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

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

1 Project Justification

Drexel University, in collaboration with Old Dominion University (ODU), requests 149,657.19 for a two-year National Leadership Planning Grant. We propose to study the current state of web archives' holdings of online advertisements. This project presages future larger-scale work that would propose methods for improving how we preserve online advertisements and other types of dynamic web elements. Our *primary audience* for this work comprises web archivists and other information professionals such as librarians concerned with providing complete collections for scholars. Our *secondary audience* includes scholars in the humanities and social sciences studying the historical impact of online advertisements. This project aligns with the IMLS National Leadership Grants Program **Goal 3** and **Objective 3.2** by improving libraries' ability to provide access to web archive collections that include contextual advertisements.

1.1 Why Advertisements?

Illuminating the mores and norms of a time and place [1, 2, 3], advertisements (ads) provide foundational source material for social, cultural, and business history, especially in unpacking vital research questions concerning race, ethnicity, gender, and socioeconomic class. Figure 1 illustrates this with mainstream print ads from the mid-1900s expressing ideas that would be considered offensive today. Moreover, the juxtaposition of ads and content, as Figure 2 demonstrates, can enhance our understanding of the interplay between the two, and thus of targeting and market segmentation and their continuity or change over time.

The existence of numerous library collections suggests the importance of historical advertising materials. These include Penn State University Libraries' "Advertising: History and Archives" [7], Cornell University Library's "Unexpected Images: Ads" [8], The Smithsonian National Museum of American History's Advertising collection [9], and multiple collections at the Library of Congress [10, 11, 12]. In addition, several dedicated web sites provide the public with browsable access to historical print ads, such as the Vintage Ad Browser [13] and The Advertising Archives [14].

1.2 Why Online Advertisements?

Online advertisements have similar—if not greater—cultural significance and impact as print advertisements. Just as physical ephemera in libraries, archives, and museums fuels compelling research, so too do online ads illuminate

Figure 1: Print ads depicting social norms of their time: 1946 (Camel Cigarettes) [4], 1953 (Alcoa Aluminum) [5], 1965 (Parker Pens) [5]

the contemporary objectives of advertisers, social norms, viewpoints, and ideals in ways that carefully curated news stories cannot. For example, embedded ads for masks have proliferated in our web browsing since the Spring of 2020 (Figure 3). The mere presence of these ads in articles unrelated to COVID-19 underscores the pervasive nature of masks in our daily lives in a way that stating "there were mask mandates and companies sold cloth masks during the COVID-19 pandemic" does not.

But the Internet Archive and other major public web archives are failing to capture most embedded advertisements in their archived pages (discussed further in Section 1.4). Formally studying the extent of this gap is a fundamental goal of this project. The images in Figure 3 underline the impact of this loss. The first two images are screenshots of an online article containing embedded ads captured at different times in March 2021. The first shows an ad for cloth masks, which uniquely identifies its time period in history and provides insight into what the world was like in 2020-2021. The second image shows the same article viewed by a different user after they visited web pages related to mortgage refinancing. First, this is a prime example of personalized ads based on a user's browsing behavior. Second, although mortgage refinancing ads have and will be around for longer than mask ads, this particular ad shows current mortgage refinancing rates, which orients the ad to a certain moment in time. Finally, the third image,an archived copy of the front page of CNN.com from June 17, 2009, shows what happens when such embedded ads are not archived. There are two missing ads, indicated by the red arrows. Since the ads were not archived, we will never know what their presence might reveal about 2009.

