Exploring Software, Tools and Methods used in Web Archive Research

| **Schmid, Katharina** | **Healy, Sharon** | **Byrne, Helena** |
| --- | --- | --- |
| *Bavarian State Library*  *Germany*  *0000-0001-6057-6640* | *Maynooth University*  *Ireland*  *0000-0003-3493-0938* | *UK Web Archive*  *United Kingdom*  *0000-0002-0966-4685* |

|  |  |  |  |
| --- | --- | --- | --- |

**Abstract – This paper is one part of a larger research project, titled, Web Archives - Researcher Skills and Tools (WARST). In this poster we focus on the data from the WARST study which examines the software, tools and methods used in the web archive research lifecycle.**

**Keywords –** web archive research, web archiving, web archive creators, web archive users

**Conference Topics** – Innovation; Community

1. Introduction

In this poster we explore the landscape of software, tools and methods used in web archive research. We consider web archive research to be inclusive of web archiving, curation, and the use of web archives and archived web content for research or other purposes [1]. We maintain that web archive research is representative of the processes and activities described in the Archive-It’s web archiving lifecycle model from appraisal and acquisition, to replay, access, and use [2]. We suggest that there will always be a need to keep examining the roles of skills, tools, and methods associated with the web archiving lifecycle as long as internet, web and software technologies keep advancing, upgrading, and changing.

* 1. *Background*

This poster is one part of a larger research project, titled, Web Archives - Researcher Skills and Tools (WARST). The WARST project focuses on individuals around the globe who participate in web archive research, and explores the skills, tools and knowledge ecologies in the web archive research lifecycle. Please see Healy et al. for a full documentary of their methodology [1]. This poster focuses on the data from the WARST study which examines the tools, software and methods used in web archive research. We use Gephi, to show a network analysis of the software, tools and methods in line with two communities of practice (i) libraries, archives, and web archive environments (n=30) and (ii) academic, scholars, students, or professionals working in IT/web environments (n=14). Through a network analysis we provide some understanding of the environment and its connections.

* 1. *Related Literature*

Several other studies have done substantive work in this area focusing on web archiving initiatives and practises, users of web archives, awareness and engagement with web archives, scholarly use of web archives, or examining both web archiving practises and the challenges and opportunities for using web archives for research [3] [4] [5] [6] [7]. We build on these studies to foster discussion about the current state of collaboration and communications between web archiving initiatives and users/researchers.

1. Findings & Discussion
   1. *Data Collection*

Overall there is significant overlap between the types of tools and methods the two groups of respondents use to collect data. In the library, archive and web archive environment, however, crawling software which produces data in the standard WARC format clearly dominates. The tools and methods used by participants from a scholarly or academic environment seem to be more diverse as they are influenced by the specific research question and methodology, for example when data is collected manually for close reading.

Changes in web technologies also clearly influence tools and methods for data collection. Social media platforms, for example, are difficult to archive with traditional crawling software and generally require platform-specific software. This is reflected in the use of tools like Instaloader and Twarc to collect data directly from an API. Additionally, both groups use browser-based crawling software alongside traditional crawlers like Heritrix to capture dynamic websites that rely heavily on technologies like JavaScript.

* 1. *Data Analysis*

In the responses of both groups, we see a broad range of methods of analysis. They include manual and computer-assisted forms of analysis as well as qualitative and quantitative analyses, sometimes used in combination. From the software mentioned we can infer that input includes text and network data as well as metadata from the crawls. Tools for processing visual, audio or audio-visual data are not mentioned explicitly.

In the responses from the library, archive or web archive environment, there is a clear focus on tools and methods for search and information retrieval. They include tools for metadata search like CDX queries but also full-text search like Apache Lucene or Apache Solr. This reflects ongoing efforts to improve search capabilities and turn “web archives [...] from mere document repositories into accessible archives” [8].

Respondents from the library, archive and web archive community also refer to tools for digital forensics and digital preservation, which are not mentioned by respondents from the academic and research community. The same is true for software used specifically to process large amounts of data. This may point to fields of expertise in the library and archive community, from which other communities could benefit.

Notably, respondents from an academic or scholarly environment did not report using any of the user interfaces offered by web archiving institutions. These include tools for replaying archived web content as well as user interfaces that offer limited analytical functionalities like the SolrWayback. Further research is required to find out whether this observation holds true beyond the scope of this survey and to determine the underlying causes.

Instead, respondents seem to prefer stand-alone tools that are not specific to web archive content. As the responses show, tools like Voyant Tools, IramuteQ and Gephi that have been developed and are widely used in the digital humanities and social sciences are also in part taken up by the library and archive community. This indicates an ongoing fruitful exchange between the two communities.

Spreadsheets are another type of standard software that is used by both communities of practice. Respondents report using it to collect and manage data as well as to conduct analyses. Both communities could therefore benefit from collaborations in developing training materials for spreadsheet softwares.

1. Conclusion

In this poster presentation we specifically focus on data from the WARST project which examines skills tools and methods used in web archive research. We surmise that the landscape is heavily influenced by changes in web and software technologies and therefore merits continuous reappraisal through studies like this.

The WARST project highlights shared practices and commonalities between different communities that are involved in web archive research. By visualizing the data through a network analysis, we can examine the environment to see exactly where the commonalities are in terms of software, tools and methods. This could serve as a starting point to foster discussion for the development of training in skills, tools and methods for web archive research. For example, the findings hint at further opportunities for collaboration and knowledge exchange with regard to user interfaces for web archive collections, and training in the use of spreadsheet softwares for both collection and analysis.

REFERENCES

1. Healy *et al.*, *WARCnet Special Report: Skills Tools and Knowledge Ecologies in Web Archive Research*. WARCnet Aarhus, Denmark, Apr. 2022. (forthcoming).
2. M. Bragg and K. Hanna, ‘The Web Archiving Lifecycle Model’, The Archive-It Team, Internet Archive, USA, Apr. 2013.
3. J. Bailey et al., ‘Web Archiving in the United States: A 2013 Survey’, NDSA Report, National Digital Stewardship Alliance (NDSA), USA, , Sep. 2014. Available:<https://osf.io/h4e6z/>..
4. M. Costa and M. J. Silva, ‘Understanding the information needs of web archive users’, in *Proceedings of the 10th International Web Archiving Workshop (IWAW 2010),* Vienna, Austria, September 22-23, 2010, pp. 9–16. Available:<https://web.archive.org/web/20110723173820/http://www.iwaw.net/10/IWAW2010.pdf>.
5. M.-D. Costea, ‘Report on the Scholarly Use of Web Archives’, NetLab, Aarhus, Denmark, 2018. Available:<http://netlab.dk/wp-content/uploads/2018/02/Costea_Report_on_the_Scholarly_Use_of_Web_Archives.pdf>.
6. H. Hockx-Yu, ‘Access and Scholarly Use of Web Archives’, *Alexandria*, vol. 25, no. 11, pp. 113–127, 2014, doi: [10.7227/ALX.0023](https://doi.org/10.7227/ALX.0023).
7. Truman, G., ‘Web Archiving Environmental Scan’, Harvard Library, Massachusetts, USA, 2016. Available:<https://dash.harvard.edu/handle/1/25658314>.
8. M. Costa, ‘Full-Text and URL Search Over Web Archives’, in *The Past Web: Exploring Web Archives*, D. Gomes *et al.*, Cham: Springer International Publishing, 2021, 71–84. doi: [10.1007/978-3-030-63291-5\_7](https://doi.org/10.1007/978-3-030-63291-5_7).