# Research Statement

Mat Kelly
Department of Computer Science
Old Dominion University, Norfolk, VA 23529
mkelly@cs.odu.edu

My PhD research has focused on studying the dynamics of preserving the data and experience of the World Wide Web that may otherwise be lost in time. Because the Web acts as a contemporary medium of cultural expression and a historical record, accurate preservation of it is critical. While efforts by large-scale archiving institutions such as the Internet Archive ensure the preservation of popular content on the Web, more obscure content and Web pages that are inaccessible to these institutions (e.g., restricted access social media pages) are often not preserved.

My research has had an evolving yet consistent theme: the exploration of approaches toward amalgamating distributed and isolated efforts to generate a more useful whole. Previous projects that led me to the research topic include my studies of isolated wireless sensor networks for an improved, aggregated picture of worldly conditions (e.g., fire, smoke, movement) for emergency responders [4]. My more recent studies have investigated integrating peer-to-peer distributed systems [3] into Web archives for resilience and persistence – necessary archiving traits. These studies have integrated concerns for privacy in the aggregation procedure to produce a more comprehensive picture of the *Web that was* without the cost of loss of personal and private archived information.

In my early work in Web archiving [1], I created and publicly released open-source[1] archiving tools for individuals to enable them to preserve the parts of the Web they care about. These tools leveraged platforms with which potentially non-technical users would already be familiar, like their Web browser. The interfaces to these tools were intentionally simple and intuitively designed to obscure technical details while facilitating good practice in the technical backend. A goal for the tools was to make the process of Web archiving accessible to Web users instead of being solely delegated to institutions. This enabled an ethos of "Archive What I See Now"[2] by Web users so as to mitigate the reliance on other parties to capture the cultural history as conveyed on the Web. By encouraging individuals to preserve their own personal and private parts of the Web, I surfaced the research question of how one's captures would be temporally aggregated with others' captures, inclusive of those from archiving institutions. In aggregating private, personal, and public Web archives, privacy and access would need consideration to ensure that which is private or sensitive remains within the control of the individual.

Through my dissertation research I have attempted to appeal to well-established and standardized protocols and systems like the Web archiving (WARC, ISO 28500) format for personal archive storage and the Memento Framework (RFC 7089) for interacting with Web archives in time, and attempted to do so without the user needing to become familiar with these more technical systems. I feel that through facilitation and the increased accessibility of Web archiving, a more representative historical record will be preserved.

## Publishing

My current research has been in resolving outstanding issues of systematic access and aggregation of personal, private, and public Web archives [2]. My publications[3] (17 peer-reviewed conference/workshop papers and 2 journal articles) have been cited 120 times[4], resulting in an h-index[5] of 6. I have collaborated with 8 different institutions[6] as co-author or co-presenter. My research has been supported by the National Science Foundation (NSF), National Endowment for the Humanities (NEH), and the Institute of Museum and Library Services (IMLS), among others.

[1] Source code for my various open-source project available on GitHub, `https://github.com/machawk1`

[2] For which my research for exploration was subsequently funded by the NEH.

[3] Comprehensive list at `https://matkelly.com/pubs`

[4] Calculated by Google Scholar, `https://scholar.google.com`

[5] `https://en.wikipedia.org/wiki/H-index`

[6] List of co-authors and institutions at `https://www.cs.odu.edu/~mkelly/postdoc.html#coauthors`

I wrote 25 blog posts for my PhD research group to disseminate my research and facilitate discussion among the academic community and practitioners. I have been a peer reviewer for the ACM/IEEE Joint Conference on Digital Libraries (JCDL) and Access conferences as well as a peer reviewer for the International Journal on Digital Libraries (IJDL). I have also continually engaged in the Web archiving community by serving on the conference organizing committee for JCDL. My contributions are formally acknowledged by-name in the IETF (Internet Engineering Task Force) standard RFC (Request for Comments) 7089 for using HTTP Framework for Time-Based Accessed to Resource States (Memento)[7].

### Future Research

I anticipate research in Web archiving to be further applicable in other areas of research. Additionally, I feel that the Web archiving community can continue to learn and apply research and advancements from these fields. My initial explorations into applying concepts of peer-to-peer, distributed systems [3] has proven applicable to personal Web archiving while opening the research questions of retaining privacy at the benefit of data persistence. I have preliminarily investigated [5] the integration of leveraging conventional live Web encryption concepts and practice (e.g., OAuth 2) onto this use case of integrating p2p distributed systems and Web archiving. Encryption, security, and dissemination of these captures will help to ensure the preservation of private, personal, and controversial information.

My research has also facilitated the potential for exploration of archival negotiation in dimensions beyond time. Because this process is temporally, spatially, and computationally expensive, it is ripe for further applicability of machine learning techniques. One such example lies in identifying the facets of Web pages (e.g., dynamic content, personalized information, topical aboutness) on the live Web versus what has been captured. These facets could then be compared to previous captures and the sufficiency of previous representation evaluated. This may inform subsequent preservation policies and technical focuses to emphasize additional preservation to maximize the qualitative representation of the Web.

My research in using the live and archived Web as corpora has informed my process of analysis. However, I hope to explore the applicability of additional fields of study like machine learning and AI as applicable to Web archiving. As with my previous transition from sensor networks to Web archiving, I also hope to apply my research and understanding in the field on Web archiving to have a greater societal impact.

### References

[1] Mat Kelly and Michele C. Weigle, “WARCreate - Create Wayback-Consumable WARC Files from Any Webpage,” In Proceedings of the ACM/IEEE Joint Conference on Digital Libraries (JCDL), Washington, DC, June 2012, pp. 437-438.

[2] Mat Kelly, Michael L. Nelson, and Michele C. Weigle, “A Framework for Aggregating Private and Public Web Archives,” In Proceedings of the ACM/IEEE Joint Conference on Digital Libraries (JCDL), Fort Worth, Texas, June 2018, pp. 273-282.

[3] Mat Kelly, Sawood Alam, Michael L. Nelson, and Michele C. Weigle, “InterPlanetary Wayback: Peer-To-Peer Permanence of Web Archives,” In Proceedings of the International Conference on Theory and Practice of Digital Libraries (TPDL). Hannover, Germany, September 2016, pp. 411-416.

[4] Michael Ruffing, Yangyang He, Mat Kelly, Jason O. Hallstrom, Stephan Olariu, and Michele C. Weigle, “A Retasking Framework For Wireless Sensor Networks,” In Proceedings of the IEEE Military Communications Conference (MILCOM). Baltimore, Maryland, October 2014, pp. 1066-1071.

[5] Mat Kelly and David Dias, “A Collaborative, Secure, and Private InterPlanetary Wayback Archiving System using IPFS,” Presented at the International Internet Preservation Consortium (IIPC) Web Archiving Conference (WAC) 2017, London, England, 15 June 2017.

[7] `https://tools.ietf.org/html/rfc7089`