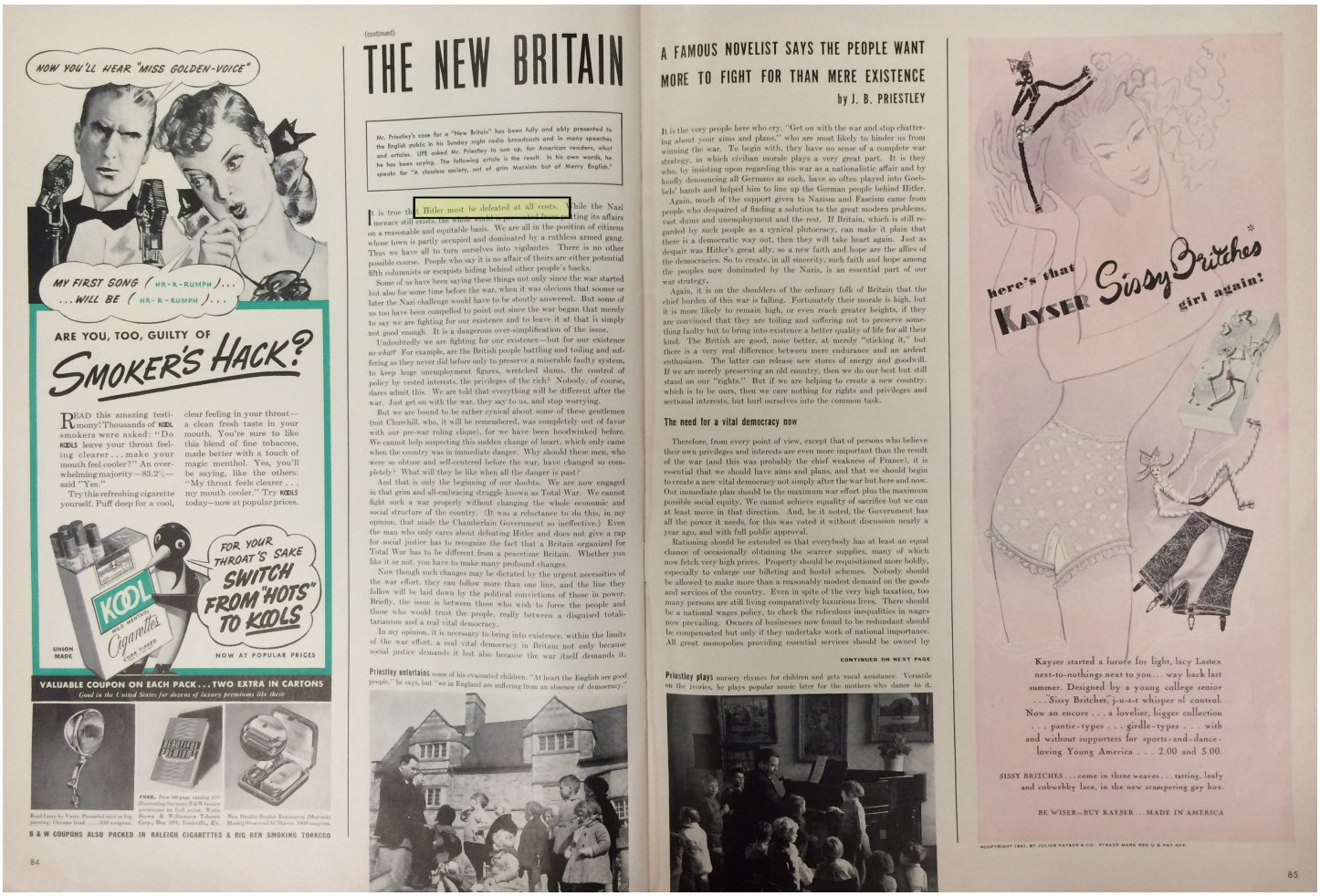


Figure 2: Meaningful juxtaposition showing a cigarette ad to the left, an underwear ad to the right, and an article in the middle of the page regarding Hitler prior to American involvement in World War II. Note the similarity between the juxtaposition of the main article and ancillary content in this physical representation and in web pages of today. [6]

