Preserving Personalized Advertisements for More Accurate Web Archives

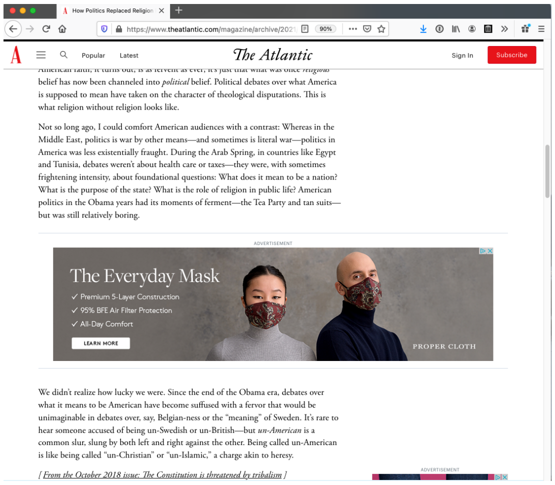
Drexel University's College of Computing and Informatics (Mat Kelly and Alex H. Poole), in collaboration with Old Dominion University's (ODU's) Department of Computer Science (Michele C. Weigle and Michael L. Nelson), respectfully requests $398,927 for a two-year National Leadership Applied Research Grant. Building on the PIs' our National Leadership Planning Grant project, "Saving Ads" (2022-2024), the proposed project homes in on personalized ads that elude conventional web archiving methods. It will thereby augment scholarly and lay users' capability to access and use web pages and advertisements by integrating insights derived from quantifying archival loss. In particular, this analysis will iteratively inform 1) the formulation of best practice guidelines and recommendations and 2) the development of more effective tools both to reconstruct advertisements from already archived content and to archive future web content. By improving archives' capacity for providing access to and use of collections of born-digital web content, this project aligns with IMLS Program **Goal 3** and **Objective 3.1** (Support collections care and management) and **3.2** (Promote access to museum and library collection).

Preserving Personalized Advertisements for More Accurate Web Archives extends a vital, timely interdisciplinary and collaborative initiative that engages both the technological shortcomings of and human negligence in archiving advertisements embedded in web pages. The project focuses on personalized ads – those viewed on the live web by users but seldom if ever captured by archival crawlers. It tackles three core research questions (RQs):

- RQ1: To what extent do institutional web archives capture personalized advertisements on the web?
- RQ2: Do scholarly and lay users prefer (re)using archived web pages that include personalized ads, that include a generic comprehensive capture, or that do not include web pages with ads?
- RQ3: How might the strategic use of diverse personas illuminate web content that would otherwise go unarchived?

1 Project Justification

To effectively study not only the past web but the recent past itself, we need web archives to preserve and present authentic, accurate, reliable, and diverse documentation. For example, web advertisements disseminated during the COVID-19 pandemic (as in Figure 1) offer visceral cultural, social, and political traces of that deeply traumatic event. But despite the web's overriding significance as a medium of public discourse, it is often overlooked as a scholarly and layperson documentary source—as is the deluge of advertisements users encounter in our daily web use. These advertisements fundamentally affect our overall user experience. Online advertising also reflects and reinforces changing societal mores concerning diversity, equity, and inclusion (e.g., race, ethnicity, gender, class, geography, nationality, and religion).


How Politics Replaced Religion
https://www.theatlantic.com/magazine/archive/2021/
90%
Popular
Latest
The Atlantic
Sign In
Subscribe
ADVERTISEMENT
The Everyday Mask
✓ Premium 5-Layer Construction
✓ 95% BFE Air Filter Protection
✓ All-Day Comfort
LEARN MORE
PROPER CLOTH
ADVERTISEMENT

American faith, it turns out, is as fervent as ever; it's just that what was once *religious* belief has now been channeled into *political* belief. Political debates over what America is supposed to mean have taken on the character of theological disputations. This is what religion without religion looks like.

Not so long ago, I could comfort American audiences with a contrast: Whereas in the Middle East, politics is war by other means—and sometimes is literal war—politics in America was less existentially fraught. During the Arab Spring, in countries like Egypt and Tunisia, debates weren't about health care or taxes—they were, with sometimes frightening intensity, about foundational questions: What does it mean to be a nation? What is the purpose of the state? What is the role of religion in public life? American politics in the Obama years had its moments of ferment—the Tea Party and tan suits—but was still relatively boring.

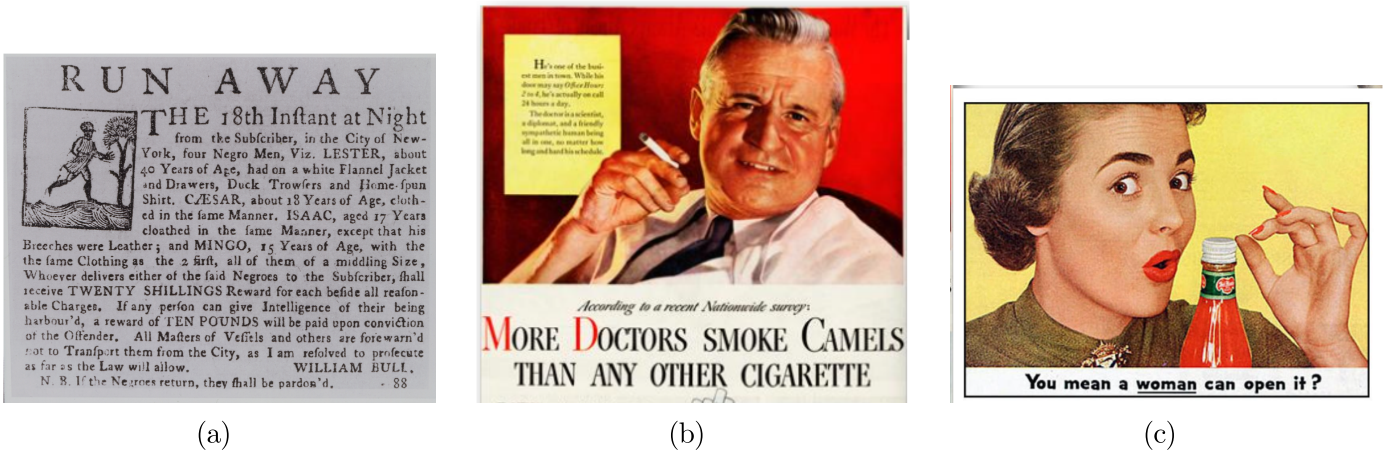
We didn't realize how lucky we were. Since the end of the Obama era, debates over what it means to be American have become suffused with a fervor that would be unimaginable in debates over, say, Belgian-ness or the "meaning" of Sweden. It's rare to hear someone accused of being un-Swedish or un-British—but *un-American* is a common slur, slung by both left and right against the other. Being called un-American is like being called "un-Christian" or "un-Islamic," a charge akin to heresy.

