



“What You See No One Saw”

Mat Kelly



Drexel University
College of Computing and Informatics (CCI)
mkelly@drexel.edu – @machawk1
In collaboration with ODU WS-DL



Archive-It Partner Meeting

Philadelphia, PA
June 26, 2025

slides:

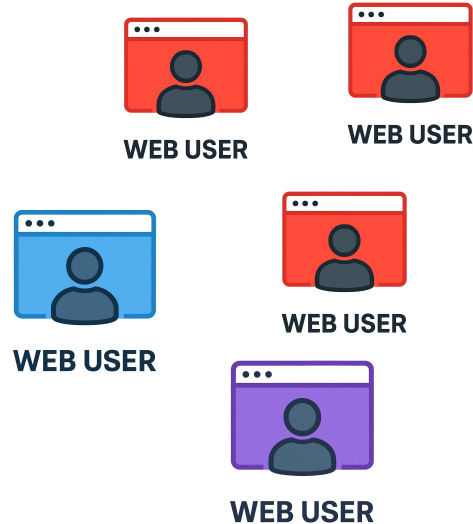
bit.ly/ait2025

Supported by:

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

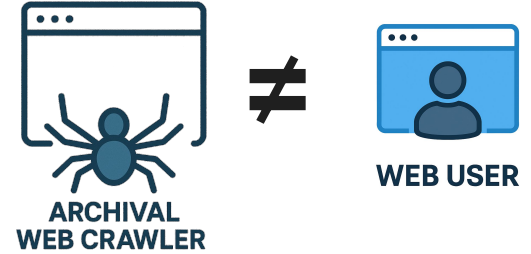


The Past Web Saved Is Not The Web That Was



Crawlers 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

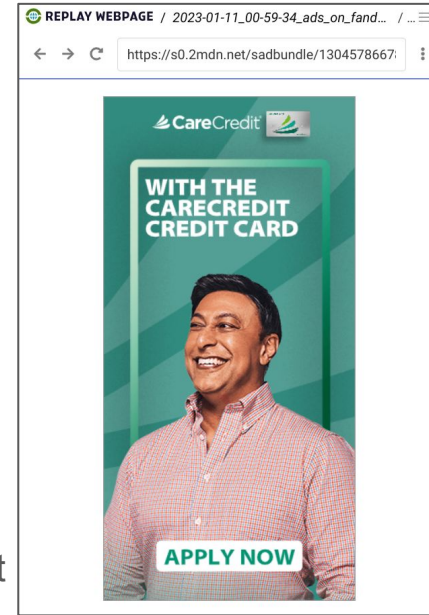


A Valid Perspective, Just Not A Web Users'

- 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



<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 Web 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

PHILLIES
Phillies opening day: Phils begin new season in Washington, pitcher traded to Marlins; roster includes three new key players
Updated 5 minutes ago
• MLB's experiencing outage issues on Opening Day
• Trea Turner, Kyle Schwarber to take turns hitting leadoff
• Ranger Suarez remains in Clearwater as he rehab
• Ryan Howard recalls facing Zack Wheeler while he was still with the Mets

PHILLY
The Flyers didn't fire John Tortorella. He fired himself. And his timing was right.
Mike Siebeki | 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

PHILLY
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 Kyle Gorman, is already getting food advice from Howie Roseman 35 minutes ago
Will the Trust Championship help land the Philadelphia area more PGA Tour events? 41 minutes ago
Philly school district leaders expected to spend \$4.4 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 3 years but the fish land one beloved turtle swim on its back an hour ago
Butter up Philly's Things to Do 2 hours ago

VS

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

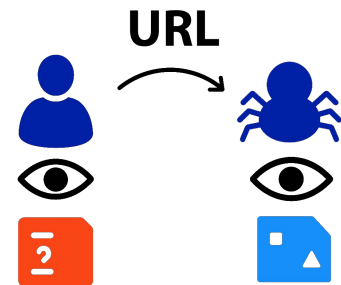
DELTA
LIMITED TIME OFFER.
40,000 → 80,000
BONUS MILES WITH DELTA SKYMILES® GOLD.
Offer ends 4/2/25. Minimum purchase & taxes apply. LEARN MORE

PHILLIES
Phillies opening day: Phils begin new season in Washington, pitcher traded to Marlins; roster includes three new key players
Updated 5 minutes ago
• MLB's experiencing outage issues on Opening Day
• Trea Turner, Kyle Schwarber to take turns hitting leadoff
• Ranger Suarez remains in Clearwater as he rehab
• Ryan Howard recalls facing Zack Wheeler while he was still with the Mets

