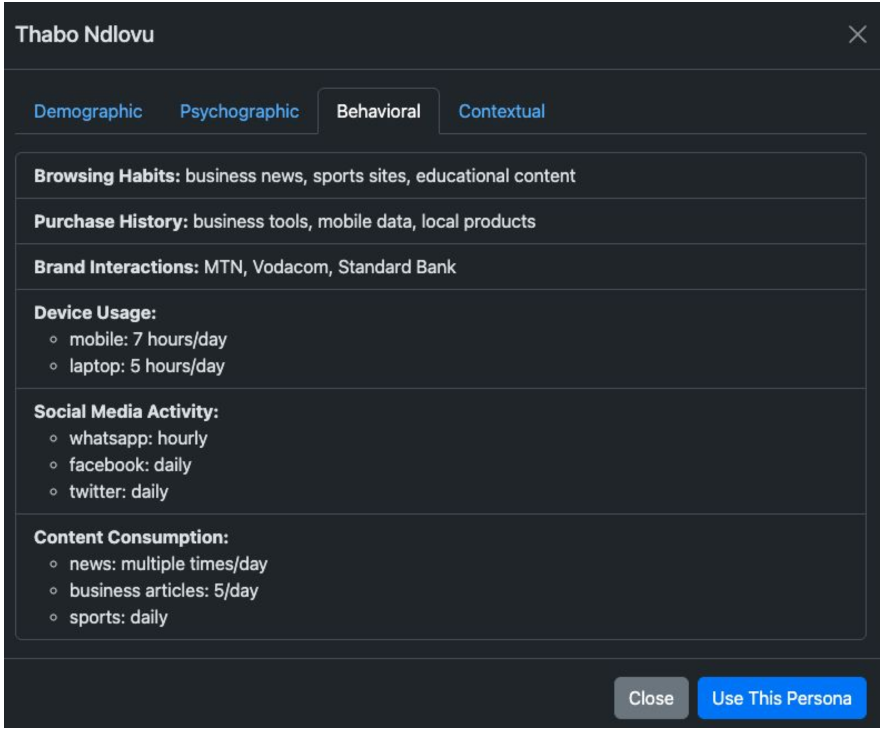
Connection Information

Online Status	Online
Connection Type	Not available
Effective Connection Type	Not available
Downlink Speed	Not available

Your Location

Parameters of Perspective / Personalization

- Demographic
  - ex: Location, Language
- Psychographic
  - ex: interests, values
- Behavioral
  - ex: browsing habits, social media activity
- Contextual
  - ex: time of day, weather, browser



Thabo Ndlovu

Browsing Habits: business news, sports sites, educational content

Purchase History: business tools, mobile data, local products

Brand Interactions: MTN, Vodacom, Standard Bank

Device Usage:
- mobile: 7 hours/day
- laptop: 5 hours/day

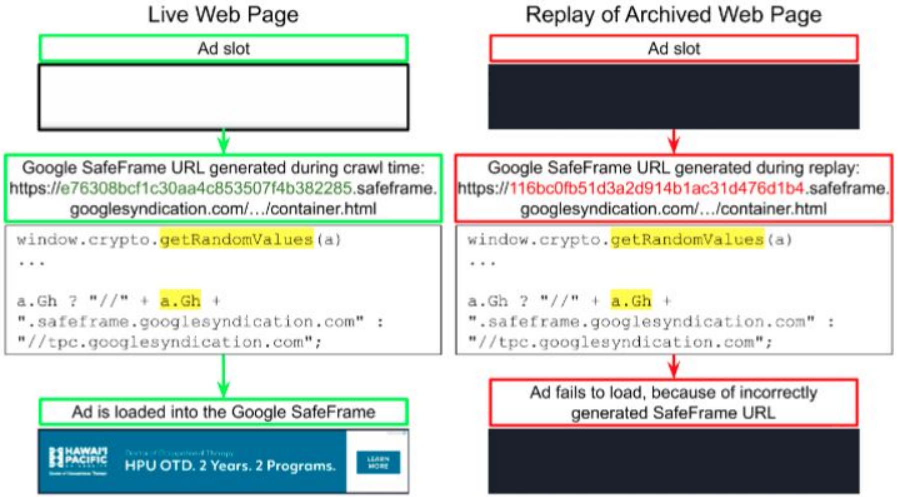
Social Media Activity:
- whatsapp: hourly
- facebook: daily
- twitter: daily

Content Consumption:
- news: multiple times/day
- business articles: 5/day
- sports: daily