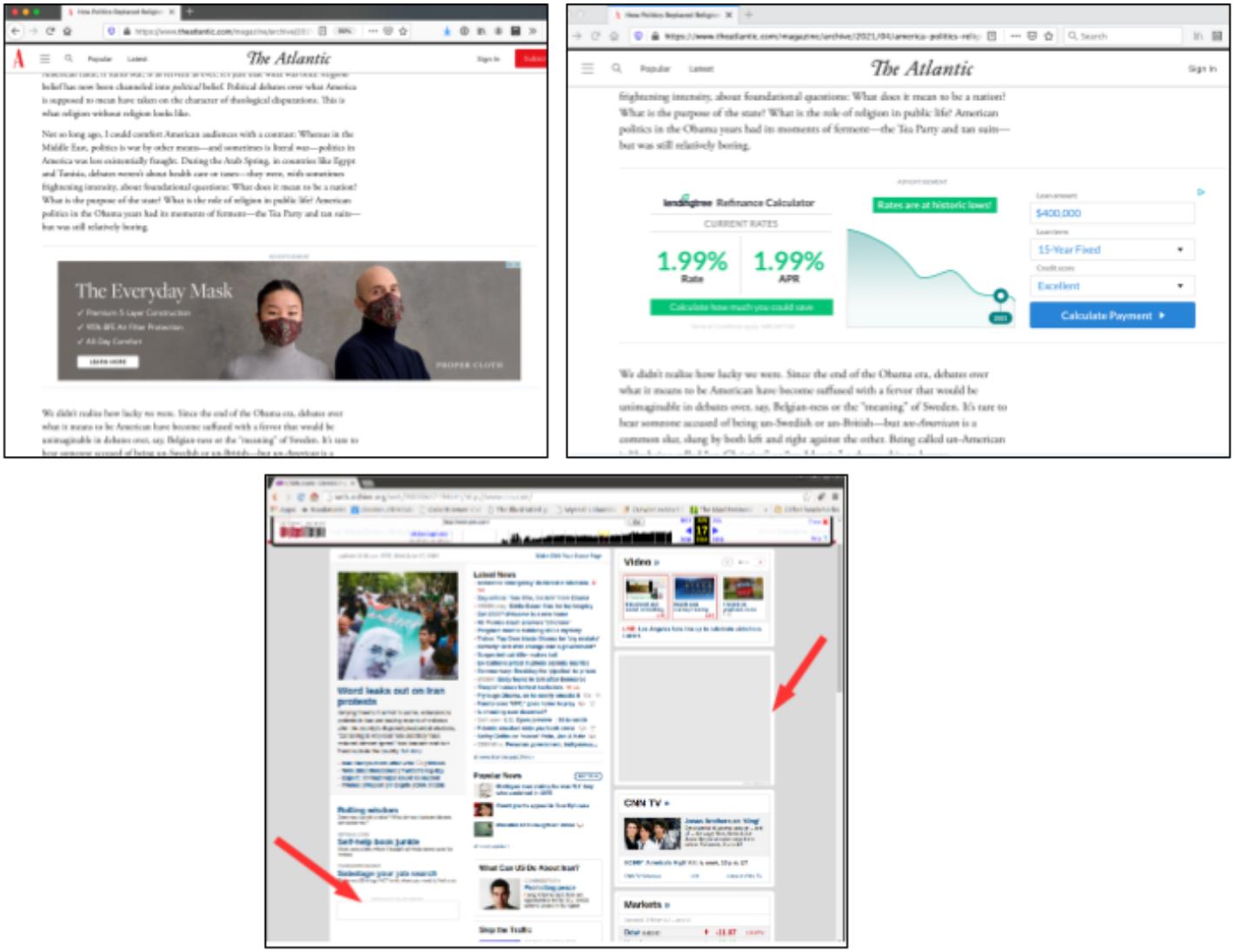


Figure 3: Contemporary ads and missing ads: ad for masks in March 2021, ad seen by a different person for mortgage rates in March 2021, a copy of the CNN homepage from June 17, 2009 in the Internet Archive with missing ads.

Web history remains inchoate, even though it offers unprecedented scale and scope—not only more data, but more sources of data [15, 16]. Researchers can now probe new research questions and revisit familiar ones in new ways [17]. Some historians have used web archives to study the recent past [18, 19], but online advertisements have been overlooked [20, 21, 22]. Historians' myopia comes with a cost, as histories of the 1990s—and beyond—written without web archives will remain grievously flawed [23]. Libraries and web archivists have a strong interest in being able to provide these types of resources for scholars, as scholars cannot study what has not been preserved. To highlight the need to preserve online advertisements, two scholars whose research involves studying current events on the Web, Dr. Ian Milligan (Associate Professor of History at the University of Waterloo in Canada) and Dr. Jennifer Stromer-Galley (Professor in the School of Information Studies at Syracuse University), have provided letters of support for this proposal.

There has been recent interest in analyzing and preserving some types of online advertising [24]. With the demise of Adobe Flash in 2020, there have been archives created of Flash banner ads [25, 26] and Flash advertising games [27]. In 2018, Facebook, now Meta, created a searchable Ad Library [28] of advertisements actively running on Meta properties, such as Facebook and Instagram. This ad archive, along with those created by Google[1] [29, 30] and

[1]Google's ad library contains only political ads placed in Google's search result pages.

Twitter[2] [31], was prompted by the emphasis on how political advertisements had been used during the 2016 US Presidential election and the 2016 Brexit referendum [32]. Facebook's Ad Library has already been used for multiple studies of Facebook advertising on topics including politics [33], COVID-19 [34], and migration [35]. However, these online ad collections contain only the advertisements themselves, not their surrounding context.

As evidenced in Figures 2 and 3, ads are culturally indicative of contextual aspects beyond the advertisement itself. Because of technical limitations and this likely post-hoc realization, online advertisements with their surrounding context have not been the focus of web preservation projects. Further, although a number of projects [36, 37, 38, 39] have sought to improve archiving of web pages, most digital preservation projects ignore the loss of online advertising. 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.