PHILLY
The Flyers didn't fire John Tortorella. He fired himself. And his timing was right.
Mike Siebeki | 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

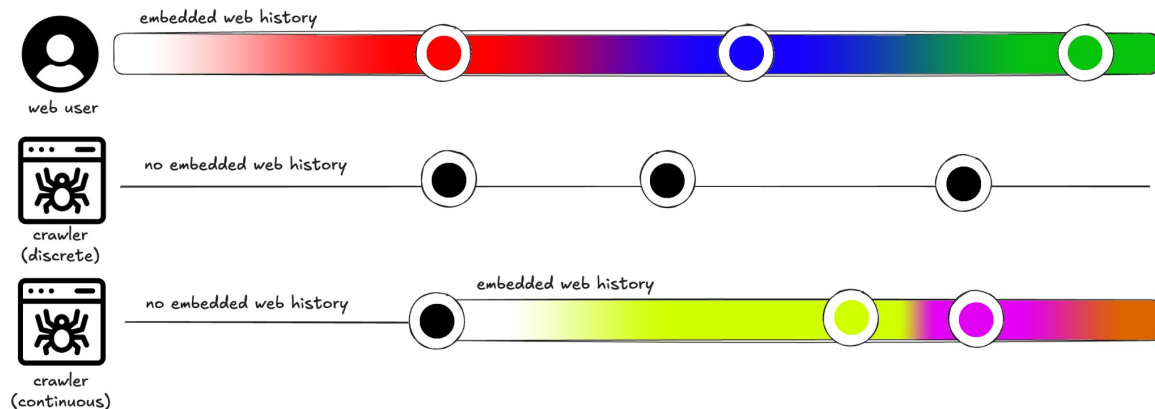
PHILLY
Take this Philly roads quiz and see if you can navigate like a local
Philly has nicknames for many of its highways and streets. Can you navigate the region based on how locals give directions?

THE LATEST
Influencer Daisy Foko, fiancé of new Eagle Kyle Gorman, is already getting food advice from Howie Roseman 35 minutes ago
Will the Trust Championship help land the Philadelphia area more PGA Tour events? 40 minutes ago
Philly school district leaders expected to spend \$4.4 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

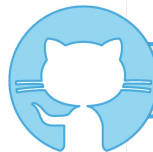
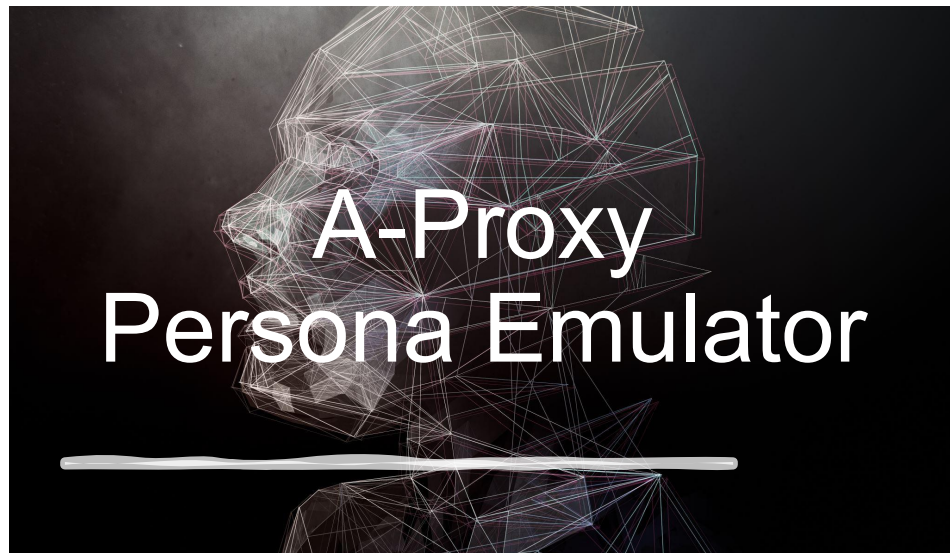
Leveraging Perspective-Based Crawling