Initial Observations and Experiments

- Examine existing crawler-derived WARCs, observe personalization traits
- Enumerate attributes like location, user-agent, language for rudimentary personalization
- Use case: capture delta of web ads
  - Difficult due to randomization, dynamic
  - Requires replay amendment
  - Web ads are more than just images
    - e.g., video, combo, interactive

BROWSERTRIX



Live Web Page

Ad slot

Google SafeFrame URL generated during crawl time: https://e76308bcf1c30aa4c853507f4b382285.safeframe.googlesyndication.com/…/container.html

```
window.crypto.getRandomValues(a)
...

a.Gh ? "//" + a.Gh +
".safeframe.googlesyndication.com" :
"//tpc.googlesyndication.com";
```

Ad is loaded into the Google SafeFrame

Replay of Archived Web Page

Ad slot

Google SafeFrame URL generated during replay: https://116bc0fb51d3a2d914b1ac31d476d1b4.safeframe.googlesyndication.com/…/container.html

```
window.crypto.getRandomValues(a)
...

a.Gh ? "//" + a.Gh +
".safeframe.googlesyndication.com" :
"//tpc.googlesyndication.com";
```

Ad fails to load, because of incorrectly generated SafeFrame URL

See arXiv:2502.01525, 2025

“What You See Now One Saw”

WEB USER

WEB ARCHIVE USER

Project supported by INSTITUTE of Museum and Library SERVICES

- Sufficient personalized context is lost when delegating to a crawl by URI
- That which we consider the historical web was captured through the lens of a perspective-agnostic crawler
- Project underway, interpolating personas, gathering data, building A-Proxy
- Ongoing dev work, data at https://github.com/savingads

See our recent tech report on archiving web ads! arXiv:2502.01525, 2025. →

Slides: bit.ly/iipcwac2025

Archiving and Replaying Current Web Advertisements: Challenges and Opportunities

TRAVIS REID, Old Dominion University, USA
ALEX H. POOLE, Drexel University, USA
HYUNG WOOK CHOI, Drexel University, USA
CHRISTOPHER RAUCH, Drexel University, USA
MAT KELLY, Drexel University, USA
MICHAEL L. NELSON, Old Dominion University, USA
MICHELE C. WEIGLE, Old Dominion University, USA

Although web advertisements represent an inimitable part of digital cultural heritage, serious archiving and replay challenges persist. To explore these challenges, we created a dataset of 279 archived ads. We encountered five problems in archiving and replaying them. For one, prior to August 2023, Internet Archive’s Save Page Now service excluded not only well-known ad services’ ads, but also URLs with ad related file and directory names. Although after August 2023, Save Page Now still blocked the archiving of ads loaded on a web page, it permitted the archiving of an ad’s resources if the user directly archived the URL(s) associated with the ad. Second, Brozzler’s incompatibility with Chrome prevented ads from being archived. Third, during crawling and replay sessions, Google’s and Amazon’s ad scripts generated URLs with different random values. This precluded archived ads’ replay. Updating replay systems’ fuzzy matching approach should enable the replay of these ads. Fourth, when loading Flashtalking web page ads outside of ad iframes, the ad script requested a non-existent URL. This, prevented the replay of ad resources. But as was the case with Google and Amazon ads, updating replay systems’ fuzzy matching approach should enable Flashtalking ads’ replay. Finally, successful replay of ads loaded in iframes with the `src` attribute of “`about:blank`” depended upon a given browser’s service worker implementation. A Chromium bug stopped service workers from accessing resources inside of this type of iframe, which in turn prevented replay. Replacing the “`about:blank`” value for the iframe’s `src` attribute with a blob URL before an ad was loaded solved this problem. Resolving these replay problems will improve the replay of ads and other dynamically loaded embedde

1 INTRODUCTION

Brewster Kahle, founder of the Internet Archive,
of valuable scientific, cultural and historical inf
characterized the web in similar terms, but also
for the study of almost every possible aspect of t
scholars, however, web content has been hemor
Whether impelled by legal obligation, business
and/or historical research, web archiving involve
to content [6, 14, 51, 53]. Web archives may be u
about the period in which the archived content
Because the web depends upon advertising re
dynamic content. Just as physical ephemera in l

Authors’ addresses: Travis Reid, Department of Computer S
Alex H. Poole, Department of Information Science, Drexel U
Department of Information Science, Drexel University, Phila
Information Science, Drexel University, Philadelphia, PA, 1
Drexel University, Philadelphia, PA, 19104, USA, mkelly@d
University, Norfolk, VA, 23529, USA, mln@cs.odu.edu; Mic
Norfolk, VA, 23529, USA, mweigle@cs.odu.edu.