LG-252362-OLS-22, Drexel University (College of Computing and Informatics)

Saving Ads: Assessing and Improving Web Archives' Holdings of Online Advertisements

Project Justification: Drexel University, in collaboration with Old Dominion University (ODU), respectfully requests $149,432.21 for a two-year National Leadership Planning Grant to systematically evaluate the current state of the art of archiving advertisements (ads) embedded in web pages, and to better understand why and how web archiving technology must improve to meet the current and future needs of researchers. This project aligns with Program Goal 3 and Objectives 3.1 and 3.2 by improving archives' capacity for providing access to and use of collections of born-digital web content.

Advertisements have been indicators of cultural heritage and norms through time (Lears 1995, Gardner 2006, Silvulka 2011) and provide foundational source material for social, cultural, and business history, especially in unpacking vital research questions concerning race, ethnicity, gender, and socioeconomic class. Online advertisements are no less important for this purpose than traditional print versions. However, the Internet Archive and other major public web archives are failing to capture most, if not all, embedded advertisements in their archived web pages (Weber 2017, Brügger 2017). Historians have used web archives to study the recent past (Milligan 2013, Milligan 2019), but online advertisements have been overlooked. For example, while the web archive collections at Library of Congress (LC) may have web ads intermixed, our exploratory discussions with representatives from LC (e.g., Abigail Grotke, Web Archiving Team Lead at LC) have indicated that they have not done any focused collecting in the area. Further, while a number of projects have sought to improve archiving of web pages, most digital preservation projects ignore the loss of online advertising.

Our vision for this foundational work is to formally study the extent of this gap through both a quantitative and a qualitative analysis of the current shortcomings of advertisements embedded in archived web pages and the importance of these advertisements to researchers. Our project will explore the paradox of having an abundance of information online at the same time as we face the disappearance of digital artifacts. Embedded advertisements provide historical context, and the loss of digital artifacts not only can prevent reproduction of such artifacts in the future, but also results in the loss of the intertextuality of a web page. When ads vanish, the "original" (per the experience of the user) is gone forever.

This project will serve as planning and preparatory work for a larger study of online advertisements that would investigate how we can better archive advertisements, and, by extension, other types of dynamic web elements, in the future. Providing the ability to create collections of online advertisements will be essential to future scholars who want to study our current times and recent past. The outcomes for this project include 1) two data sets of archived online advertisements, 2) a public interface for browsing the data sets, 3) a baseline quantitative analysis of what ads have been archived and not archived through time, and 4) a qualitative assessment of the importance of what is missing from web archives and the needs of scholars for the future.

One primary goal of this project is to contribute to the involvement of libraries and archives in the role of preserving and providing access to a more complete cultural record within the nation and beyond. Our work will help to develop new technologies to improve how they collect, preserve, and provide access to these potentially incomplete cultural collections to include the cultural context that re-experiencing historical ads provides.

Project Work Plan: Our project team consists of faculty from Drexel Information Science (PI Mat Kelly, Co-PI Alex Poole) and ODU Computer Science's Web Science and Digital Libraries (WS-DL) research group (Co-PI Michael Nelson, Co-PI Michele Weigle) along with PhD student researchers.

While the task of archiving and re-experiencing the dynamic web instilled in advertisements is an ongoing and difficult preservation task, we have scaled the scope of this project to be accomplished within the two-year time frame by focusing our preliminary studies on the archived web corpora to four (4) tasks:

- **Task 1** (lead: Co-PI Nelson) - We will create two data sets of archived advertisements. The first data set will be gathered from existing web archives to map the trends in the rise and fall of ads in archived web pages beginning in 1996 through present day. The second data set will be collected from the live web and archived using browser-based tools, such as Conifer.
- **Task 2** (lead: Co-PI Weigle) - We will develop a public interface for browsing the data sets of advertisements from Task 1.
- **Task 3** (lead: PI Kelly) - We will perform a quantitative analysis of ads found in the data sets from Task 1 to establish a baseline of what has been archived and what is missing. This will include a historical analysis of what has been archived in the past as well as an analysis of how well we can archive today's ads.
- **Task 4** (lead: Co-PI Poole) - We will perform a qualitative analysis based on information gathered from humanities stakeholders to assess how scholars across the disciples in the humanities value web ads, utilizing the interface developed in Task 2.

PI Kelly will lead and coordinate the overall effort, and he and the Drexel PhD student will be involved in all tasks.