In physical archives, what is admitted to be included in an archive reflects the values of particular institutions (and funding availability), as well as the professional judgment of archivists. Decisions made at the macro and micro level (institutional and personal) affect the cultural and historical record. 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. With time, this loss of preserved ads will denote a gap in the cultural record. Because of the ephemeral and temporal nature of the web, timely preservation of ads is a high priority challenge faced by archives. For a long time, ads included via JavaScript and iframes (i.e., the modern delivery mechanism for ads) were not crawled by conventional crawlers such as wget and Heritrix, in part because they do not execute JavaScript, and the HTML file (i.e., the Document Object Model (DOM)) never has a chance to have the ads dynamically inserted. This was not without benefit to the web archive: in previous work we measured that using a browser-based crawler (instead of wget or Heritrix) and executing JavaScript and interacting with page (e.g., hovering or clicking on all the menus and buttons) can acquire nearly 16X more web resources and cause the crawler to run 39X more slowly [37, 40, 41, 42]. When storage was expensive and networks were slow, this additional penalty often meant that ads were *something to be avoided* in the web archive – why waste time and space on ads when there are "real" pages left uncrawled? Fortunately, while no archive ever has enough storage, resource constraints are no longer the forcing function they once were. Reflecting this, there is recent increased interest in deploying browser-based crawling (e.g., Browsertrix, Brozzler), to enable JavaScript execution so the entire DOM is built and all URLs in a page are revealed, crawled, and archived [39]. Thus, the time is right for us to consider returning to crawling ads, the forgotten web pages that were eschewed in the interest of short-term efficiency.

1.3 Studying the Gap

The **major goal** of this 2-year Planning Grant is to study the identified gap by analyzing the need for and feasibility of archiving advertisements embedded in web pages. The needs will be assessed through a mixed methods study of librarians, archivists, and humanities scholars to learn what aspects of online advertising in the recent past are important and to gain insight into what aspects of online advertising future scholars might be interested in studying. The feasibility will be assessed through a quantitative assessment of how well different types of ads (e.g., static images vs. dynamically-loaded) have been archived through time and an evaluation of the challenges faced in archiving current and future advertisements. We anticipate this project will serve as planning and preliminary work for larger-scale research and development projects, namely, a proposal for a longer-term IMLS Project or Research in Service to Practice Grant or an NEH/IMLS Digital Humanities Advancement Grant. We want to involve libraries and archives in the role of preserving and providing access to a more complete cultural record. Our work will provide a baseline for the future development of new technologies to improve how these institutions collect, preserve, and provide access to these valuable cultural collections, including their cultural context.

The **outcomes** of this project will include two data sets of online advertisements and their surrounding web pages to be used in this research that can also be shared for further analysis (**outcome 1**), a quantitative baseline and categorization of what ads are and are not archived through time (**outcome 2**), a qualitative assessment of the

[2]Twitter has banned political and issue advertising, but has released text files of ads that ran between May 2018 and Nov 2019.

significance of what is missing from web archives and needs for the future (**outcome 3**), and a summary document that will provide the foundation for future work, including future research questions that could be addressed based on this initial exploration (**outcome 4**).

We see these outcomes as springboards for future work. The data sets will provide a basis for evaluation of the extent to which the ads reflect the mores and proclivities of advertisers of their time. The quantitative analysis could lead to further analysis of the evolution of generalized vs. personalized advertising messages. The qualitative analysis will help us better understand the requirements for future capture technologies.

This project will explore the paradox of having an abundance of information online at the same time as we face the disappearance of digital artifacts. 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 project can have far-reaching impact on archival practice beyond our focused, seminal, exploratory efforts.

1.4 Background: The Interplay Between Web Archiving Technology and Ad Delivery Technology

Most current institutional archiving workflows rely on applications called crawlers, the most popular being the Internet Archive's Heritrix crawler [43]. These crawlers begin with an initial set of resources, search for outgoing links, and then add those references as future archival targets before moving to the next. This process leverages the interconnectivity of sites and pages. However, lacking a clear representation of a user's experience, the crawler only captures the static, stateless nature of pages without the ability to interact with or archive dynamically loaded content. But until approximately 2005, when JavaScript became popularized as an ad delivery method, ads mainly consisted of single static image files that were embedded into web pages based on their content. For instance, a web page about cars might contain ads for a car parts store, like AutoZone. We anticipate that traveling back 15+ years in the archives to pre-2005 yields a fairly decent record of advertising on the web.

