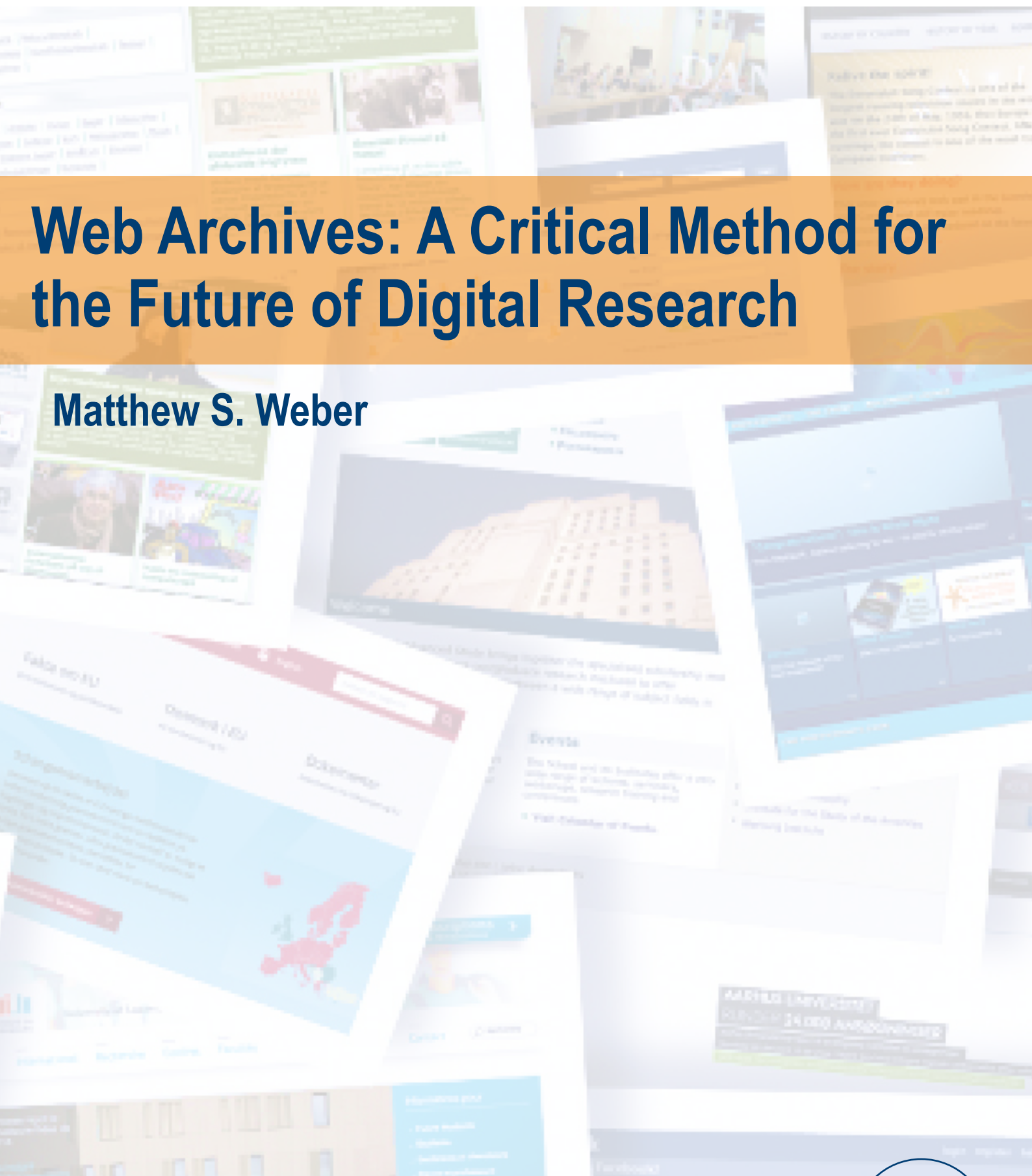
Web Archives: A Critical Method for the Future of Digital Research

Matthew S. Weber



WARCNET PAPERS

WARCnet
web archive studies

Web Archives: A Critical Method for the Future of Digital Research

Matthew S. Weber

msw@umn.edu

WARCnet Papers
Aarhus, Denmark 2020

WARCnet Papers ISSN 2597-0615.

Matthew S. Weber: Web Archives: A Critical Method for the Future of Digital Research

Published by the research network WARCnet, Aarhus, 2020.

Editors of WARCnet Papers: Niels Brügger, Jane Winters, Valérie Schafer, Kees Teszelszky, Peter Webster, Michael Kurzmeier.

Cover design: Julie Brøndum
ISBN: 978-87-94108-01-0

WARCnet
Department of Media and Journlism Studies
School of Communication and Culture
Aarhus University
Helsingforsgade 14
8200 Aarhus N
Denmark
warcnet.eu

This WARCnet Paper has gone through a process of single blind review.

The WARCnet network is funded by the Independent Research Fund Denmark | Humanities (grant no 9055-00005B).