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



WIP: Persona-based Web Archiving

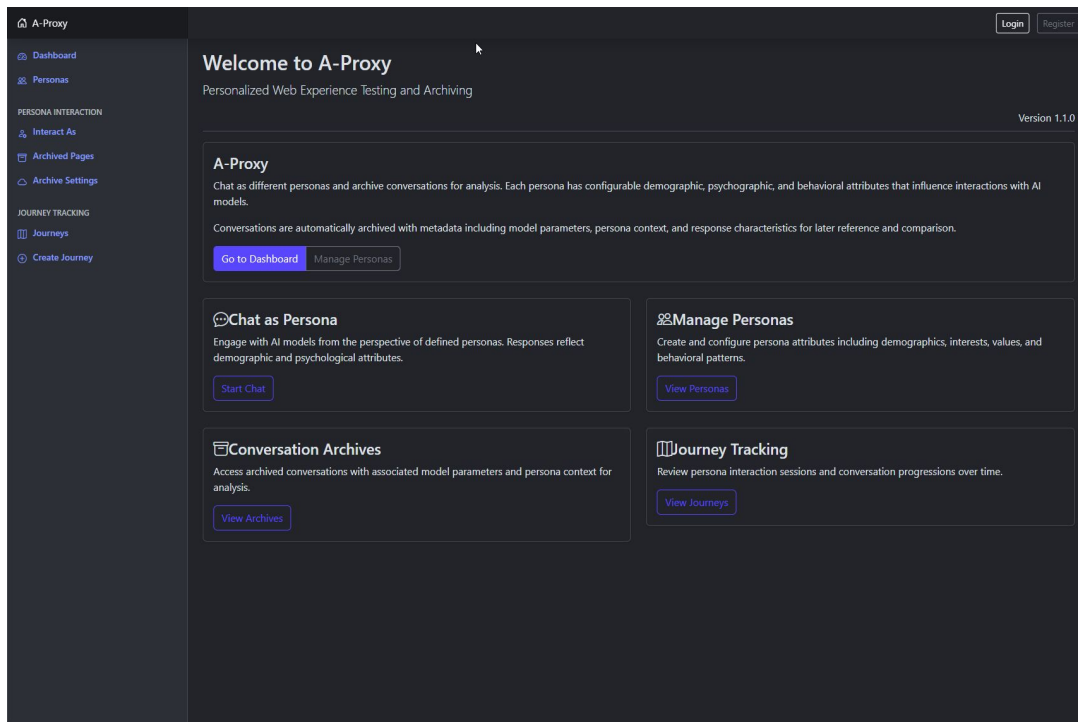
- Rapidly prototyped crawl director
- Side-load Chrome profile with Selenium WebDriver
- UI for user to specify crawl profile attributes
- Based on Andy Jackson's Sliver
 - <https://github.com/anjackson/sliver>



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

A-Proxy Dashboard

- SQLite database for persistence
- Integrated persona management (no separate microservice)
- Comprehensive data relationships and indexing



A-Proxy Usage Flow

AKA the feedback loop

1. Persona generation

- Archivists create initial personas across four categories (w/ LLM assistance)

2. Persona development through chat

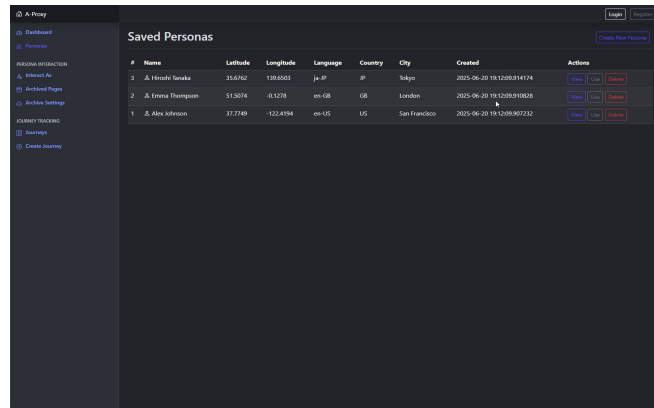
- Archivists chat w/ personas to develop their attributes over time or give directives in a certain format (e.g., you have a job in a restaurant on 36th Street)

3. Targeted Crawling (Waypoints)

- Browser variables reflect persona characteristics, cookies set context, synthetic history files represent past activity

4. Dual Archival value

- **For web archiving:** refined personas enable crawls from specific perspectives
- **For LLM research:** preserved chats document how LLMs respond to increased personalization



#	Name	Latitude	Longitude	Language	Country	City	Created	Actions
3	A. Hiroshi Tanaka	35.6762	139.6903	ja-JP	JP	Tokyo	2025-06-20 19:12:09.914174	View Edit Delete
2	A. Emma Thompson	51.5074	-0.1278	en-GB	GB	London	2025-06-20 19:12:09.918038	View Edit Delete
1	A. Alex Johnson	37.7749	-122.4194	en-US	US	San Francisco	2025-06-20 19:12:09.907232	View Edit Delete

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

The screenshot shows a user profile for 'Thabo Ndlovu'. It features four tabs: 'Demographic', 'Psychographic', 'Behavioral' (which is selected), and 'Contextual'. The 'Behavioral' tab displays several data points:

- 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

At the bottom right, there are two buttons: 'Close' and 'Use This Persona'.