With the emergence of JavaScript and other client-side technologies around 2005, ad delivery (and much of web content delivery in general) became dynamic [44, 45]. Because web archiving crawlers of the time did not execute JavaScript, many advertisements never appeared to the crawler and thus were not captured; this prevented such applications from effectively preserving the representations that users experienced. We expect that traveling back in the web archives to a more recent time (0–15 years) shows a less-complete record of online advertising. In more recent developments introduced by the Internet Archive's Brozzler [46] project and the initiatives of Webrecorder [47] and Conifer [48], web archiving crawlers can use an underlying headless web browser or other browser-based technology to execute JavaScript and preserve dynamic resources that were missed by prior archiving tools [49].

Beyond the technical complications of preserving dynamic ads, the *type* of ads placed on web pages has changed over time. With the emergence of third-party cookies used for tracking users as they traversed the web, much of the advertising content targeted the user, i.e., personalized ads. So, even if the page's contents are completely archived, it is only one version of the page and not the canonical version of the page.

Today, we are on the cusp of yet another major change in the delivery of advertisements. Google [50] is preparing to remove the use of third-party cookies from its Chrome web browser, which will likely lead to the demise of third-party cookies. These web elements have been vital to the delivery of personalized advertising through tracking. At this time, it is not clear what will replace third-party cookies and how personalized advertisements will be delivered in the future, but we should be prepared to study the impact of this change.

2 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 one PhD student researcher from each university. The PI Kelly will lead and coordinate the overall effort, and he and the Drexel PhD student, who will be funded for the full 2 years, will be involved in all tasks.

This project will be carried out through the completion of four main tasks to be accomplished in the two-year time frame. Each task is mapped to one of the planned outcomes. We anticipate that some tasks will be able

to run concurrently following the initial creation of data sets in Task 1. The specific schedule of these phases and task dependencies are illustrated in the Schedule of Completion document.

Here we summarize each task, including the lead, time frame, and associated outcome. After providing an overview of all of the tasks, we will discuss details and evaluation criteria for each task in subsequent subsections.

Task 1 - Develop multiple URL collections containing advertisements

- Leads: PI Nelson and PI Weigle (ODU)
- Time frame: August 1, 2022 – May 31, 2023
- Outcome: Publicly available data sets of URLs of advertisements (outcome 1)

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 [48, 51]. Our goal is to collect at least 500 unique advertisements between the two data sets. Although this task will be led by ODU, PI Kelly and the Drexel PhD student will also be involved in this task.

Task 2 - Quantitative evaluation and analysis of archived advertisements from Task 1

- Lead: PI Kelly
- Time frame: January 1, 2023 – August 31, 2023
- Outcome: Analysis and categorization of the advertisements from Task 1 (outcome 2)

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. We anticipate that the analysis performed during this task will lead to a categorization of the types of ads that have been available through time, from static images that were placed on web pages based on the web page's content to today's personalized dynamic ads placed based the user's previously visited web pages. We have overlapped Tasks 1 and 2 in time because the analysis of archived advertisements could proceed during the collection of live Web advertisements.

Task 3 - Survey of and interviews with archivists and scholars

- Lead: PI Poole
- Time frame: June 1, 2023 – December 31, 2023
- Outcome: Qualitative assessment of the significance of what is missing from web archives and needs for the future (outcome 3)

We will perform a mixed-methods analysis based on information gathered from web archivists and humanities and information sciences stakeholders to assess how scholars across disciplines value web ads. This task will be led by Co-PI Poole based on his prior expertise in this domain. We will investigate the importance of disappearing online advertisements 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 web archivists and humanities scholars is needed to garner a more robust assessment of the significance of the problem. To collect digital ephemera relevant to scholars, we need to establish the types of material they find most useful. This task will, first, explore the obstacles and ambitions of web archivists in their effort to ensure a complete cultural record and to curate robust materials for scholarship. Second, it will assess how scholars value web ads. The research design involves a survey followed by semistructured interviews with key survey respondents. This task complements the quantitative analysis in Task 2.

Task 4 - Project Results Dissemination

- All PIs
- Time frame: December 1, 2023 – July 31, 2024
- Outcome: Dissemination of work accomplished, provides foundation for future work, formalizes research questions

In this task, we will write a white paper to serve as a baseline to inform further research, including the preparation of a longer-term IMLS or NEH grant proposal. Throughout the project, all PIs will be involved in ensuring that all outcomes and deliverables are publicly accessible and will participate in the writing and release of the findings documents from the assessments. The initial formalization of this effort will be performed during Task 2 and Task 3, while a concentrated effort will be put forth in the final group effort (Task 4) to close the work.