DANMARKS FRIE FORSKNINGSFOND
INDEPENDENT RESEARCH FUND DENMARK

WARCnet Papers

Niels Brügger: *Welcome to WARCnet* (May 2020)

Ian Milligan: *You shouldn't Need to be a Web Historian to Use Web Archives* (Aug 2020)

Valérie Schafer and Ben Els: *Exploring special web archive collections related to COVID-19: The case of the BnL* (Sep 2020)

Niels Brügger, Anders Klindt Myrvoll, Sabine Schostag, and Stephen Hunt: *Exploring special web archive collections related to COVID-19: The case Netarkivet* (Oct 2020)

Friedel Geeraert and Nicola Bingham: *Exploring special web archives collections related to COVID-19: The case of the UK Web Archive* (Nov 2020)

Friedel Geeraert and Barbara Signori: *Exploring special web archives collections related to COVID-19: The case of the Swiss National Library* (Nov 2020)

Matthew S. Weber: *Web Archives: A Critical Method for the Future of Digital Research* (Nov 2020)

Valérie Schafer, Jérôme Thièvre and Boris Blanckemane: *Exploring special web archives collections related to COVID-19: The case of INA* (Aug 2020)

Niels Brügger, Valérie Schafer, Jane Winters (Eds.): *Perspectives on web archive studies: Taking stock, new ideas, next steps* (Sep 2020)

Friedel Geeraert and Márton Németh: *Exploring special web archives collections related to COVID-19: The case of the National Széchényi Library in Hungary* (Oct 2020)

Friedel Geeraert and Nicola Bingham: *Exploring special web archives collections related to COVID-19: The case of the IIPC Collaborative collection* (Nov 2020)

Caroline Nyvang and Kristjana Mjöll Jónsdóttir Hjörvar: *Exploring special web archives collections related to COVID-19: The Case of the Icelandic web archive* (Nov 2020)

All WARCnet Papers can be downloaded for free from the project website warcnet.eu.

Web Archives: A Critical Method for the Future of Digital Research

Matthew S. Weber

Abstract: Web archives have evolved considerably in the past twenty years both as a means of data collection and as a process for conducting scholarly research. This essay examines how Web archives have gained legitimacy as a method for research in academia, and focuses on core areas for future development. Specific focus is given to the building of a common knowledge base, the need to address issues of accessibility and scalability, the interdisciplinary nature of Web archive research, and the future challenges of reliability and validity when engaged with Web archive studies.

Keywords: Web archives, research methods, reliability, validity

INTRODUCTION

Web archiving is the practice of preserving Web content in a curated manner to preserve content for future access. Web archiving has been in existence for decades now, dating back to the 1990s and the establishment of the Internet Archive. In the late 1990s a variety of nonprofits and universities began to archive various aspects of digital content, but the formation of the Internet Archive in 1996 launched the first comprehensive Web archiving program in the world (Toyoda & Kitsuregawa, 2012). Early Web archiving was sporadic in nature, as collections were often built from donated datasets or were constructed in an ad-hoc nature by sampling across web domains. In the subsequent decades the degree of Web archiving has accelerated and become far more comprehensive (Milligan, 2016; Schneider et al., 2003). Web archives are more than a source of data; increasingly Web archives are a central source of data for scholars — be it in the form of archived Web pages, Twitter content, news articles, online communities, or the countless other forms of online data. Indeed, while Web archives have traditionally been thought of as archived records of website pages, the proliferation of content on the Web has meant that the nature of archived Web content has evolved. While many still access rich repositories of archived Web pages, researchers are just as likely to work with subsets of Twitter content or a record of a single

community of Web pages, such as Microsoft's archive of Usenet interactions.[1] The form of Web archive research is continuing to change and evolve, but Web archives are emerging as a central and critical method for a wide variety of scholarship across domains. Indeed, even in writing about Web archiving, I have moved beyond a discussion of purely Web-centric content; domains such as Usenet, and more modern platforms including WhatsApp and Weibo, interact with the Web but are not directly a part of it. The discussion herein focuses on the Web but incorporates a discussion of some of these key components as they relate to Web archiving (e.g. discussions of Twitter archiving, which relies increasingly on a mobile Web environment).

My own experiences with Web archives stretch back to the late 2000s. At that time, Web archiving was well established as a data preservation practice, but the notion of Web archives as a research method was still in its infancy. In summer of 2008 I was invited to attend the Oxford Internet Institute's Summer Doctoral Program. Attending this program was a privilege, and one I was fortunate to take advantage of in my career. The research program afforded me two weeks of focused time at Oxford University to work with senior mentors and other doctoral students to develop my dissertation research. My dissertation examined the evolution of the news media industry in the United States across a 10-year period, and ultimately leveraged Internet Archive data to trace how these media organizations changed the presentation of their content on the Web. I was doubly fortunate in that the 2008 program focused on the emerging research area of Web Science, and I spent my days in discussion with the likes of Tim Berners-Lee, Wendy Hall, Jonathan Zittrain and William Dutton, among others. In addition, I had the opportunity to meet and talk with Kris Negulescu, who was the Director of the Web Group at the Internet Archive at the time.

Meeting and talking with Negulescu about the data that was being captured by the Internet Archive[2] opened my eyes to the breadth and depth of data contained in Web archives. At the time, I was a doctoral student with a passion for understanding how and why digital media was evolving in the context of news production and distribution. But I also realized that in order to study that evolution I would need to be able to understand what digital media was becoming and what it had been. Tracking what digital media had started as, and what it had evolved into, meant that I needed to be able to either recreate or revisit prior generations of Web content. Archived Web content clearly provided a path to being able to analyze the patterns that I was aiming to examine. It is fair to say that in lieu of time travel archived Web content provides about the only consistent and freely available mechanism for examining prior iterations of Web technology.