[From the October 2018 issue: The Constitution is threatened by tribalism]

Figure 1: Online advertisement for face masks in March 2021.

Online advertisements are no less markers of a society than print advertisements have been for centuries. Advertisements shape and are shaped by societal norms. Consider the print ads in Figure 2. All of these were mainstream and uncontroversial at the time they were published, but none of them would be acceptable to publish today. Historians have used advertisements such as these to uncover the mores and perspectives of a time and place [Lears, 1995, Gardner and Brandt, 2006, Sivulka, 2011]. Similarly, historians studying the 21st century will need to access contemporary online ads to illuminate social norms, viewpoints, and ideals as well as the objectives of advertisers themselves.

Surprisingly, even those institutions that engage in web archiving often exclude advertisements, erroneously deeming them ephemeral, inconsequential, and wasteful of resources instead of the fundamentally revealing documentation that they are. Absent this content, however, snapshots of the past Web and thus of past culture, society,



RUN AWAY

THE 18th Inſtant at Night from the Subſcriber, in the City of New-York, four Negro Men, Viz. LESTER, about 40 Years of Age, had on a white Flannel Jacket and Drawers, Duck Trowſers and Home-ſpun Shirt. CÆSAR, about 18 Years of Age, clothed in the ſame Manner. ISAAC, aged 17 Years cloathed in the ſame Manner, except that his Breeches were Leather; and MINGO, 15 Years of Age, with the the ſame Clothing as the 2 firſt, all of them of a middling Size, Whoever delivers either of the ſaid Negroes to the Subſcriber, ſhall receive TWENTY SHILLINGS Reward for each beſide all reaſonable Charges. If any perſon can give Intelligence of their being harbour'd, a reward of TEN POUNDS will be paid upon conviction of the Offender. All Maſters of Veſſels and others are forewarn'd not to Tranſport them from the City, as I am reſolved to proſecute as far as the Law will allow. WILLIAM BULL.

N. B. If the Negroes return, they ſhall be pardon'd. - 88

(a)

He's one of the busiest men in town. While his door may say *Office Hours 2 to 4*, he's actually on call 24 hours a day.

The doctor is a scientist, a diplomat, and a friendly sympathetic human being all in one, no matter how long and hard his schedule.

According to a recent Nationwide survey:
MORE DOCTORS SMOKE CAMELS
THAN ANY OTHER CIGARETTE

(b)

(c)

Figure 2: Print ads depicting social norms of their time:
2a: 1763 (runaway slave) [Schomburg Center for Research in Black Culture, 1997],
2b: 1946 (Camel Cigarettes) [Center for the Study of Tobacco and Society, 2019],
2c: 1953 (Alcoa Aluminum) [Brown et al., 2018]

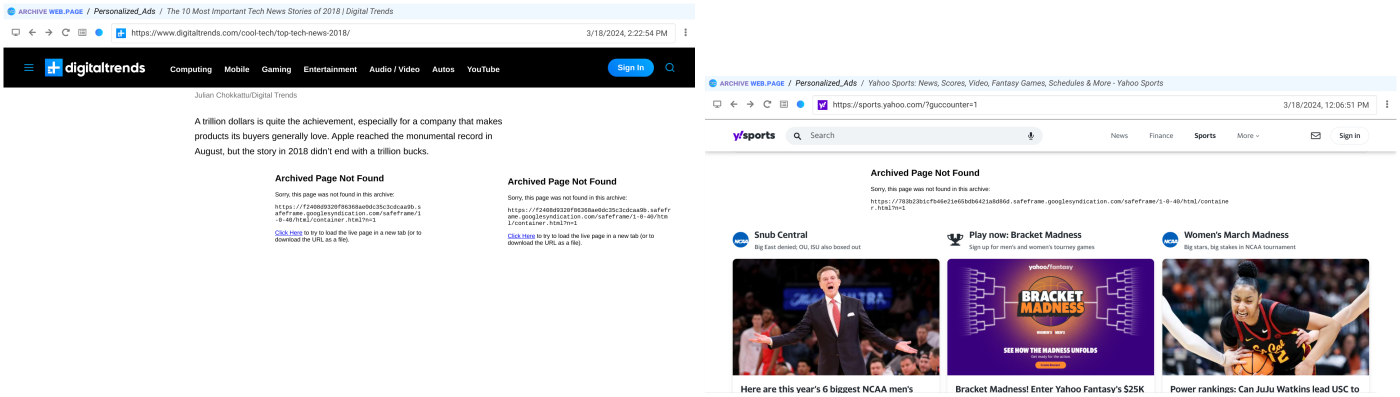
and politics will remain incomplete, even misleading. Their preservation is therefore crucial for scholars and laypersons exploring historical contexts and media ecologies. Recognizing the importance of advertisements in context as digital cultural artifacts can enhance not only historical research, but also archival principles and practices and public engagement with cultural heritage and memory. This will allow future generations of researchers and laypersons to understand different periods' social, cultural, and political influences and legacies, to make informed, empathetic historical comparisons, to channel those comparisons into practice, even policy, and to apply the novel principles and practices derived from web ads to other dynamic content.

The technical challenges of archiving web ads—in fact, of archiving all dynamic web content—are multifaceted. Most notably, web archiving requires not just the storage of the past web but also the ability to accurately replay the capture in future browsers. Enabling future replayability of these ads demands new tools that embody archival principles and facilitate archival best practices. As part of this tool development, the project team will construct a dataset of online advertisements that reflects a broad range of social, cultural, and political perspectives.