2.1 Task 1 - Develop multiple URL collections containing advertisements

The work proposed will include creation of two data sets for exploring the evolution of ads on the web, which could be of immediate value to archivists as well as humanities and social science researchers. Anecdotally we know that advertisements are missing, but the extent of missing advertisements has never been quantified. The data sets will be based on an existing list of over 16,000 URLs (web addresses) gathered as part of our previous research [52] that were sampled from the 500 top domains on the web in 2019, URLs available in HTTP Archive [53], and URLs gathered by the Web Archives for Historical Research Group [54]. We also have lists of known ad server hostnames [55, 56] and will query web archives for a sample of archived resources from these hosts. Our first data set will include archived web pages dating back to 1996 derived from public web archives, and will provide a baseline for advertisements present or missing in the current web archives. The second data set will be created using Conifer [48] (previously Webrecorder.io [51]), a browser-based web archiving tool created by Rhizome, and will serve as a more contemporary collection of web ads still available on the live web as of the date of capture in 2022. We will also explore the use of other browser-based crawlers, including Browsertrix and Brozzler (e.g., [57]). The second data set will inform future crawling strategies, balancing the additional resources required to obtain the missing advertisements vs. the prevalence and reuse of advertisements across pages (cf. [37]). Our goal is to collect at least 500 unique advertisements between the two data sets. We note that because the same advertisement may be represented by multiple URLs (due to the addition of tracking parameters like `utm_source`), we may end up with more than 500 URLs for the 500 advertisements.

The first data set, from public web archives, will be used in the Task 3 analysis of how well advertisements have been archived over time. This is important because the archivability of embedded resources (like advertisements) is often affected by the technology used to create the web page (for instance, if the page is built largely using JavaScript or with standard HTML) [44, 45]. We will identify the embedded resources in these web pages and filter for those that are likely linking to advertisements. There are two methods we will use to determine if an embedded resource is likely an advertisement. First, for resources that are not archived (and thus cannot be rendered), we will determine the likelihood based on the URL itself [58, 59], referencing lists of known ad server hostnames [55, 56]. For resources that are archived, we will inspect the resource and make a determination as to whether it is an advertisement or not. Because there are no existing classifiers for this task, we will perform this classification manually by inspection. Developing a machine learning based classifier could be a task for a follow-on proposal, but is out of scope for a planning grant.

The second data set will be collected from the live web and will be archived using Conifer (or other browser-based crawlers). We will use the same original data set as above, but load the pages on the live web instead of from web archives. This will surface current advertisements as opposed to advertisements placed in the web pages in the past. We will use the same manual inspection method as described above to determine if a resource is an advertisement or not.

Evaluation: The presence or absence of advertisements in web archives will be assessed via replay (rendering). The next step will be to examine the missing resources and then infer from their URLs (e.g., doubleclick.net), placement in the page, and method of inclusion (e.g., HTML iframes) the likelihood that they were ads. We will assess the different types of advertisements collected and the domains from which the advertisements were served. We will consider ourselves successful in this task if we can produce a high-quality, gold-standard data set of online advertisements covering the time period from 1996 to 2022. The successful outcome of this task will allow us to ask further questions about how well advertisements have been and are being archived and lead to research we will explore in future proposals.

2.2 Task 2 - Establish a baseline for what's archived and what's missing

In Task 2, we will map trends in the rise and fall of advertisements in archived web pages, beginning in 1996 through the present day (although the web archives are sparse until approximately 2007). We hypothesize that the early days of the web, before JavaScript was the predominant method of including web advertisements, will

have most of the ads in web pages preserved, followed by a decline in ads as JavaScript gains prominence. This task will establish a quantitative baseline for assessing what, where, and when web ads are archived.

We will assess the percentage of ads found in web archives out of the total number of references to those ads in our dataset. The data set we create in Task 1 will contain advertisements that were found in web archives as well as advertisements that were referred to by archived web pages but that were not themselves archived. Just because a URL appears in the HTML source of an archived page does not mean that the content of that referenced URL is itself archived. In addition, in Task 1 we will have collected a set of recent advertisements and archived those with Conifer. We will evaluate how well these recent advertisements have been archived in traditional public web archives, like the Internet Archive's Wayback Machine and others.

To determine technological impact, we will analyze the percentage of advertisements archived through time so that we can determine how technology changes, both in the delivery of advertisements and in web archiving, affect this percentage over time. There is an interplay between advertising technology and web archiving technology, as discussed in Section 1.4. We will investigate how changes to these technologies over time have impacted the ability to archive advertisements and other dynamic content on the web. For instance, the Internet Archive's "Save Page Now" tool started using Brozzler [46], a headless browser that executes JavaScript, in October 2019 [60], so we expect to see more ads archived after that change.