One of the central challenges that has plagued Web archive research is the balance between archiving practices and research protocols. While there is a tremendous amount of data available, there have generally been significant hurdles associated with moving the data from a repository to a research ready format. Thus, my early conversations in 2008 launched my work on developing new approaches to utilizing Web archives as a research methodology for data collection and analysis.

---

[1] For details on the Usenet archive see https://archive.org/details/usenet

[2] For more information on the Internet Archive see https://archive.org

In the past decade, Web archiving has become increasingly interconnected with emerging research on digital media, and has become more important than ever as a method for data collection. The goal of this essay is to examine what it means to use Web archives as a method for digital research. In unpacking what it means for Web archives to be considered as a digital method for research, the following proceeds by first discussing what Web archives are as a research method, and what it means to be established as a methodology. Subsequently, four key challenges facing Web archiving are examined from a research perspective. These challenges include the need to continue developing a knowledge base, the importance of increased accessibility and scalability, the role of developing intersections with existing domains of research, and the need for approaches that aide in establishing validity and reliability of research conducted via Web archives. This discussion is presented, in part, in the context of my own experience examining news media through the lens of Web archives. In the closing sections of this essay, I explore what I see as future directions for research in this space.

DEFINING A RESEARCH METHOD

A common problem plaguing emerging research methods is that the method itself is evolving as the discipline continues to emerge. For example, this is the exact issue that has proven problematic in trying to define mixed methods as a research methodology; thus, it is imperative to realize that definitions will evolve over time(Johnson et al., 2007). Thus, research methods are broadly defined as methods by which, through the careful and exhaustive investigation of all the ascertainable bearing upon a definable problem, we research a solution to that problem (Connaway & Powell, 2010). In this way, Web archives have taken on a life as an exhaustive process for examining a range of research questions. Research methods are a driving force in scholarship, and the choice of research method is an important decision in the trajectory of any given research project (and career). The selection of an appropriate method is central to the process of conducting research — especially given the plethora of methods that are available today (Lather, 1986). A research method ultimately is a systemic plan for conducting research (Patton, 1990), whether that be through semi-structured interviews designed to lead to an in-depth thematic analysis or a survey designed for subsequent statistical analysis.

In the case of Web archives, I postulate in this essay that the emerging method of Web archives is a type of research method by which a researcher or team of researchers utilize quantitative or qualitative approaches to examining archived Web data for the broad purposes of developing a broader understanding of a variety of phenomena related to the development of the Web (see Winters, 2017, for an exploration of the role of Web archives as a critical tool in the social science and history scholarship). This is posed as an initial evaluation of the state of Web archive research; the following examines some of the ways in which this arena of research has continued to emerge and to garner legitimacy.

WEB ARCHIVES AS AN EMERGING METHOD

Today, Web archiving is well established as a tool for archivists seeking to collect and archive Web data (Arms et al., 2006). Increasingly, Web archiving is a complicated space for research, spanning from qualitative content analyses to the world of big data (Schafer & Musiani, 2015). Indeed, the Internet Archive, one of the best known web archiving examples, was founded in 1996, and similar efforts are well established, such as the British Library's Web archiving efforts that started in 2004 and the Danish National Library's efforts that launched in 2005. These substantial archiving efforts have each established notable trajectories as research resources. These are just a few of the many national and non-profit efforts that have launched in the past 20 years to support archiving efforts related to Web data.

Early research efforts related to Web archiving approached the topic from an archival perspective (Ben David, 2016). Indeed, the early research community that formed as part of the International Internet Preservation Consortium's (http://netpreserve.org) work was focused on studying the effectiveness and utility of archiving efforts, or reporting back on the status of ongoing archiving efforts. But as the field has grown, these communities have evolved as well.

From a research perspective, it is only in the past decade that scholars have been able to access the breadth and depth of archival Internet data for research purposes. As that has happened, relevant scholarship has opened up new domains of research for scholars. The Buddha Project out of Oxford, UK, sought to open up the UK web to researchers by demonstrating its utility through a series of case studies projects. The work was supported by funding from the UK's Arts and Humanities Research Council. In another vein, in 2020 the United States National Library of Medicine was able to expand its Global Health Events web archive to rapidly capture and archive emerging resources related to the COVID-19 pandemic. The increasing flexibility to adapt Web archiving practices to archive an emerging event is further indicative of the way in which Web archiving is developed as a methodology for researchers. In preserving COVID-19 resources, the National Library of Medicine served to create an important research artifact, but also specified a method for analysis — Web archives — as the avenue for studying the digital presentation of information related to COVID-19. More recently the WEB90 project was funded by the French National Research Agency with the goal of developing projects focused on Web data collected in France from the 1990s. Others, including the Internet Archive and the British Library, have started related collections (see, for example, the related WarcNet papers available at https://cc.au.dk/en/warcnet/warcnet-papers/ that detail out national collections including Luxembourg, France and Hungary, among others); collectively, this demonstrates the ability to operationalize this research methodology in order to enable the study of real-time events as they unfold and evolve.

In thinking about the development of web archiving as a method for digital research, there are a number of key challenges that impact the growth of this methodology within academic. Four primary challenges must be addressed in establishing Web archiving as a methodology. First, the knowledge base regarding how best to utilize Web archive in

research is still developing. Second, there is a need for increased accessibility of Web archives and improved scalability of existing tools in order to leverage the full capacity and depth of Web archives. Third, in order to continue developing Web archiving as a method there is a need to build intersections with existing domains of research in order to expand the applicability of the research method. Fourth, as the use of the method continues to expand it will be important to establish new indicators of validity and reliability.