1.1 Web Archiving Practices: Current Progress and Future Priorities

Unlike print ads, web platforms cater ads to user types. But even institutions that have performed web archiving have neglected this diversity; their software-based crawlers usually capture only a single experience–and not even one that any human likely ever encountered. Building upon our research on personalized representation in web archives and work supported by our IMLS Planning Grant,[1] we have begun to remedy this remarkable oversight [Kelly et al., 2018, Kelly et al., 2013a, Berlin et al., 2017, Weigle et al., 2017, Kelly, 2016, Kelly, 2017, Kelly et al., 2013b]. Indeed, preliminary findings from our Planning Grant underline the severity of this loss. Not only do some of the largest web archiving initiatives eschew web ads, but even archived ads often cannot be (re)displayed in the context of their containing web page due to contemporary replay systems' technical limitations. What is more, many online ad services deploy client-side randomization and/or web address obfuscation to truncate the reusability of ads loaded from contexts outside the original (e.g., ads captured and replayed by a web page). As a result of these obstacles, our record of the past Web and therefore of the past three decades of culture, society, and politics, especially issues centering on race, ethnicity, gender, sexuality, ability, class, region, and religion, remains – and potentially will remain – incomplete, unbalanced, and even misleading. The impending further loss of these cultural artifacts demands immediate action. Confronting the twin challenges of ongoing neglect and rampant personalization, "Preserving Personalized Advertisements for More Accurate Web Archives will surface

[1] NLG-LIBRARIES-FY22 LG-252362-OLS-22


ARCHIVE WEB.PAGE / Personalized_Ads / The 10 Most Important Tech News Stories of 2018 | Digital Trends
https://www.digitaltrends.com/cool-tech/top-tech-news-2018/
3/18/2024, 2:22:54 PM
digitaltrends
Computing
Mobile
Gaming
Entertainment
Audio / Video
Autos
YouTube
Sign In
ARCHIVE WEB.PAGE / Personalized_Ads / Yahoo Sports: News, Scores, Video, Fantasy Games, Schedules & More - Yahoo Sports
https://sports.yahoo.com/?guccounter=1
3/18/2024, 12:06:51 PM
y!sports
Search
News
Finance
Sports
More
Sign in
yahoo!fantasy
BRACKET MADNESS
WOMEN'S MEN'S
SEE HOW THE MADNESS UNFOLDS
Get ready for the action.
Create Bracket

Julian Chokkattu/Digital Trends

A trillion dollars is quite the achievement, especially for a company that makes products its buyers generally love. Apple reached the monumental record in August, but the story in 2018 didn't end with a trillion bucks.

Archived Page Not Found

Sorry, this page was not found in this archive:

```
https://f2408d9320f86368ae0dc35c3cdcaa9b.s
afeframe.googlesyndication.com/safeframe/1
-0-40/html/container.html?n=1
```

Click Here to try to load the live page in a new tab (or to download the URL as a file).

Archived Page Not Found

Sorry, this page was not found in this archive:

```
https://f2408d9320f86368ae0dc35c3cdcaa9b.safefr
ame.googlesyndication.com/safeframe/1-0-40/htm
l/container.html?n=1
```

Click Here to try to load the live page in a new tab (or to download the URL as a file).

Archived Page Not Found

Sorry, this page was not found in this archive:

```
https://783b23b1cfb46e21e65bdb6421a8d86d.safeframe.googlesyndication.com/safeframe/1-0-40/html/containe
r.html?n=1
```

Snub Central
Big East denied; OU, ISU also boxed out

Play now: Bracket Madness
Sign up for men's and women's tourney games

Women's March Madness
Big stars, big stakes in NCAA tournament

Here are this year's 6 biggest NCAA men's

Bracket Madness! Enter Yahoo Fantasy's $25K

Power rankings: Can JuJu Watkins lead USC to

Figure 3: Two archived pages where advertisements were not captured and could not be replayed due to the use of Google SafeFrame ad delivery; cryptic "Archived Page Not Found" segments are shown instead.

the sociotechnical challenges involved in the preservation of and subsequent access to web advertisements. This exploration will build seamlessly on our Planning Grant.

Efforts to archive web content must consider the original context and interactivity of web pages and advertisements. Preservation strategies include emulating original contexts, combining archival captures with comprehensive web crawls, and employing browser-based crawlers to manage dynamic content. Recognizing the importance of advertisements in context as cultural artifacts can enhance archival practices and historical research, allowing future researchers to make informed historical comparisons and understand different periods' cultural influences.

Our recent efforts showcase significant progress in the preservation of dynamic web advertisements. We have created a dataset of 279 replayable ads [Rauch et al., 2024] and demonstrated the ability to capture and replay these ads in their original context. Jumping off this work, we are exploring the extent to which existing archives capture elements of dynamic advertisements but by default, do not display them. One instance involves ads that are loaded into a Google SafeFrame [Google Ad Manager, 2024]. We have demonstrated that many of these ads have been archived, but because of the nuances of how the ads are loaded into the frames when rendered, they are not able to be replayed from the archive. Figure 3 shows two such examples. These webpages were captured with a browser-based crawler, but the advertisements could not be replayed because of Google's SafeFrame delivery mechanism. This is indicated by the presence of "safeframe.googlesyndication.com/safeframe" in the replay error message. By analyzing WARC[2] files from the Internet Archive and Common Crawl dataset, we will reconstitute a larger set of original archived ads, addressing the lack of coverage and expanding access to historical web content. This work both enhances the preservation of advertisements and applies more globally to *all* forms of dynamic content, auguring a generalizable approach to processing existing archived content.