The potential users of the data sets span a number of fields from English (particularly rhetoricians and textual studies scholars), to cultural studies, sociology, media studies and communication, history, and political science. We anticipate Tasks 2-4 to be able to be run concurrently following the creation of the data sets in Task 1. The PI and Co-PIs have extensive experience in collecting (Task 1), visualizing (Task 2), and analyzing (Task 3) data from web archives (Ainsworth 2012, Kelly 2013, Aturban 2019, Mabe 2020).

In Task 4, we will investigate the importance of disappearing online advertising and the challenges associated with recording, preserving, and curating such artifacts. While researchers can make tentative claims about the size and scope of the problem, qualitative data from scholars across a number of disciplines in the humanities is needed to garner a more robust assessment of the significance of the problem. Through the use of surveys and interviews, we will better understand what information is required to do their work, and how further progress can be made. This qualitative study will be led by Co-PI Poole based on prior expertise in this domain. The inductive analysis of survey results will involve a coding process based upon grounded theory. The recursive process of grounded theory coding involves valuing a researcher's interpretive strategies and acknowledging the relationship between methods and emergent theory. Furthermore, grounded theory supports transparency in research methods. We expect to identify emergent themes regarding researchers' perceptions of the project data and its analysis as it is relevant to their individual areas of study, and translate these themes as key points in the findings document.

Diversity Plan: We will consider diversity and inclusion in hiring of our graduate research assistants and selecting the pool of scholars for our Task 4 survey. Selection of the graduate research assistants at both Drexel and ODU will be based on a process that prioritizes diversity based on the inclusion of female and minority students. Since 2017, Drexel CCI has had an ongoing strategic effort to increase the number of women across the college through its Women in Tech initiative. In the ODU WS-DL research group, 7 of the 18 PhD students are women and 3 of the 18 are from a minority group. In the larger Department of CS at ODU, 30% of the graduate students are female, and ODU is considered a "minority serving institution" with an enrollment that reflects the demographics of the region. In our qualitative survey, we will strive to include a diverse pool of scholars to ensure that many views are heard and included.

Project Results: The tangible products of this work will be the data sets (Task 1), public interface (Task 2), and the findings documents (Tasks 3 and 4), which are all intended to frame future research and feature development. Drexel and ODU will lead in the creation of data sets, quantitatively analyze these data sets, and be responsible for at least two data sets that can be shared publicly with a permissive license on GitHub to encourage reuse and further exploration in the domain. Our ad data sets will be released with a Creative Commons license in JSON and/or CSV format, with the specific syntax and semantics to be developed during the course of the project. In addition to publishing the data on GitHub during development, we will publish it with Zenodo, where it will receive a DOI, at the end of the project. We will also publicize the data sets and developed interface through blog posts, social media, and presentations at workshops and conferences. The findings and results of this work will be described in blog posts, preprints, and conference papers. In the past, we have used archived email lists, most recently Google Groups, to establish a place for community discussion as well. Also, relevant data and research findings will be disseminated at conferences and through publications; proposed conferences include the Joint Conference on Digital Libraries (JCDL) and THATcamp, The Humanities and Technology Camp. We also plan to submit publications to the International Journal on Digital Libraries (IJDL) and the Journal of Digital Humanities (JDH). We will use our standing in the digital library community to demonstrate and present the work at workshops such as International Internet Preservation Consortium (IIPC) and the annual Archive-It Partners Meeting. Both Drexel and ODU WS-DL are especially attuned to and qualified for the long-term preservation of the resulting data sets and findings. Multiple project team members have a deep knowledge of digital preservation methods and will apply best practices as needed.

While some may question the importance of contemporary advertisements, ads from the past are culturally indicative of contextual aspects beyond the advertisement itself. Because of technical limitations and this likely post-hoc realization, online advertisement are not being preserved to the same extent as other web contents. In time, this loss will denote a gap in the cultural record. Because of the ephemeral and temporal nature of the web, timely preservation of ads fulfills a high priority challenge faced by archives. Initially identifying problematic results of the past, we can inform a more complete cultural record in the future. By proactively exploring the complications and facets of preserving online advertisements, this work can have far-reaching impact on archival practice beyond our focused, seminal, exploratory efforts.

Budget Summary: The budget request of $149,432.21 accounts for all anticipated costs. Direct costs are $65,416.61 for salaries of one PhD student at each of Drexel ($53,166.61) and ODU ($12,250.00), $8,736.05 for PI/Co-PI salaries at Drexel ($6,792.05) and ODU ($1,944), and $18,209.10 for tuition remission for the Drexel student for two years. Fringe costs are $1,657.25 (Drexel - PIs) and $1,935.00 (ODU - PIs, student). Indirect costs total $44,607.20 (Drexel) and $8,871.00 (ODU).