BUILDING A KNOWLEDGE BASE

I have worked in my own research to develop a framework for incorporating Web archive research as a central methodological approach to studying local news media. Contextually, much of my research focuses on the study of change in news media organizations, and in recent years that work has shifted to emphasize the evolution of local news organizations. Web archiving is particularly apt to the study of this space in part because news organizations are increasingly reliant on the Web as a primary platform for dissemination of information. As a result, the archiving and replay of Web pages is a critical methodological tool for being applied to replicate how news media content is disseminated via the Web. Moreover, the acceleration of tool development has meant that it is increasingly easier to extract and analyze data at scale in order to extrapolate and analyze data.

Tools for analysis are part of the overall development of this domain of research, but it is equally important that researchers establish a robust tradition of scholarship in this space. Web archiving has begun to take hold as a central component of research across the academy. An analysis of Web of Science and Google Scholar citations shows steady year-over-year growth of the mention of “web archive” or “web archiving” in researching abstracts from 1996 through to 2019, indicating a growth in digital research utilizing Web archives. The domains of scholarship vary, ranging from computer science to information science to imaging research to communication and the social sciences.

The establishment of a knowledge base is more evident when one looks at the growing number of academic conferences that feature discussions related to Web archives. First, a number of Web archiving oriented conferences feature robust research tracks, including the International Internet Preservation Consortium’s Web Archiving Conference, the Personal Digital Archiving Conference, the Web Archiving and Digital Libraries Workshop at the Joint Conference on Digital Libraries, the International Conference on Digital Preservation (iPres), the Digital Preservation conference and the Digital Library Federation conference, among others. But beyond the host of Web archiving conferences that have emerged over the years, academic conferences such as the Annual Conference of the International Communication Association, the Annual Meeting of the American Sociological Association, and the Annual Meeting of the American History Association each feature divisions or tracks that routinely feature research focused on Web archiving. The development of the scholarship surrounding Web archiving will continue to enhance Web archives as a method in coming years. Further, the ongoing development of niche research events focused specifically on research related to Web archives has helped to create a space for further development of the methodology.

In my own contributions to research in this space, I have focused on analyzing news media data (Weber & Napoli, 2018). News archives are a particularly interesting space for Web archival research because the topical space is constantly evolving in order to keep pace with changing consumer technology. Moreover, the breadth and depth of news content creates challenges from the research perspective, especially in terms of capturing and analyzing relevant data (Boss & Broussard, 2017). In particular, I have focused on an exemplar case in my own research, working with a team to create a sample of local news websites for 100 communities in the United States. This research was undertaken as part of the study with the goal of understanding the health and robustness of local community news and the role of related sociodemographic in affecting the robustness of local news (see e.g. Napoli et al., 2017).

Archival news media has been shown by many to be a useful tool for understanding media and society. For instance, early research using this method focused on the transmission of news articles across different media platforms (Leskovec et al., 2007). Scholars have further utilized archival Internet data to recreate hyperlinking patterns of news media organizations online, and to assess the evolution of news media flow over time (Weber, 2012; Weber & Monge, 2014), as well as to recreate patterns of social movements and collective action (Bennett, 2005). The aggregate experience, especially my own work with local news archives, has further helped to shape my understanding of the key challenges facing Web archives as a research method.

ACCESSIBILITY AND SCALABILITY

In large part, the growth of scholarship in pertaining to Web archiving is driven by growing awareness of Web archiving technology and data. Archivists and librarians deserve significant credit for their work in promoting Web archives, and working with research to improve access to archives. Simultaneously, as others have documented, the growth in computing power and programming capacity has helped to improve accessibility and the ability to deal with data at scale. The last 10 years has seen a rapid growth in the number of Web archiving research projects, each of which has made notable contributions to the accessibility and scalability of Web archives.

For example, the Memento project (http://timetravel.mementoweb.org) has helped to make the custom creation of Web archiving more accessible to individual scholars. Memento launched a browser plugin that allows individuals to access Web archives via their browser interface, and also to build custom collections. Ultimately, the project aims to make Web archived content more discoverable by users by allowing a user to view a Web page as it appeared on any given day based on access through one's browser. The protocols users for the Memento program access publicly available Web archives, and the interface is available to the end user via a Google Chrome plugin.

Memento focuses on making Web archived content easily searchable at the user level; at the other end of the spectrum the Archives Unleashed project aims to make Web archiving more accessible to researchers. In the past few years, the Archives Unleashed project team has helped to create a suite of server-side tools, as well as a graphic user

interface, in order to improve researchers' ability to access Web archive data. The research tools (including a toolset based on Apache Spark, and a set of accompanying Jupyter notebooks) include substantial documentation and use cases; overall the aggregate platform is designed to increase research accessibility and to improve the integration of Web archive data into established research and graduate student training. The Archives Unleashed Cloud project allows researchers to connect to datasets hosted on archive-it.org and to both visualize and analyze their datasets, as well as to create derivate datasets based on research needs. In addition, there are numerous other projects that have emerged over the years, further contributing to the legitimacy and viability of these models for improving access to Web archives.

Each of these projects underscores the growing recognition that new tools are needed in order to improve both archiving practice, and the access to archive data. In the work my team led on local news, we started our data collection by limiting data to communities with a population between 20,000 and 300,000. Using US Census data, this helped to generate a list of 493 communities. We did not intend to collect each and every community, nor each and every website, within those 493 communities. Nevertheless, we generated a list of nearly 800 news outlets that we intended to study. Collecting a constructed week sample during 2016, we generated a dataset that contained 1.6 million documents (html files, pdfs, images, audio files, etc.) and 2.2 terabytes of total data based on the seed set of local outlets (split across print, television, radio and online-only news outlets). A process of manual evaluation of the front pages of the archived local news sources found that the archive contains just over 20,000 distinct news stories.