An alternative approach to archiving websites using capture tools is to create screenshots during crawl time. The Wayback Machine offers this option as part of the "Save Page Now" feature that allows users to submit a specific URL for capture. If the user opts in, a capture of the home page is created. Other data collections stored by the Internet Archive also contain versions comprising only screenshots. We have collected an illustrative sample of these screenshots.[3] Although the interactivity of the original page is lost, these images provide a way to compare the visual representation of an advertisement in the screenshot against the replayed web page as reassembled by the Internet Archive. We have provided the URLs to the Internet Archive's closest matching archived page representing the homepage of various sites, along with links to the corresponding replayed version that was reconstituted from captured files closest in time.[4]

[2] WARC (Web ARChive) is the standard format for storing the results of web crawls. [International Internet Preservation Consortium, 2022]

[3] `https://github.com/savingads/toolkit/blob/main/data/screenshot_ads.md`

[4] `https://github.com/savingads/toolkit/blob/main/data/compare_screenshots.md`

Additionally, because the Internet Archive and most web crawlers capture content in a hierarchy that resembles the URL of the embedded resource, it is possible to view ads in isolation by visiting the archived versions of these ad server content networks. For instance, it is possible to view archived ads from their source in an ad network, i.e. independently of the pages that contain them. We have collected an exemplary set of archived URLs from ad servers.[5]

Complementing these activities, we are in the process of creating guidelines to assist archivists in utilizing existing tools to capture both the advertisements themselves and salient contextual information related to their placement and audience engagement. This includes metadata capturing the ad server's attempts at personalization (as presented to the crawler), and the identification of elements that vary according to set conditions. We are further developing tools focusing on the reconstruction of pages containing ads that are present in already archived data, but not displayed because of the restrictions mentioned below. Building on this progress and confronting the twin challenges of ongoing neglect and personalization, the proposed project will surface the technological challenges involved in the preservation of and subsequent access to web advertisements.

An initial part of the project grapples with ongoing neglect. Statista [Statistia Research Development, 2023] estimates that 35.7% of web users use ad blockers, software that amends a browser's functionality to prevent ads from being loaded and viewed. As a result, most users rarely if ever experience unadulterated ads (e.g., unblockable ads embedded in videos). To gain insight into users' perspectives on this issue, we will engage these users through surveys and interviews (see Section 2.3.1). Foreground a user-centric perspective, these research activities will discern key features of importance (such as attraction and annoyance) from the user's perspective that contribute to users' interaction (or lack thereof) with web advertisements, as well as the technical methods that enable them.

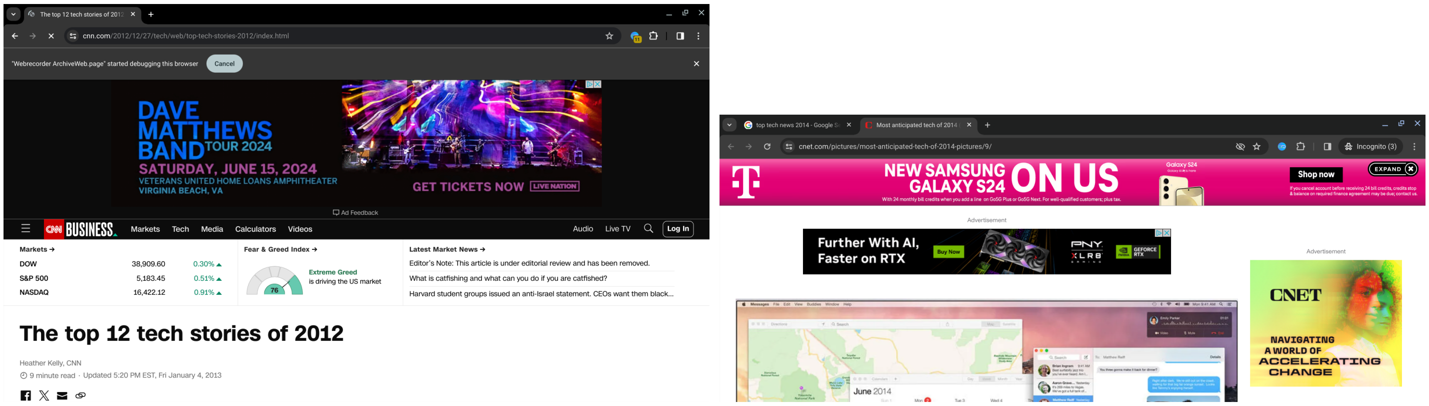

The top 12 tech stories of 2012
cnn.com/2012/12/27/tech/web/top-tech-stories-2012/index.html
"Webrecorder ArchiveWeb.page" started debugging this browser
Cancel
DAVE MATTHEWS BAND TOUR 2024
SATURDAY, JUNE 15, 2024
VETERANS UNITED HOME LOANS AMPHITHEATER
VIRGINIA BEACH, VA
GET TICKETS NOW
LIVE NATION
Ad Feedback
CNN BUSINESS
Markets
Tech
Media
Calculators
Videos
Audio
Live TV
Log In
Markets →
DOW 38,909.60 0.30% ▲
S&P 500 5,183.45 0.51% ▲
NASDAQ 16,422.12 0.91% ▲
Fear & Greed Index →
76
Extreme Greed is driving the US market
Latest Market News →
Editor's Note: This article is under editorial review and has been removed.
What is catfishing and what can you do if you are catfished?
Harvard student groups issued an anti-Israel statement. CEOs want them black...
The top 12 tech stories of 2012
Heather Kelly, CNN
9 minute read · Updated 5:20 PM EST, Fri January 4, 2013
top tech news 2014 - Google Se
Most anticipated tech of 2014
cnet.com/pictures/most-anticipated-tech-of-2014-pictures/9/
Incognito (3)
T
NEW SAMSUNG GALAXY S24 ON US
With 24 monthly bill credits when you add a line on Go5G Plus or Go5G Next. For well-qualified customers; plus tax.
Galaxy S24
Shop now
EXPAND
If you cancel account before receiving 24 bill credits, credits stop & balance on required finance agreement may be due; contact us.
Advertisement
Further With AI, Faster on RTX
Buy Now
PNY XLR8 GAMING
GEFORCE RTX
Advertisement
CNET
NAVIGATING A WORLD OF ACCELERATING CHANGE
June 2014

Figure 4: Two examples of personalized ads: one is location-based (a 2024 concert in Virginia for a CNN.com page from 2013) and one is based on browsing history (a current NVIDIA graphics card for a 2014 article).

