LG-256695-OLS-24 - Drexel University (College of Computing and Informatics)

Preserving Personalized Advertisements for More Accurate Web Archives

Introduction: Drexel University's College of Computing and Informatics (Mat Kelly and Alex H. Poole), in collaboration with Old Dominion University's (ODU) Department of Computer Science (Michele C. Weigle and Michael L. Nelson), respectfully requests $407,724 for a two-year National Leadership Applied Research Grant. The proposed project will extend a vital, timely initiative that addresses both the technological shortcomings of and human negligence in archiving advertisements embedded in web pages. The project will focus on personalized ads – those viewed on the live web by users and typically not surfaced for capture by archival crawlers. 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 (Advance Collections Stewardship and Access) and Objectives 3.1 (Support collections care and management) and 3.2 (Promote access to museum and library collections). This project will tackle three primary research questions (RQ):

- RQ1: To what extent do institutional web archives capture personalized advertisements on the web?
- RQ2: Do scholarly or lay web users prefer (re)using archived web pages including personalized ads, a generic comprehensive capture, or a capture with web ads missing?
- RQ3: In what ways might the strategic use of diverse personas surface types of web content that would otherwise go unarchived?

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, and diverse documentation. For example, web advertisements disseminated during the COVID-19 pandemic, e.g., for vaccines, offer visceral cultural, social, and political traces of that deeply traumatic event. Despite the web's overriding significance as a medium of public discourse, however, it is often overlooked as a scholarly and layperson documentary source, as is the deluge of advertisements we encounter in our daily web use. These advertisements fundamentally affect our overall user experience. Exacerbating matters, even those institutions that engage in web archiving often exclude advertisements, erroneously deeming them ephemeral and inconsequential instead of the fundamentally revealing cultural, social, and political documentation that they are.

Even those committed to archiving ads faced an unprecedented challenge. 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 crawler usually capture only a single experience, and not even one that any human likely encountered. Building upon our research on personalized representation in web archives and supported by an IMLS Planning Grant, 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[1] 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 web archive 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 a context outside the original (e.g., ads captured and replayed by a web page). As a result of these exclusions and 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 will remain unless we act—incomplete, unbalanced, and even misleading. The impending further loss of these cultural artifacts demands immediate action. Confronting the twin challenges of ongoing neglect and personalization, the proposed project will surface the sociotechnical challenges involved in the preservation of and subsequent access to web advertisements. These ongoing challenges demand additional exploration. This exploration will build seamlessly on our Planning Grant.

Project Work Plan: Our project will be executed in 4 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.

In *Phase 1*, we will iteratively create diverse web user personas 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, thus recording the distinctive characteristics each ad features for each persona. In *Phase 2*, we will quantify the lack of coverage that our persona analysis of Phase 1 revealed, thereby permitting us to answer research question 1. In *Phase 3*, we will recruit through purposive sampling demographically diverse

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

web archivists and users to evaluate findings drawn from previous phases. More specifically, this sample population will juxtapose earlier phases' captures to judge the interpretive impact of viewing one of these archived representations, on the one hand, with that of an identically preserved web page denuded of web ad, 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. 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 genuine and complete picture of the past. Further advancing IMLS's goals for advancing collections stewardship and access, we will license all data through Creative Commons.

PI Kelly will extend his previous work on identifying personalized representations on the past web in Phase 1 [Kelly et al., 2013a]. PIs Weigle and Nelson will lead Phase 2 with progressive overlap from the previous phase to inform the extent of persona-driven data collection. In Phase 3, PI Poole will interview diverse archivists, researchers, and lay web users to enable holistic, interdisciplinary input and feedback. In Phase 4, all PIs will integrate the personalized web ads data collected (Phase 1), identify the extent of our data's uniqueness (Phase 2), evaluate the representativeness of our collection (Phase 3), and make our data widely available.

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

Diversity Plan: As in our Planning Grant project, which features two RAs from underrepresented populations, we will prioritize diversity, equity, and inclusion in recruiting 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 Dean's DEI Council (on which PI Poole serves), and ODU's ongoing commitment as a Minority-Serving Institution. Every grant activity will include a diverse and inclusive pool of students, professionals, researchers, and users to ensure a true diversity of voices and views.

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. Not only the dataset, but formal and informal publications and presentations illuminating ongoing progress, research findings, lessons learned, and emerging best practices for web archiving will be disseminated to the library and information science and computer science communities throughout the project's duration. Overall, this project will advance knowledge and benefit society. 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.

Budget Summary: The budget request of $407,724 accounts for all anticipated costs. Direct costs are $335,241 for two years' salaries of one PhD student at Drexel ($64,560), one hourly graduate student at Drexel ($32,280), three GRAs (two in year 1 and one in year 2) over two years at ODU ($82,667), $23,219 for PI/Co-PI salaries at Drexel ($11,746.00) and ODU ($20,085), and $47,328 for tuition remission for each PhD student at Drexel ($19,920) and ODU ($27,408) for two years. Fringe costs are $4,111 (Drexel - PIs) and $6,374 (ODU - PIs, student). Indirect costs are $72,482 (Drexel) and $62,219 (ODU).