The importance of improved accessibility and scalability cannot be understated. In my own research, the ability to generate derivate datasets through the Internet Archive's Archive-It program, and to do initial analysis using the Archives Unleashed Cloud platform, contributed immeasurably to my ability to move forward in the research initiatives that I was working on at the time. Access to sufficient computing resources is part of the challenge for scholars, but a number of the newer tools allow scholars to work with subsets of data, or to do processing on cloud based platforms. In addition, many of the graphic user interfaces that are available allow for qualitative research without needed access to substantial computing power. As web archiving expands as a method, the requirements for doing research in this space are continuing to evolve. Many of the recent iterations of research tools and platforms for access have helped to lower the barrier to entry.

INTERSECTIONS WITH DOMAINS OF RESEARCH

As noted in the discussion of the emerging knowledge base pertaining to Web archives, the research space of Web archiving has grown significantly over the past 20 years and has continued to develop as an inherently multi-disciplinary space. While inherently multi-disciplinary, the research space of Web archiving continues to struggle with the challenges of defining both interdisciplinarity boundaries and connections.

Large-scale data are increasingly being used by researchers to explore ecosystems of interaction, examining the way a large body of entities interacts with one another, while

still being able to focus in on individual actors. For instance, large-scale data are utilized for a wide array of social science investigations, including research examining social movements (Driscoll et al., 2013), healthcare and health messaging (Chawla & Davis, 2013; Emery et al., 2014), collective action (Agarwal, 2014), news media (Leetaru, 2011; Weber, 2012) and even as a tool for qualitative research (Bisel et al., 2014).

Archived Web data has taken many forms; notably, in recent years, archived Twitter data has proved popular for examining a host of phenomena. Twitter data as a form of archived Web content have been utilized to explore a number of mimetic processes focuses on the diffusion and replication of information (Jungherr, 2014; Park et al., 2014). Much of the emergent research of Twitter platforms has focused on political issues, including agenda setting and information spread (Vargo et al., 2014). Twitter research in recent years has exploded, and covers a range of topics, including notable work covering terrorist attack. Smyrnaios and Ratinaud (2017) archived more than 2 million Tweets in order to analyze the types of groups that emerged in the wake of the Charlie Hebdo attacks in Paris, France in 2015. And Schafer and colleagues (2019) looked at the same period of terror attacks in France in order to examine the real-time archiving practices of the National Library of France (Bibliothèque nationale de France, BnF) and the National Audio-visual Institute. In other instances, there is a robust body of research examining the use of Twitter by consumer to find health information (Park et al., 2016), and as a platform for the dissemination of health information (Chung, 2016). Much of this health related work has emerged out of public health and health communication scholarship. This is much further afield from the origins of Web archiving studies, but the underlying point is that archived Web content is increasingly utilized in a wide array of scholarly domains.

Other forms of intersectional work using Web archives include the study of images portrayed on the Web (Ben-David et al., 2018), the use of Web archives to study history (Gorsky, 2015), and Web archives as fertile ground for computer science scholarship (AlSum & Nelson, 2013), among others. These topics all further bolster the growing important of Web archives to researchers, and point to increasing connections between disciplines.

VALIDITY AND RELIABILITY

As research continues to incorporate Web archives, and as the method of research continues to evolve, it is increasingly critical that the Web archiving community establishes benchmarks for the validity and reliability of data and data analysis relevant to Web archives. Some of the questions of validity and reliability are inherently interwoven into the act of archiving. For example, studies of subsets of Web archives in the United States are plagued by the lack of a national preservation scheme. Comparatively, countries such as Portugal and Demark have the advantage of a national preservation mandate that has enabled the creation of national Web archives (see arquivo.pt and netarchive.dk for the respective Web portals of these collections). Despite the differences in crawl scoping and crawl completeness, almost every Web crawl faces some degree of uncertainty with

regards to the completeness of the crawl, and in turn, the validity and reliability of the data extracted from a given Web archive.

Looking at extant research on Web archive completeness, one relevant study found that the average life of a website is three years (Agata et al., 2014); in turn, this means that websites preserved today are often fragmented representations of the website as it existed in the past. Given that there are billions of web pages available online, this also suggests that while a website might be archived, the outlinks from that website will often be missing in an archive. In turn, it is clear then that Web archive data has a high probability of being incomplete; crawlers generally observe robots.txt records, they struggle to handle dynamic web content, often cannot capture social media data and are simply unable to fully capture the scope of the entire World Wide Web (Ainsworth et al., 2011; McKay, 2004). These issues of data completeness and degradation plague social media data as well. For example, a study looking at three years of Twitter data found that after two years only 41% of the original content had been archived future use, and 27% of the data had already been lost to future users (SalahEldeen & Nelson, 2012).

In order to address some of these limitations of working with Web archive research, scholars have proposed focusing archival efforts on capturing data that changes most frequently (Spaniol et al., 2009). The idea behind this approach is to capture the data that are evolving most rapidly. Scholars also note that crawling strategies should prioritize archival efforts based on the size and relative position of websites within their larger ecosystems (Song et al., 2004). More importantly, it is clear that there is a need for transparency in the provenance of a crawl such that researchers are able to clearly understand the choices that were made in creating a crawl. Limitations with regards to Web archiving data are inevitable, but making those limitations transparent is critical in order for researchers to be able to control for the limitations that are present.