Second, we will address the challenges posed by personalization. Representations of personalized preserved content are those that an archival crawler retrieved [Kelly et al., 2013a]. These representations differ, however, from what a live web user originally saw. Most displayed content, including web ads, is personalized for the real-time user. An archival crawler that lacks a baseline browsing history (e.g., visits to Amazon.com), would never receive these personalized, tailored resource representations (e.g., an image advertisement for the last thing you searched Amazon for). Personalization therefore produces a unique fingerprint. If two users, while logged onto to other web services (e.g., logged into Google, even if not visiting Gmail) visit the same web page at the same time, or follow up on a previous browser history in that same session, they will receive different resource representations (e.g., different images as advertisements). As an example, Figure 4 shows ads that were observed and captured by a computer science student who lives in Virginia Beach, VA and who had been shopping for a new NVIDIA graphics card while logged into his personal Google account. After logging into a different Google account, he no longer encountered the NVIDIA ad, even when browsing the same webpages.

[5] https://github.com/savingads/toolkit/blob/main/data/wayback_examples.md

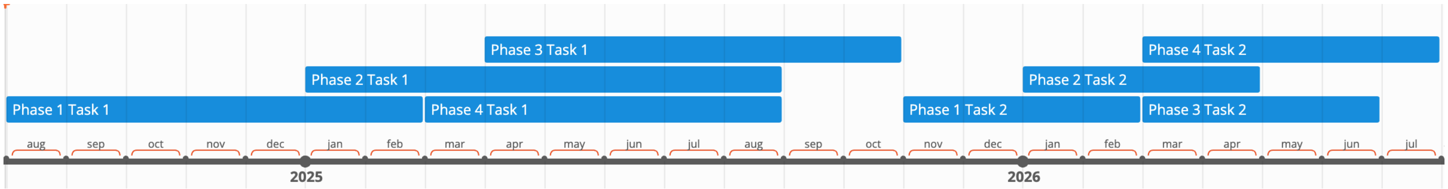


Figure 5: The four phases each contain two tasks with the latter of the two in each phase able to be informed by the findings from the completion of Task 1 for each phase.

Personalization, in short, precludes the existence of one canonical representation of a web page. As a result, each composite representation that each user, both present and past, has been served – including the one served to the Internet Archive's archival crawler – is equally valid. The accumulation of the varying aspects of these representations (e.g., the different ads that are served to different users), increases the aggregate representations' richness. Preservation strategies must accommodate this richness to capture a more comprehensive documentary record of the web. We will do so through human subjects research using a persona-based approach (see Section 2.1.1).

Our efforts to date have primarily focused on overcoming the technical barriers to archiving embedded web advertisements. Even in the absence of explicit institutional policies against ad archival, the tools currently available to web archivists are inadequate for capturing and presenting ads in a manner that historians can readily utilize. Our attempts at archiving and replay have resulted in the capture of only a fraction of most ad components. Even when ads are successfully archived, they often cannot be replayed without extensive customization of the replay environment and the application of manual adjustments. This makes them largely inaccessible to historical analysis. Therefore, the requirements for preserving and replaying web advertisements require additional investigation that fruitfully extends our Planning Grant approaches and deliverables.

Phase 1: Persona development and data collection
- Task 1 (P_1T_1): August 1, 2024 – February 28, 2025 Persona development and data collection
- Task 2 (P_1T_2): November 1, 2025 – February 28, 2026 Informed persona supplementation & collection based on Phase 3

Phase 2: Evaluation of data relative to web archives' holdings
- Task 1 (P_2T_1): January 1, 2025 – August 31, 2025 Evaluation of P_1T_1 data set re: archival holdings
- Task 2 (P_2T_2): January 1, 2026 – April 30, 2026 Evaluation of P_1T_2 data set of archival holdings

Phase 3: Evaluation of captures based on users' expectations
- Task 1 (P_3T_1): April 1, 2025 – October 31, 2025 Survey/Interview design and execution
- Task 2 (P_3T_2): March 1, 2026 – June 30, 2026 P_1T_2 user validation

Phase 4: Results dissemination and technical web archive supplementation
- Task 1 (P_4T_1): March 1, 2025 – August 31, 2025 Packaging of preliminary results for publication & dissemination
- Task 2 (P_4T_2): March 1, 2026 – July 31, 2026 Packaging of final results, report on RQs answered, detail future work

2 Project Work Plan

During our Planning Grant, we learned that the initial goal of creating a list of advertisement URLs for users to visit based on existing archives was more difficult than originally envisioned. Since the majority of public archives either do not preserve ad content or are unable to replay it [Rauch et al., 2024], we have built a preliminary dataset of web ads in the context of web pages as complete WACZ representations [Kreymer and Summers, 2021]. A newly developed standard format, the WACZ file format permits further context to be retained (beyond that provided for by WARC). This work suggests to other web archivists how they might capture web ads and share them through file exchanges or downloads. Extending this initiative, the proposed project will develop a modifiable replay system based on existing tools. We will host the system on a GitHub repository, enabling other stakeholders to create reproducible copies.

Preserving Personalized Advertisements for More Accurate Web Archives will comprise four overlapping phases: (1) persona-driven longitudinal data collection; (2) quantitative analysis to determine the extent of missing advertisements; (3) qualitative, user-informed evaluation of unique representations; and (4) a research study on the opportunities and challenges of archiving the personalized web and disseminating archived representations of it.

2.1 Phase 1: Persona development and data collection

In phase 1, PI Kelly will extend previous work [Kelly et al., 2013a] on identifying personalized representations on the past web. The project team will iteratively create diverse web user personas[6] to highlight ways in which various demographic, social, and cultural factors impact various communities' encounters with web archives. These personas will represent the ways in which users from various racial, gender, class, ability, regional, and religious backgrounds engage the contemporary live web. We will then capture the web ads served to each persona, recording the distinctive characteristics of each ad. This latter task will be iteratively informed by the other three phases (see Figure 5).

2.1.1 Phase 1 Task 1 (P_1T_1): Persona development and data collection

Time Frame: August 1, 2024 – February 28, 2025

The first task of the project's first phase will determine the extent to which personalized captures exist in web archives (RQ1). It will also unpack what features characterize this personalization and how these features compare to the representativeness of the archival crawler as served to a contemporary web user. Developing these perspectives via personas will provide for diversity of representation in archives' holdings (RQ3).