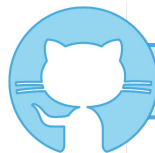
Example: Demographic & Psychographic

The screenshot shows the A-Proxy interface for a user named Hiroshi Tanaka. The top navigation bar includes 'A-Proxy', 'Login', 'Register', 'Enter URL', 'Preview', and 'Archive'. The left sidebar lists 'Dashboard', 'Persons', 'PERSONA INTERACTION', 'Interact As', 'Archived Pages', 'Archive Settings', 'JOURNEY TRACKING', 'Journeys', and 'Create Journey'. The main content area is titled 'Hiroshi Tanaka' and includes a 'Geolocation' field with coordinates 39.682048, -74.2293504 and a 'Language' field set to 'ja-JP'. Below this, there are four tabs: 'Demographic', 'Psychographic', 'Behavioral', and 'Contextual'. The 'Demographic' tab is active, displaying a table of demographic information and a map of Tokyo.

Demographic Information	
Name:	Hiroshi Tanaka
Country:	JP
City:	Tokyo
Region:	Kanto
Language:	ja-JP
Age:	35
Gender:	Male
Education:	Bachelor's Degree
Income:	Medium-High
Occupation:	Business Analyst

The screenshot shows the A-Proxy interface for the same user, Hiroshi Tanaka, but with the 'Psychographic' tab selected. The top navigation bar and left sidebar are identical to the previous screenshot. The main content area shows the 'Psychographic' tab, which includes a title 'Psychographic' and a subtitle 'Psychological attributes, interests, and values'. Below this, there are four sections: 'Interests', 'Personal Values', 'Attitudes', and 'Lifestyle'. Each section contains a list of tags representing the user's psychological profile.

Psychographic	
Interests:	anime, gaming, technology, traditional culture
Personal Values:	respect, efficiency, tradition
Attitudes:	detail-oriented, respectful, technology-forward
Lifestyle:	Urban traditional-modern blend
Personality:	Methodical, respectful
Opinions:	quality-focused, tradition-respecting



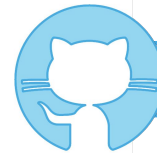
<https://github.com/savingads/a-proxy>

Example: Behavioral and Contextual

The screenshot shows the A-Proxy interface for a user named Hiroshi Tanaka. The interface is dark-themed with a sidebar on the left containing navigation links: Dashboard, Personas, PERSONA INTERACTION, Interact As, Archived Pages, Archive Settings, JOURNEY TRACKING, Journeys, and Create Journey. The main content area has a header with 'Hiroshi Tanaka' and an 'Edit' button. Below the header, there's a section for 'Geolocation: 39.6820824, -74.2378597' and 'Language: ja-JP'. The 'Behavioral' tab is selected, showing 'Online behavior and usage patterns'. It includes sections for 'Browsing Habits' (tech reviews, gaming sites, news), 'Purchase History' (electronics, games, traditional items), 'Brand Interactions' (Sony, Nintendo, Uniqlo), 'Device Usage' (mobile: 8 hours/day, desktop: 4 hours/day, gaming: 3 hours/day), 'Social Media Activity' (line: daily, twitter: daily, instagram: weekly), and 'Content Consumption' (videos: 4 hours/day, manga: 2 hours/day, news: 1 hour/day).