Testing the limitations that exist in Web archives, I took the approach in prior research of looking for the traces of Web pages to see what was present in an archive dataset versus what should be present based on outlinks from Web pages. This work focused on analyzing records for the 109th through the 112th Congresses that were collected by the Internet Archive across both the senate.gov and house.gov domains (Weber & Nguyen, 2015). For example, the senate.gov domain records contain 26,965,770 captures representing 8,674,397 unique URLs. An analysis of outlinks from Web pages in the senate.gov domain across the four Congressional sessions shows that on average 25% of outlinks were missing from the dataset.

Further, this research used curve fitting to try and approximate how the degree of missing data changed over time. The curve fit alone did not prove particularly interesting from a research perspective, as it is simply a representation of the data across time. The function, however, provides an extrapolated estimation of the degree of difference over time in a given sample. From a research perspective, the idea is that by being able to describe the "amount of error" with a function researchers can better control for that error. Again, work is needed to advance research in this space, but increased attention to the completeness of and degree of error in Web archives can help to enhance the validity and reliability of Web archive research. As the method of Web archiving evolves over time it is

clear that more scholarship in this space, in partnership with archivists and librarians, will be needed to advance the field.

RESEARCH DESIGN AND THEORY: FUTURE DIRECTIONS

While Web archiving is well established as a data preservation practice, the preceding highlights many of the ways in which Web archiving is still emerging as a mature and legitimate research method. Researchers working in this domain tend to one of two camps; the first conduct their own archiving and use those results in their research, and the second rely on collections built by others. There are pros and cons to both, and a further discussion of the various approaches in detailed by Weber and Napoli (2018). Those who create their own archives tend to lean towards work in emergent fields such as computational social science. But more broadly, a wide group of traditionally-trained humanities and social science researchers may encounter challenges when approaching web archives. In part, many of the doctoral programs that these students are trained by are not yet embracing computational approaches to research. The big data barrier is significant, and there are associated challenges related to the access and storage of these data. I have focused on research challenges, but as a field we also have a commitment to continue to advocate for advances in educational opportunities. As I have noted, programs such as Archives Unleashed provide a starting point, but an ongoing effort is needed to sustain this effort.

Beyond the issue of education, the discussion in this essay focused on the growing knowledge base that has helped to establish the legitimacy of Web archive research, the need for ongoing efforts to continue improving accessibility of Web archives and scalability of established research methods, continued develop of Web archiving as a method through interdisciplinary connections, and the need for improved indicators of validity and reliability of Web archives.

When pursuing studies of digital phenomena on the Web, the method of Web archiving provides a number of advantages compared to other approaches. Web archiving is an unobtrusive approach to data collection and analysis, it is a relatively stable approach, when collected using open-source means it is an accessible approach, and the degree of coverage is relatively comprehensive. As Web archiving continues to expand, the strong foundations for research in this space will serve as a springboard that others can build upon. That said, it is clear that there are some limitations to Web archiving as a research method. Web archiving is limited by technical barriers to entries (see Ian Milligan's work in this series published here), aforementioned questions about completeness, and uncertainties about the cost-effectiveness of data collection via this approach to research. Web archiving is particularly limited with regards to research examining the foundational years of the Web during the 1990s; the archival technology at the time was relatively limited, and the Web was growing at a scale that made it challenging to accurately capture the complete Web. These types of limitations are inherent to the artifact being studied; as archiving technology and Web archives as a method continue to evolve researchers will be better situated to address these challenges.

In order to move beyond Web archiving as an approach to data collection it is increasingly important that broader consideration be given to the research design and the way in which the act of collection (or access to collections) is integrated into the broader research design. In part, addressing the previously outlined issues will help to address the complicated challenges of research design. In part, the integration of Web archiving data with robust theoretical studies will help to better integrate Web archives into research practice. Web archiving continues to be a young methodology and a young practice, and it is clear that there is substantial room for this arena of research to grow in the future.

REFERENCES

Agarwal, S. D., Bennett, W. L., Johnson, C. N., & Walker, S. . (2014). A model of crowd enabled organization: Theory and methods for understanding the role of twitter in the occupy protests. *International Journal of Communication, 8*(27), 646-672.

Agata, T., Miyata, Y., Ishita, E., Ikeuchi, A., & Ueda, S. (2014, 8-12 Sept. 2014). Life span of web pages: A survey of 10 million pages collected in 2001. Digital Libraries (JCDL), 2014 IEEE/ACM Joint Conference on,

Ainsworth, S. G., AlSum, A., SalahEldeen, H., Weigle, M. C., & Nelson, M. L. (2011). *How much of the web is archived?* Proceedings of the 11th annual international ACM/IEEE joint conference on Digital libraries, Ottawa, Ontario, Canada.

AlSum, A., & Nelson, M. L. (2013). ArcLink: optimization techniques to build and retrieve the temporal web graph. Proceedings of the 13th ACM/IEEE-CS joint conference on Digital libraries,

Ben David, A. (2016). What does the Web remember of its deleted past? An archival reconstruction of the former Yugoslav top-level domain. *New Media & Society, 18*(7), 1103-1119. https://doi.org/10.1177/1461444816643790

Arms, W., Aya, S., Dmitriev, P., Kot, B., Mitchell, R., & Walle, L. (2006). A Research Library Based on the Historical Collections of the Internet Archive. *D-Lib Magazine, 12*(2).