To identify various crawler perspectives, we will leverage the collections of advertisements and their respective contexts developed and disseminated in "Saving Ads." Given the difference between a software-based crawler and a web user as exhibited through a user-agent (e.g., web browser), we anticipate the base persona to be an informative benchmark for developing subsequent perspectives and personas.

2.1.2 Phase 1 Task 2 (P_1T_2): Informed persona supplementation and collection based on Phase 3

Time Frame: November 1, 2025 – February 28, 2026

The second task in Phase 1 will be another iteration of (P_1T_1), informed by the findings in Task 1 of each of the other phases, i.e., P_1T_1, P_2T_1, P_3T_1, and P_4T_1. In this phase, we will learn from our ongoing methodological refinements, reflect on programmatic process to identify and evaluate the captures' representativeness, and perform a second round of data collection.

2.2 Phase 2: Evaluation of data relative to web archives' holdings

In *Phase 2*, we will quantify the lack of coverage that our persona analysis of Phase 1 revealed (RQ1). PIs Weigle and Nelson will lead Phase 2, which will draw upon the previous phase to inform the scope of persona-driven data collection. Complementing this quantification of lack of coverage, we will examine the extent to which the existing archives have captured elements of dynamic advertisements but do not (by default) display them. The goal is to find components in the archive files that permit reconstitution of the original ads from the downloaded WARC files containing the captured web pages. Because various captures of URLs tagged for retrieval may not be retrieved during the same crawl, and because some elements of advertisements might be common to other pages, the processing of the WARC files available through the Internet Archive, the Common Crawl dataset, and other archives presents a further opportunity for redressing the identified lack of coverage. Notably, if we demonstrate a relatively simple procedure to reconstitute content that would not play by default when visiting the URL of a page within a web archive, we can show that our processing approach is generalizable to other types of dynamic content and, therefore, expand access to the historical web more broadly.

[6] The number of personas is to be determined but will be initially seeded as a count between 20 and 50 and further informed by Phase 3 Task 1.

2.2.1 Phase 2 Task 1 (P_2T_1): Evaluation of P_1T_1 data set re: archival holdings

Time Frame: January 1, 2025 – August 31, 2025

The first task of Phase 2 will center on initial evaluation of the Phase 1 data set. We will focus on metrics of completeness of archival captures (i.e., composite mementos) using both manually evaluated and computationally driven approaches. The former method is particularly important for web advertisements that might require direct user interaction to surface interactive, JavaScript-driven representations for the archival and replay processes.

2.2.2 Phase 2 Task 2 (P_2T_2): Evaluation of P_1T_2 data set of archival holdings

Time Frame: January 1, 2026 – April 30, 2026

The second task of Phase 2 will reprise the first task of Phase 2, except for evaluation of the informed data collection procedure invoked in Phase 1 Task 2 (P_1T_2).

2.3 Phase 3: Evaluation of captures based on users' expectations

In *Phase 3*, we will conduct human subjects research (IRB approval is currently pending). We have developed a protocol to capture a comprehensive view of the interviewees' experiences with web archiving. Our selection of interviewees is based primarily on a systematic literature review and web searches that ferreted out stakeholders whose institutions archive dynamic websites and/or who have experience in assessing the cultural and historical value of advertisements. We have crafted both an outreach email template for targeted individuals and a more general solicitation for posting on relevant archives-related listservs such as the IIPC, the Digital Preservation Coalition, and the Archive Team.

Through purposive sampling, PI Poole will conduct qualitative interviews of diverse (e.g., race, ethnicity, gender, class, nationality, ability, geographic location) archivists, researchers, and lay web users to enable holistic, interdisciplinary input and feedback regrading previous phases' findings. In particular, this sample population will juxtapose earlier phases' captures to shed light on the interpretive impact of viewing one of these archived representations, on the one hand, with that of an identically preserved web page denuded of ads, on the other. We will subsequently compare the data generated by the Phase 1 personas with that generated through the interviews and surveys to inform iterative development of additional diverse personas. These additional personas and use cases will constitute a second Task (P_1T_2) in the scope of the Phase 1 data collection; they will supplement the data and evaluation conducted in P_3T_1. Upon completion of the first Task of Phase 3 (P_3T_1), which will help to answer RQ3, we will execute another iteration of Phases 1 and 2 (Tasks P_1T_2 and P_2T_2 respectively). A validation procedure of this complementary data, informed by P_3T_1, will allow us to answer RQ2.

2.3.1 Phase 3 Task 1 (P_3T_1): Survey/Interview design and execution

Time Frame: April 1, 2025 – October 31, 2025

Involving actual web users as opposed to a web browser, phase 3 will be the basis for answering RQ3. Our preliminary findings [Rauch et al., 2024] suggested that the latter may be problematic. In other words, when the same data and information is passed using HTTP transaction clients in two different browsers, users may experience different replay representations. To counter this, we will consult users directly to evaluate their sentiments and their perspectives on the potential importance of advertisements on the live and archived Web.

In advance of this process as informed by our persona development in P_1T_1, the PIs will secure approval from our respective Institutional Review Boards (IRB). This will ensure that our work with human subjects meets the highest ethical standards.

This phase, led by PI Poole, will inform P_1T_2 for further persona development. We expect that our interviews will help inform previously neglected personas. We will interview a wide range of web users, ranging from web archiving practitioners to casual web users, and focusing on diverse demographics (age, race, ethnicity, geographic region, gender, web experience, and so forth).

The project team organized the protocol questions according to the following themes:

Establishing Basic Web Archiving Experience: These questions will elicit the interviewee's experience and expertise with different types of archives, their previous key projects, and any shifts in their archival focus over time. This sets the stage for a deeper discussion on their specific experiences and lessons learned in web archiving.

Exploring Depth and Nature of Web Archiving Experience: Tools and Practices: This section will probe the practical aspects of the interviewee's web archiving, including the use of specific tools, the impact of technological shifts, and the size and scope of collections.

Challenges and Strategies: These questions will tease out information on strategies employed and challenges encountered, especially concerning the archiving of web advertisements.

Focusing on Archiving Web Advertisements: Specific Projects and Collaborations: These questions focus on projects related to web advertisements, exploring the nature of these projects, collaborations, and the proportion of web archives dedicated to advertisements.