The screenshot shows the A-Proxy interface for a user named Hiroshi Tanaka, with the 'Contextual' tab selected. The interface is dark-themed with a sidebar on the left containing navigation links: Dashboard, Personas, PERSONA INTERACTION, Interact As, Archived Pages, Archive Settings, JOURNEY TRACKING, Journeys, and Create Journey. The main content area has a header with 'Hiroshi Tanaka' and an 'Edit' button. Below the header, there's a section for 'Geolocation: 39.6820824, -74.2378597' and 'Language: ja-JP'. The 'Contextual' tab is selected, showing 'Situational and environmental factors'. It includes a table with the following data:

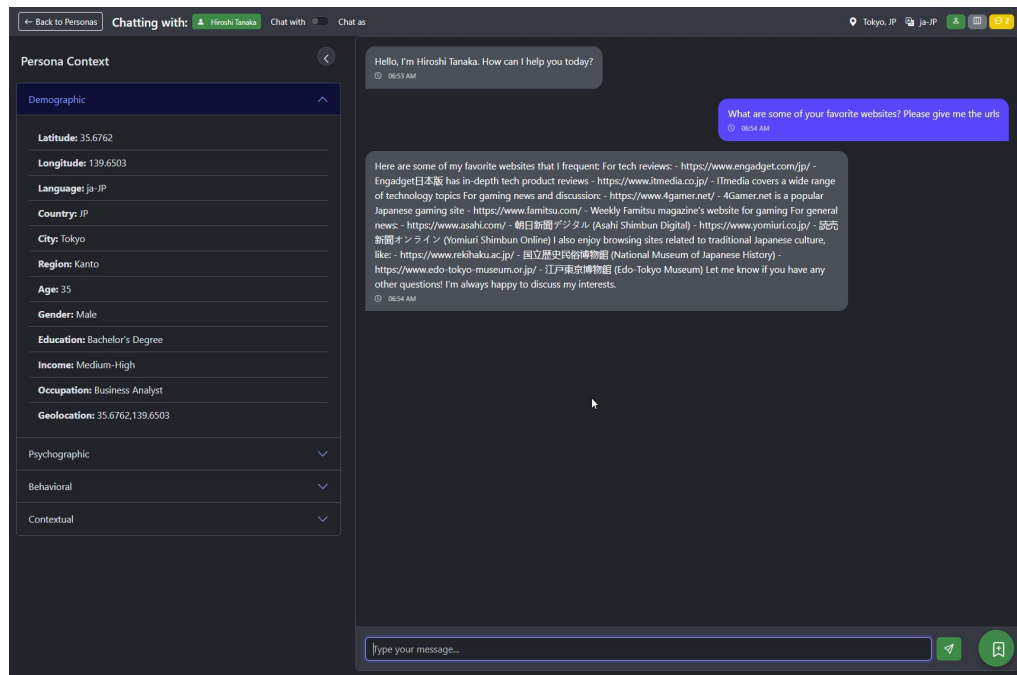
Time of Day:	evening
Day of Week:	weekday
Season:	summer
Weather:	humid
Device Type:	mobile
Browser Type:	chrome
Screen Size:	390x844
Connection Type:	5g



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

AI Chat Integration to Identify Frequented URLs

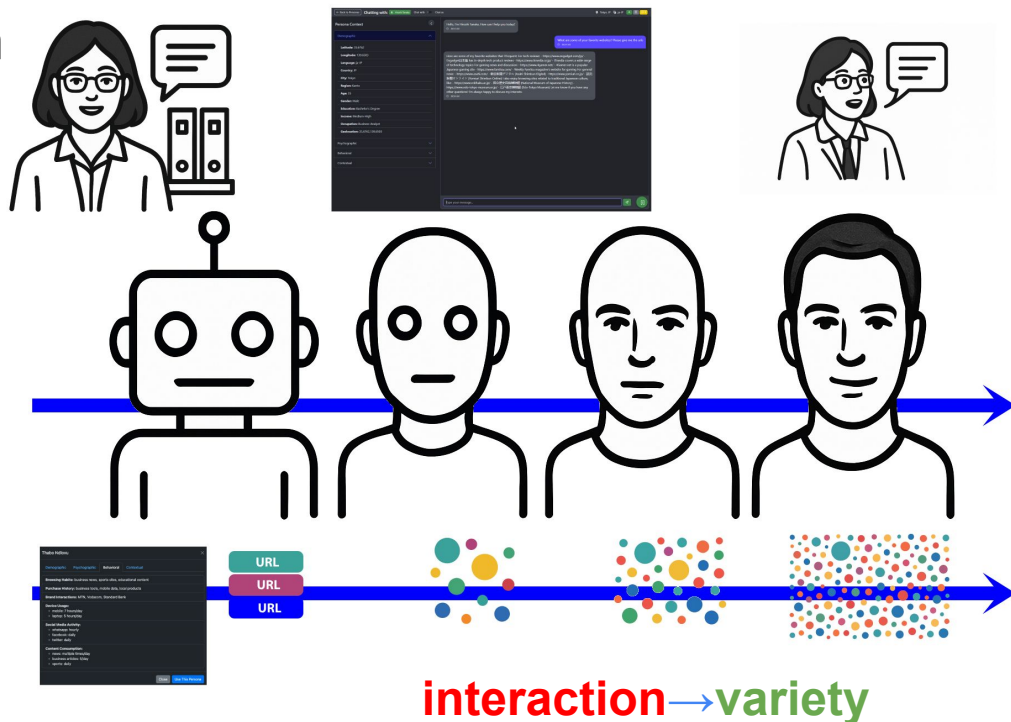
- Chat with Claude AI models as different personas
- Persona attributes influence AI responses
- Conversation history and context management
- Model parameter tracking for analysis
- These could also be supplied as seeds to a crawler or ad hoc web archiving session