Ben-David, A., Amram, A., & Bekkerman, R. (2018). The colors of the national Web: visual data analysis of the historical Yugoslav Web domain. *International Journal on Digital Libraries, 19*(1), 95-106.

Bennett, W. L. (2005). Social movements beyond borders: understanding two eras of transnational activism. *Transnational protest and global activism*, 203-226.

Bisel, R. S., Barge, J. K., Dougherty, D. S., Lucas, K., & Tracy, S. J. (2014). A Round-Table Discussion of "Big" Data in Qualitative Organizational Communication Research. *Management Communication Quarterly, 28*(4), 625-649. https://doi.org/10.1177/0893318914549952

Boss, K., & Broussard, M. (2017, 2017/06/01). Challenges of archiving and preserving born-digital news applications. *IFLA Journal, 43*(2), 150-157. https://doi.org/10.1177/0340035216686355

Chawla, N. V., & Davis, D. A. (2013, Sep). Bringing big data to personalized healthcare: a patient-centered framework. *Journal of General Internal Medicine, 28 Suppl 3*, S660-665. https://doi.org/10.1007/s11606-013-2455-8

Chung, J. E. (2016). A smoking cessation campaign on Twitter: understanding the use of Twitter and identifying major players in a health campaign. *Journal of Health Communication, 21*(5), 517-526.

Connaway, L. S., & Powell, R. R. (2010). *Basic research methods for librarians*. ABC-CLIO.

Driscoll, K., Ananny, M., Guth, K., Kazemzadeh, A., Leavitt, A., & Thorson, K. (2013). Big bird, binders, and bayonets: Humor and live-tweeting during the 2012 US presidential debates. *Selected Papers of Internet Research, 3*.

Emery, S. L., Szczypka, G., Abril, E. P., Kim, Y., & Vera, L. (2014, Apr). Are you Scared Yet?: Evaluating Fear Appeal Messages in Tweets about the Tips Campaign. *Journal of Communication, 64*, 278-295. https://doi.org/10.1111/jcom.12083

Gorsky, M. (2015). Sources and resources into the dark domain: the UK web archive as a source for the contemporary history of public health. *Social history of medicine, 28*(3), 596-616.

Johnson, R. B., Onwuegbuzie, A. J., & Turner, L. A. (2007). Toward a definition of mixed methods research. *Journal of mixed methods research, 1*(2), 112-133.

Jungherr, A. (2014). The Logic of Political Coverage on Twitter: Temporal Dynamics and Content. *Journal of Communication, 64*(2), 239-259. https://doi.org/10.1111/jcom.12087

Lather, P. (1986). Issues of validity in openly ideological research: Between a rock and a soft place. *Interchange, 17*(4), 63-84.

Leetaru, K. (2011). Culturomics 2.0: Forecasting large-scale human behavior using global news media tone in time and space. *First Monday, 16*(9). https://doi.org/10.5210/fm.v16i9.3663.

Leskovec, J., Kleinberg, J., & Faloutsos, C. (2007). Graph evolution: Densification and shrinking diameters. *ACM Transactions on Knowledge Discovery from Data, 1*(1), 1-42. https://doi.org/10.1145/1217299.1217301

McKay, C. (2004). Ephemeral to enduring: the Internet Archive and its role in preserving digital media. *Information Technology and Libraries, 23*(1), 3.

Milligan, I. (2016, 2016/03/01). Lost in the Infinite Archive: The Promise and Pitfalls of Web Archives. *International Journal of Humanities and Arts Computing, 10*(1), 78-94. https://doi.org/10.3366/ijhac.2016.0161

Napoli, P. M., Stonbely, S., McCollough, K., & Renninger, B. (2017, 2017/04/21). Local Journalism and the Information Needs of Local Communities. *Journalism Practice, 11*(4), 373-395. https://doi.org/10.1080/17512786.2016.1146625

Park, H., Reber, B. H., & Chon, M.-G. (2016). Tweeting as health communication: health organizations' use of Twitter for health promotion and public engagement. *Journal of Health Communication, 21*(2), 188-198.

Park, J., Baek, Y. M., & Cha, M. (2014). Cross-Cultural Comparison of Nonverbal Cues in Emoticons on Twitter: Evidence from Big Data Analysis. *Journal of Communication, 64*(2), 333-354. https://doi.org/10.1111/jcom.12086

Patton, M. Q. (1990). *Qualitative evaluation and research methods*. SAGE Publications, inc.

SalahEldeen, H. M., & Nelson, M. L. (2012). Losing my revolution: how many resources shared on social media have been lost? In *Theory and Practice of Digital Libraries* (pp. 125-137). Springer.

Schafer, V., Truc, G., Badouard, R., Castex, L., & Musiani, F. (2019). Paris and Nice terrorist attacks: Exploring Twitter and web archives. *Media, War & Conflict*, *12*(2), 153-170.

Schafer, V., & Musiani, F. (2015). The Historian of the Web : Crawler, Browser or Lurker? *Web Archives for Historians*. Retrieved from https://webarchivehistorians.org/2015/03/13/the-historian-of-the-web-crawler-browser-or-lurker/

Schneider, S. M., Foot, K., Kimpton, M., & Jones, G. (2003). Building thematic web collections: challenges and experiences from the September 11 Web Archive and the Election 2002 Web Archive. 3rd ECDL Workshop on Web Archives, Trondheim, Norway,

Song, R., Liu, H., Wen, J.-R., & Ma, W.-Y. (2004). Learning block importance models for web pages. Proceedings of the 13th international conference on World Wide Web,

Spaniol, M., Denev, D., Mazeika, A., Weikum, G., & Senellart, P. (2009). *Data quality in web archiving* Proceedings of the 3rd workshop on Information credibility on the web, Madrid, Spain.