Appraisal, Acquisition, and Accessioning: Selection and Integration: These questions will target the criteria for archiving web advertisements, policies guiding their capture, and the integration of these advertisements into archival systems.

Management and Preservation: Challenges and Techniques: These questions will addresses the technical, administrative, and preservation challenges associated with web archiving, with a focus on advertisements. We will also explore strategies for maintaining the fidelity and integrity of archived content.

Security, Description, Access, and Outreach: Ethical Concerns and User Engagement: These questions will explore security and privacy concerns, how online advertisements are described and accessed, and outreach efforts to promote the use of these archives.

Use of Archived Advertisements (specifically researchers): These questions will help understand how the archived advertisements are used for research, the user experience, and the potential scholarly value of web-based ads.

2.3.2 Phase 3 Task 2 (P_3T_2): P_1T_2 user validation

Time Frame: March 1, 2026 – June 30, 2026

This task will allow us to reflect on the representativeness of our own data collection and on our informed definition of personas. By utilizing the feedback loop from Phase 3 Task 1 to Phase 1 Task 2, we can then evaluate the extent to which the sampling of personas is exhaustive. Since a complete representation is difficult to evaluate, introspection on our processes and methodology for establishing and refining the personas and web users will inform additional research in the area.

2.4 Phase 4: Results dissemination and technical web archive supplementation

The project team continues to develop shareable deliverables. In addition to PI Poole's participation in a Digital Humanities panel titled "Evaluating the Value of Exploratory Tools in Digital Humanities Collections and Scholarly Projects: Discussions from Researchers, Developers, and Users' Perspectives" [Ma et al., 2023] at the 2023 Association for Information Science and Technology (ASIS&T) annual meeting in London, UK, the team was recently invited to present a workshop at the IIPC 2024 General Assembly and Web Archiving Conference at the National Library of France, Paris (April 25, 2024). In this presentation, we will describe the current state of archival practice as it relates to the preservation of web-based advertising in context, highlight the large gap in the historical record created by the failure to archive such advertisements, and suggest approaches archivists can use to ensure that the temporally and culturally relevant information sources represented by web advertisements are preserved. Attendees will become familiar with the settings in popular tools that control the scope of archival crawling, techniques to broaden the scope of capture to include advertising materials with a reasonable tradeoff to additional storage requirements, and understand how already archived data can be coaxed into replaying web advertisements that are not rendered by default.

In Phase 4, we will compare our project's captures to those of previous web advertising efforts. This will ensure our efforts add to the archival record, namely in advancing a more complete heritage and historical record. Further advancing IMLS's goals for advancing collections stewardship and access, we will license all data through Creative Commons.

2.4.1 Phase 4 Task 1 (P_4T_1): Packaging of preliminary results for publication & dissemination

Time Frame: March 1, 2025 – August 31, 2025

Our work will be disseminated in several formats. First, using our approach and modified tools, we will offer a set of WACZ archives for the websites we have successfully captured. These archives will be available for download and can be (re)used to replay the advertisements originally captured. Second, we will release the set of tools and scripts that we employed to successfully replay the content and ensure it conforms to the replay tool's requirements. Third, we will publish the scripts used to reconstitute archived pages from larger web archive collections in WARC format. This package will include both the scripts that guide the process of reassembling past web pages from a series of captured crawler transactions and the methodology used to forensically reconstruct elements of the pages not initially stored in the same WARC file. These tools will enable users to parse large data files using PySpark and other large data manipulation tools to identify content recognized as advertisement-related and consolidate it into discrete WARC packages that represent individual websites. For websites reconstituted in this manner, i.e. that have elements of temporal disagreement (certain elements were archived at a different date or time but might still be useful in understanding the advertisement context), we will present metadata indicators of the time divergence and highlight the content that was included post-crawl.

2.4.2 Phase 4 Task 2 (P_4T_2): Packaging of final results, report on RQs answered, detail future work

Time Frame: March 1, 2026 – July 31, 2026

As a product of our planning grant, we have prepared and disseminated a technical report that sets forth our preliminary conclusions concerning the processes required to replay complex archived content. This report identifies common problems in archiving and replaying technical aspects exhibited in online advertisements (and elsewhere on the web) and describes potential solutions to replicating past web experiences that contain these ads. It is a ”living” document that will continue to capture our technical progress; revisions will take place as we complete the various phases in the proposed project. We will share further new insights more broadly through publications and conferences targeted at the archives and preservation communities, in forums such as the International Conference on Digital Preservation (iPRES), the Annual Meeting of the Association for Information Science and Technology (ASIS&T) and the corresponding journal (*JASIST*), the ACM/IEEE Joint Conference on Digital Libraries (JCDL), and other key venues to computer scientists, social scientists, humanities scholars, and archival studies and library and information science scholars. We will also seek to disseminate our work to various computer and library and information science and public history educational outlets and stakeholders. Additional outreach venues will also include listservs like those from the National Digital Stewardship Alliance (NDSA), Digital Library Federation (DLF), and Coalition for Networked Information (CNI); blogging outlets like those from ODU WS-DL[7] and IIPC[8], and interactive platforms like Internet Archive's Slack communication channels. Given the relevance of our work to an extraordinarily broad range of scholarly and lay communities, our initial report will allow practitioners with similar objectives to become aware of degraded replay mitigation strategies as they are developed.