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

Journey Tracking

- Create browsing/interaction sessions
- Track waypoints and conversation progression
- Archive sessions with full metadata



Core Endpoints

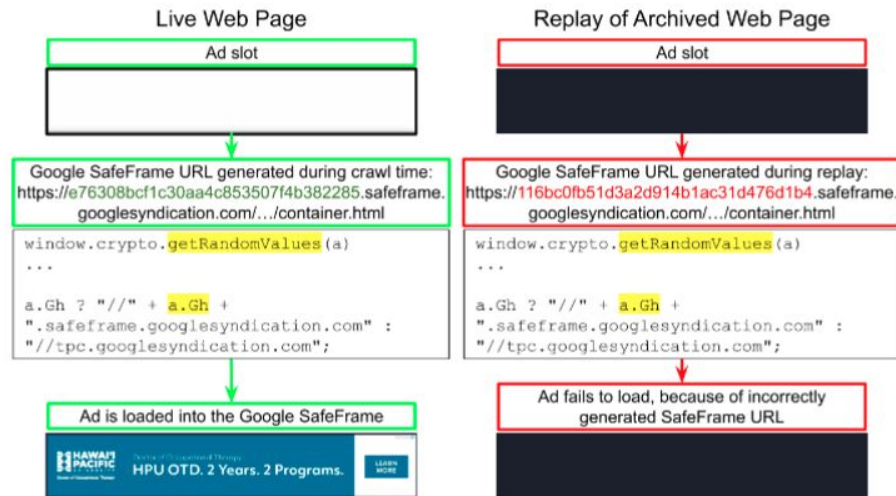
- / - Home dashboard
- /login - User authentication
- /personas - Persona management
- /interact-as - Choose persona for interaction
- /direct-chat/<persona_id> - AI chat interface
- /journeys - Journey tracking
- /archives - Conversation archives

How'd We Get to AI-informed Archiving?

- Original impetus of this work was to capture personalized ads
 - ...but it's also useful for personalized content not exposed to unpersonalized crawlers
- We created two data sets to classify existing ads in archives and developed novel approaches to facilitate capture of contemporary web ads
- Capture alone is often not enough, replay system need to be adjusted to account for runtime, generated content
 - e.g., client-side randomization to define resource location on live web may be replicated differently at replay time

Initial Observations and Experiments

- Examine existing crawler-derived WARC, 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



See arXiv:2502.01525, 2025

“What You See No One Saw”



- 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](https://arxiv.org/abs/2502.01525), 2025. →

Slides: bit.ly/ait2025

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, Brozler'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 Flashalking 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 Flashalking 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 embedded

1 INTRODUCTION

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

Authors' addresses: Travis Reid, Department of Computer Science, Old Dominion University, Norfolk, VA, 23529, USA, mreid@cs.odu.edu; Alex H. Poole, Department of Information Science, Drexel University, Philadelphia, PA, 19104, USA, apoole@drexel.edu; Hyung Wook Choi, Department of Information Science, Drexel University, Philadelphia, PA, 19104, USA, hwochoi@drexel.edu; Christopher Rauch, Department of Information Science, Drexel University, Philadelphia, PA, 19104, USA, crauch@drexel.edu; Mat Kelly, Department of Information Science, Drexel University, Philadelphia, PA, 19104, USA, mkelly@drexel.edu; Michael L. Nelson, Old Dominion University, Norfolk, VA, 23529, USA, mln@cs.odu.edu; Michele C. Weigle, Old Dominion University, Norfolk, VA, 23529, USA, mweigle@cs.odu.edu.