We will also explore the possibility of matching possible advertisements to pages where they are missing. Depending on what we find in Task 2, we believe it will be possible to identify a subset of advertisements shared among many different pages. Informed by advertisement reuse patterns in the live web, we will explore possible reuse patterns for the past web as well. If we determine this feasible, we will propose a complete strategy and evaluation plan in the full proposal.

Evaluation: The results of this task will be published in the open (blog posts, preprints, conference papers) and peer-reviewed literature. This task represents an analysis of the data gathered in Task 1. We will consider our analysis to be successful if the report of our findings is published in a peer-reviewed venue.

2.3 Task 3 - Qualitative survey and semistructured interviews

The third task phase centers on obtaining expert community user input and feedback. Our primary stakeholders are web archivists, special librarians, and scholars in the humanities and social sciences. A mixed methods approach will yield rich baseline data and ensure national impact. First, we will conduct a broad-gauged quantitative survey of humanities and social sciences scholars and archivists [61]. Taking a purposive sampling approach, we will disseminate the Qualtrics-based survey via prominent professional listservs such as the International Internet Preservation Consortium (IIPC), Humanities and Social Sciences Net Online (H-NET), the Organization of American Historians (OAH), the American Studies Association (ASA), the Modern Language Association (MLA), the American Political Science Association (APSA), the American Sociological Association (ASA), the Cultural Studies Association (CSA), the Society for Cinema and Media Studies (SCMS), and the Society of American Archivists (SAA). This survey will provide us not only with foundational data concerning scholarly and curatorial interest in the web archiving of advertisements, but will enable us to recruit participants for the qualitative part of the user study, which hinges on semistructured interviews [62, 63]. These interviews will offer multiple nuanced perspectives from both scholars and curators on the ways in which scholars might value web ads and leverage their research potential, leverage our two exemplary datasets, and develop the various affordances of the public user interface. Lessons learned will be integrated into the summative findings document. These interviews will also facilitate future national collaborations (e.g., in refining tools and methods, disseminating datasets, an encouraging research use) among the project staff and interested scholars, librarians, and archivists, as well as among their institutions' libraries, archives, and repositories. Additionally, our letter writers have expressed a willingness to assist with this task in identifying potential candidates to interview.

Through the use of surveys and interviews, we will better understand what information information professionals and researchers require to do their work, and how further progress can be made. Survey data followed by qualitative semistructured interviews will underpin the work on this task and will be led by PI Poole. More specifically, analysis of surveys will be grounded in descriptive statistics and inductive analysis, while analysis of interviews will be rooted in grounded theory [64, 65, 66]. Both iterative and ongoing, grounded theory coding involves valuing a researcher's

interpretive strategies and acknowledging the relationship between methods and emergent theory. Through initial and focused coding, we will identify key themes regarding information professionals' and researchers' perceptions of the project data and its analysis; these constitute key points in the findings document. Our mixed methods approach will give our findings both depth and reach, reliability and trustworthiness.

2.4 Task 4: Reporting and Assessment

The tangible products of this work will be the data sets (Task 1), the findings documents (Tasks 2 and 3), and the white paper and scholarly publications and presentations (Task 4), all of which will frame future research and feature development. Drexel and ODU will lead in the creation of data sets and quantitatively analyze these data sets. The two data sets will 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[3], where it will receive a DOI, at the end of the project.

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), the Web Archiving and Digital Libraries (WADL) workshop, IIPC General Assembly (GA) and Web Archiving Conference (WAC), and the annual Archive-It Partners Meeting. We also plan to submit publications to journals such as the International Journal on Digital Libraries (IJDL).

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.

3 Diversity Plan

The prior emphasis from both ODU and Drexel in facilitating diversity and inclusion will be installed into this project from the beginning. We will consider diversity and inclusion in hiring of our graduate research assistants and selecting the pool of scholars for our Task 3 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. In our qualitative survey (Task 3), we will strive to include a diverse pool of archivists and scholars to ensure that many views are heard and included.

Since 2017, Drexel's College of Computing and Informatics (CCI) has had an ongoing strategic effort to increase the number of women across the college through its Women in Tech initiative. In November 2020, Drexel CCI launched its Diversity, Equity, and Inclusion Council in an effort to ensure a welcoming, supportive, respectful, and inclusive environment. In selecting the student from Drexel to be involved in this project, the council, who has one stated goal of "advising on diversity planning", will be consulted to ensure that the selection takes into account the already diverse student population at CCI.

In the ODU WS-DL research group, 10 of our 19 PhD students are female and 3 are from a minority group. In the larger Department of Computer Science 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. With this level of representation in our pool of potential candidates, it is highly likely that the ODU graduate student hired on this project will be either female or in a minority group.