3 Diversity Plan

As in our Planning Grant project, which features two RAs from underrepresented populations, we will prioritize diversity, equity, and inclusion in recruiting both graduate research assistants and participants for our Task 3 user study. Our RA hiring efforts will be facilitated by Drexel CCI's Women in Tech initiative and its DEI Council (on

[7] `https://ws-dl.blogspot.com`

[8] `https://netpreserveblog.wordpress.com/`

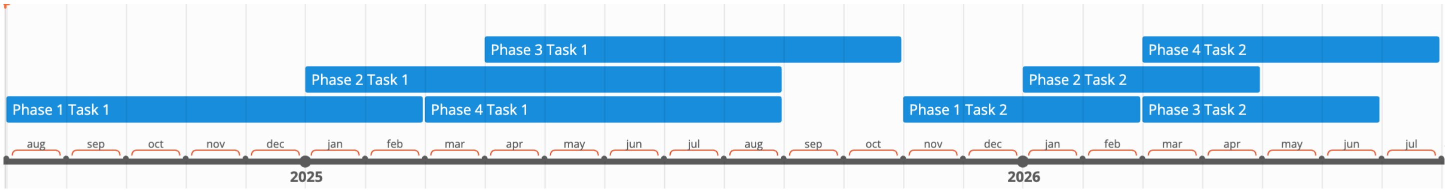
which PI Poole serves), and ODU's ongoing commitment as a minority-serving institution (38% of students from underrepresented ethnic groups).[9] Every grant activity will include a diverse and inclusive (e.g., race, ethnicity, class, gender, region, nationality, ability) pool of students, professionals, researchers, and users to ensure variegated voices and perspectives. As exemplified in Phase 1, moreover, we will emphasize diversity of web users' perspectives. This approach will promote a representative historical record, thereby surfacing portions of the web typically experienced by web users of marginalized or underrepresented groups.

4 Project Results

Through the Phase 1 development of the personas and the execution of the collection procedures, we will develop a robust data set using high-fidelity web archiving capture tools. This dataset will comprise diverse personalized web advertisements and their encapsulating web pages. As noted above, both the dataset and formal and informal publications and presentations illuminating ongoing progress, research findings, lessons learned, and emerging best practices for web archiving will be disseminated to sundry library and information science and computer science communities throughout the project's duration. Overall, this project will advance knowledge and benefit society. Addressing top IMLS priorities, it directly supports balanced stewardship of and promoting access to a rich and variegated part of the cultural, social, and political record. An authentic, inclusive, and diverse record can not only illuminate the current state of and historical reasons for polarization, but help promote constructive public sphere conversations about ameliorating it.

[9] https://www.odu.edu/about/facts-and-figures

Schedule of Completion



Phase 1: Persona development and data collection

- Task 1 (P_1T_1): August 1, 2024 – February 28, 2025 Persona development and data collection
- Task 2 (P_1T_2): November 1, 2025 – February 28, 2026 Informed persona supplementation and collection based on Phase 3

Phase 2: Evaluation of data relative to web archives' holdings

- Task 1 (P_2T_1): January 1, 2025 – August 31, 2025 Evaluation of P_1T_1 data set re: archival holdings
- Task 2 (P_2T_2): January 1, 2026 – April 30, 2026 Evaluation of P_1T_2 data set of archival holdings

Phase 3: Evaluation of captures based on users' expectations

- Task 1 (P_3T_1): April 1, 2025 – October 31, 2025 Survey/Interview design and execution
- Task 2 (P_3T_2): March 1, 2026 – June 30, 2026 P_1T_2 user validation

Phase 4: Results dissemination and technical web archive supplementation

- Task 1 (P_4T_1): March 1, 2025 – August 31, 2025 Packaging of preliminary results for publication & dissemination
- Task 2 (P_4T_2): March 1, 2026 – July 31, 2026 Packaging of final results, report on RQs answered, detail future work

Digital Products Plan

Both tasks in Phase 1 entail data collection as a portion of their digital products output. As with our Planning grant (e.g., `https://github.com/savingads/Recently_Archived_Ads`), we anticipate using GitHub to widely share both the data collected and analyzed as well as the software tools we develop to facilitate the project tasks. For example, in the Planning Grant, we needed to develop specialized software to supplement the replay capability of web archives that were captured but difficult for replay system to present to the end user. All software products we generate will be permissively licensed via the MIT license for reuse without restriction to broaden the impact of our work for further web archiving research beyond our own project. Data collected, while stored in GitHub, will also be redundantly stored in Zenodo[10] for more persistent access and to acquire persistent identifiers for referencing our data.

Digital Product 1: Collected & Analyzed Data

- Type: Web-based resource representations (HTML, JS, CSS); WARC files, WACZ files
- Availability: GitHub repository (while building collection), Zenodo (in perpetituity)
- Access: Initially restricted to assure no PII conflicts, ultimately public and liberally licensed (e.g., CC BY SA)
- Sustainability: Redundant copies on GitHub, Zenodo, and a suitable Drexel-based data repository TBD

Digital Product 2: Software Products

- Type: Python scripts in the form of free scripting and Jupyter-like notebooks for reproducibility
- Availability: GitHub repository
- Access: No access restrictions, permissively licensed (MIT)
- Sustainability: Stored on GitHub using their built-in distributed version control system. Redundantly stored on Drexel's Gitlab instance[11], which is also version controlled, has redundant backups, and is freely available to Drexel faculty, staff, and students.

[10] `https://zenodo.org/`

[11] `https://gitlab.cci.drexel.edu`

Digital/Data Management Plan

This project will generate large amounts of data in the form of archived web pages, stored in WARC and WACZ files. These data are considered “Digital Products” and are described as such in the “Digital Products Plan”. As researchers dealing in web archives, we recognize the importance of data preservation.

The amount of data collected is not initially feasible to estimate, as it is dependent on the discovery and analysis phase of the project. WARC files are typically restricted to 1 Gigabyte per file. We are likely to opt for smaller WARC files to allow for use of the free tier of GitHub without needing to incur the cost of Git’s Large File Storage (LFS) services.

It is possible that data we collect on the web, particularly personalized advertisements, contain traces of identifying information. We have anticipated this aspect of this form of web archiving and expect to have an initial embargo period of our data after the completion of Phase 1 Task 1 until the evaluation procedure of Phase 2 Task 1 can complete. We likewise expect to repeat this process on the subsequent “informed” tasks of data collection (Phase 1 Task 2) and evaluation (Phase 2 Task 2).

If personally identifiable information (PII) is encountered, we will work to anonymize the traces prior to making the data publicly available on GitHub. We expect to also follow the above embargo and anonymization processes with a deposit of the data into Zenodo and a Drexel-based data repository for data reuse. We will provide provenance information with these data sources to provide a trace on the processes that were used for sanitization as necessary.

This necessary process will be codified and documented in our procedures when disseminating the research products (i.e., papers) and revisited in the second iteration of the data-evaluation process described above for consistency of methodology.