Toyoda, M., & Kitsuregawa, M. (2012). The history of web archiving. *Proceedings of the IEEE, 100*(Special Centennial Issue), 1441-1443.

Vargo, C. J., Guo, L., McCombs, M., & Shaw, D. L. (2014). Network Issue Agendas on Twitter During the 2012 U.S. Presidential Election. *Journal of Communication, 64*(2), 296-316. https://doi.org/10.1111/jcom.12089

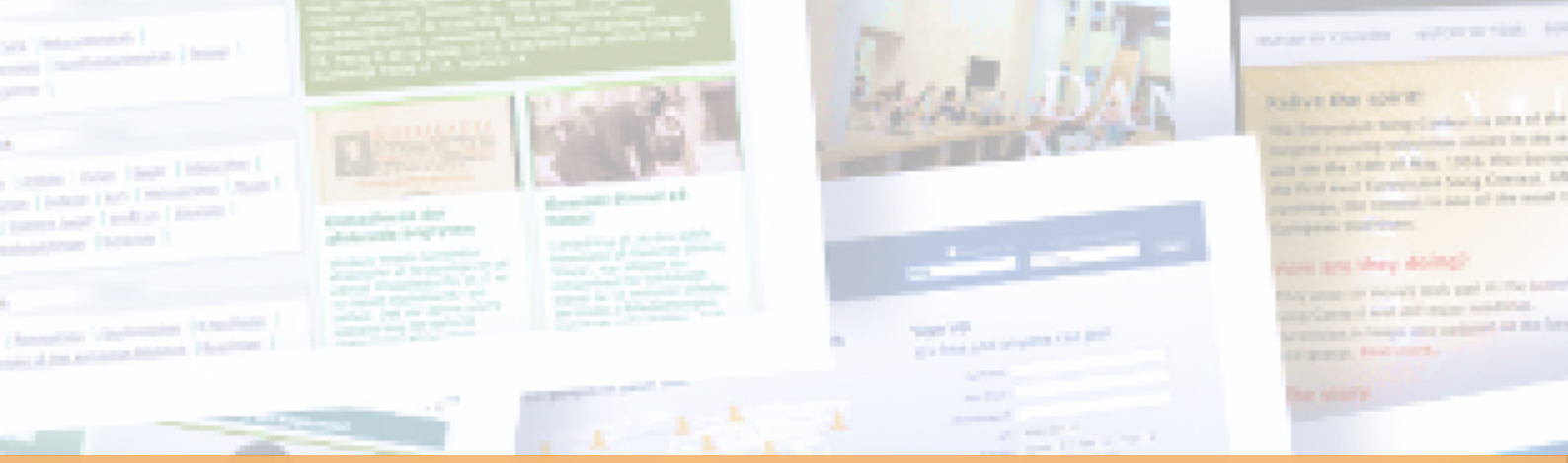
Weber, M. S. (2012). Newspapers and the Long-Term Implications of Hyperlinking. *Journal of Computer-Mediated Communication, 17*(2), 187-201. https://doi.org/10.1111/j.1083-6101.2011.01563.x

Weber, M. S., & Monge, P. (2014). Industries in turmoil: Driving transformation during periods of disruption. *Communication Research*, 1-30. https://doi.org/10.1177/0093650213514601

Weber, M. S., & Napoli, P. M. (2018). Journalism history, Web archives, and new methods for understanding the evolution of digital journalism. *Digital Journalism, 6*(9), 1186-1205.

Weber, M. S., & Nguyen, H. (2015). Big Data? Big Issues: Degradation in Longitudinal Data and Implications for Social Sciences. WebSci 2015, Oxford, UK.

Winters, J. (2017). Coda: Web archives for humanities research — some reflections. In N. Brügger & R. Schroeder (Eds.), *The Web as History: Using Web Archives to Understand the Past and the Present* (pp. 238-249). UCL Press.



WARCnet Papers is a series of papers related to the activities of the WARCnet network. WARCnet Papers publishes keynotes, interviews, round table discussions, presentations, extended minutes, reports, white papers, status reports, and similar. To ensure the relevance of the publications, WARCnet Papers strives to publish with a rapid turnover. The WARCnet Papers series is edited by Niels Brügger, Jane Winters, Valérie Schafer, Kees Teszelszky, Peter Webster and Michael Kurzmeier. In cases where a WARCnet Paper has gone through a process of single blind review, this is mentioned in the individual publication.

The aim of the WARCnet network is to promote high-quality national and transnational research that will help us to understand the history of (trans)national web domains and of transnational events on the web, drawing on the increasingly important digital cultural heritage held in national web archives. The network activities run in 2020-22, hosted by the School of Communication and Culture at Aarhus University, and are funded by the Independent Research Fund Denmark | Humanities (grant no 9055-00005B).

WARCNET PAPERS

INDEPENDENT RESEARCH FUND DENMARK

warcnet.eu warcnet@cc.au.dk twitter: @WARC_net facebook: WARCnet
youtube: WARCnet Web Archive Studies slideshare: WARCnetWebArchiveStu