4 Project Results

The major goal of this Planning Grant is to analyze the need for and feasibility of archiving advertisements embedded in web pages. The first step is to assess how well online advertisements have been archived in the past. Our intuition is that ads from the early days of the Web, when many ads were delivered as embedded images, may be well-preserved, but that as ads began to be delivered more dynamically through the use of JavaScript, we will

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

discover gaps in the archival record. **Task 1** in our plan focuses on building a collection of archived advertisements, to discover what is available. This will include a collection of historical archived advertisements and a new collection that we will build of current advertisements, using the latest high-fidelity Web archiving technologies. This will result in **outcome 1**, two publicly available data sets for URLs of advertisements.

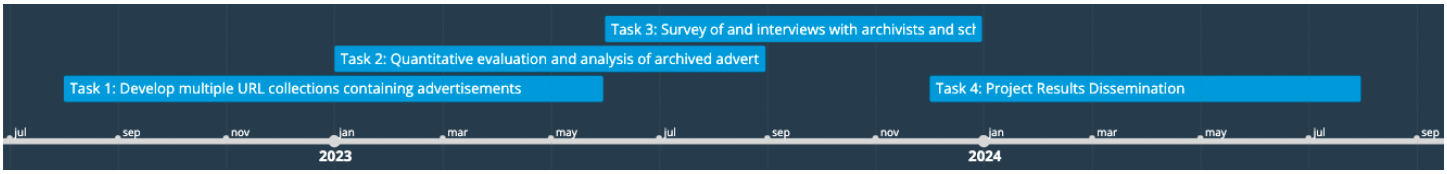
Task 2 will focus on analyzing the data gathered in Task 1, to establish a baseline of what has been archived and what is missing. This analysis task will test our stated intuition and help to guide recommendations for what needs to be done in the future to better archive online advertisements. Through this task, we will produce **outcome 2**, an analysis and categorization of the advertisements collected in Task 1. The results of this analysis will be published and could serve as the basis for future work, either by our team or by others.

Once we have an idea of the types of advertisements that are missing from the archival record, we want to assess the need from archivists and scholars to do a better job of preserving advertisements going forward. This leads to our **Task 3** where we will perform a qualitative analysis consisting of surveys and interviews with archivists and scholars. This task will produce **outcome 3**, an assessment of the significance of what is missing from web archives and the needs for the future.

Our final task, **Task 4**, has the goal of pulling together the findings from the previous three tasks and writing a white paper (**outcome 4**) that allows us to widely disseminate the findings from our preliminary studies. This white paper will also serve as the foundation for future work and include research questions that could be investigated based on our initial exploration.

Ads from the past reflect contextual aspects beyond the advertisements themselves. Because of technical limitations and this lack of awareness, online advertisement are not being preserved to the same extent as other web content. 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.

Schedule of Completion



Task 1 - Develop multiple URL collections containing advertisements
August 1, 2022 – May 31, 2023

Task 2 - Quantitative evaluation and analysis of archived advertisements from Task 1
January 1, 2023 – August 31, 2023

Task 3 - Survey of and interviews with archivists and scholars
June 1, 2023 – December 31, 2023

Task 4 - Project Results Dissemination
December 1, 2023 – July 31, 2024

Digital Products Plan

Type

- Two data sets of online advertisements and their surrounding web pages to be used in this research that can also be shared for further analysis.
- A quantitative baseline and categorization of what ads are and are not archived through time.
- A qualitative assessment of the significance of what is missing from web archives and needs for the future.
- A summary document that will provide the foundation for future work, including future research questions that could be addressed based on this initial exploration.

Availability

Data sets will be shared publicly on GitHub to encourage reuse and further exploration in this domain.

Access

Data sets will be shared using a permissive Creative Commons license. Code products will likewise will be shared using an applicable permissive license like The MIT License.

Sustainability

In addition to publishing the data on GitHub during development, we will publish it with Zenodo[1], where it will receive a DOI, at the end of the project.

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

Organizational Profile

Drexel University is one of a few U.S. institutions with a college that encompasses both the depth and breadth of computing and informatics – and the important interplay between those fields – under one roof. With a crosscutting, interdisciplinary curriculum, Drexel's College of Computing & Informatics (CCI) is uniquely poised to meet the changing needs of society and underpin the transformational roles of information and technology in today's economy. CCI is home to one of the oldest continually ALA-accredited library and information science programs in the country: the Library and Information Science major in the College's Master of Science in Information degree program. The College is a founding member of the iSchools Caucus of 29 prominent colleges dedicated to advancing the information field in the 21st Century.