REQUIREMENTS AND DESIDERATA FOR THE SCHOLARLY USE OF WEB ARCHIVES

Eveline Vlassenroot
imec-mict-UGent, Ghent, Belgium

Sally Chambers
Ghent Centre for Digital Humanities, UGent, Ghent, Belgium

Friedel Geeraert
Royal Library and State Archives of Belgium, Brussel, Belgium

Peter Mechant
imec-mict-UGent, Ghent, Belgium

Introduction

The web and online information has become of utmost importance. However, the short lifespan of online data (with 40% of content being removed after 1 year) (Brügger, 2005, p.15) poses serious challenges for preserving and safeguarding digital heritage and information. Hence, web or media historians, sociologists or digital scholars must learn to "dig" in online sources such as the Internet Archive or national web archives in order to find relevant research material.

Thus, gaining insight into the needs and requirements of users of web archives is essential especially for web archives that are currently taking shape (as is the case in Belgium). However, web archiving institutions often do not have a lot of information about the use of their respective web archives (Bailey, Grotke, McCain, Moffatt, & Taylor, 2017). In addition to projects that took a qualitative approach in researching needs in relation to web archiving, see e.g. Stirling et al. (2012) , Dougherty & Meyer (2014) or Fernando et al. (2018), a number of studies have been published that analyse the needs of web archive users.

The perceived challenges of using web archives for research are well-documented in recent publications and research initiatives such as the BUDDAH project and RESAW. One of the most important challenges web archiving institutions need to overcome is the lack of awareness of the existence of web archives in the research community (Winters, 2017). Given that archived web content is relatively new research material, new skills need to be acquired to work with this content which is not something evident or something every researcher is willing to do (Gebeil, 2016; Winters, 2017).

Yakel & Thores (2003) point to three distinct forms of knowledge required to work effectively with primary data sources: (i) domain (subject) knowledge, (ii) artifactual literacy, and their own concept of (iii) archival intelligence. Domain, or subject, knowledge is an understanding of the subject being researched. Artifactual literacy is the "practice of criticism, analysis, and pedagogy that reads texts as if they were objects and objects as if they were texts." Archival intelligence is a person's knowledge of archival principles, practices, and institutions, and an understanding of the relationship between primary sources and their surrogates (Yakel & Thores, 2003).

Methodology

In this paper, we explore the requirements of researchers working with web archives and outline how they perceive the limitations and possibilities of using the archived web as a data resource, using survey data (n=154)[1]. The research question underlying this study is: "What are scholars' requirements for working with web archives as primary sources?". We asked researchers with and without experience of working with web archives for, amongst others, the search functionalities, descriptive metadata and selection and access criteria they require. The survey also requested basic demographic information (e.g. gender, nationality, highest level of education, current job role, area of study) and the prevalence of use of web archives. In order to assess the relationship between researcher's domain (subject) knowledge, artifactual literacy, archival intelligence and the prevalence of use of web archives, we elaborated on the work of Yakel and Torres (2003) and Van Deursen et al. (2014).

Our survey ran for two months and was disseminated in English, French and Dutch. The survey was spread through mailing lists, newsletters, websites and social media channels of the institutions participating in the project, targeted emails to library, archives and web archiving professionals as well as to researchers in the humanities and social and political sciences in Belgium and abroad. A dedicated effort was made to send personal survey-invites to relevant researchers and research groups.

Preliminary results

Preliminary results show that only 68% of the researchers are aware of the existence of web archives. When asking respondents with no experience of working with web archives why they are not using web archives the most popular answer was that they

were not aware of the existence of web archives. This corresponds with the findings of Costea (2018) who suggests that "a significant segment of the research community is still unaware of web archives and that many do not know exactly what they contain or how they can be used."

With regard to the specific requirements for scholars to work with web archives, there is no solution that can cater to all the wishes and requirements of all scholars in every domain (Hockx-Yu, 2014). Every scholar has different requirements depending on the research they are conducting, the skills they have and the tools available. Nevertheless we identified text/discourse analysis, exploratory research and retrieving lost documents/ web pages as the most important methods of analysis. Furthermore we found that scholars were most interested in the following topics: news websites (online journals and articles) and websites about culture/cultural history. Keeping these results in mind can prevent us from relapsing in a 'one size fits nobody' trap.

Few researchers have explored this paradigm of user expertise that was described by Yakel and Torres (2003) and found that users need a good deal of domain knowledge and artifactual literacy to process collections, create value-added finding aids, and fulfill user needs. Archival intelligence is what users learn in archival education programs, by reading and analyzing the archival literature and through years of practice in repositories (Duff, Yakel, & Tibbo, 2013). Nevertheless we found in our research that years of practice is not relevant to acquire archival intelligence but the frequency of use of web archives is. Furthermore, we would like to propose an extension to Yakel and Torres' 'archival intelligence' by introducing a new concept of 'web-archival intelligence'.

In addition to arriving at significant findings that demonstrate the relationships between researcher's domain (subject) knowledge, artifactual literacy, archival intelligence and use prevalence of web archives, this study discusses the limitations of using the archived web as a data resource and concludes with actions to overcome these hurdles and fulfill the desiderata of scholars.

References

Bailey, J., Grotke, A., McCain, E., Moffatt, C., & Taylor, N. (2017). *Web Archiving in the United States: A 2016 Survey*. National Digital Stewardship Alliance.

Brügger, N. (2005). Archiving Websites. Aarhus: Centre for Internet Research.

Costea, M.-D. (2018). *Report on the Scholarly Use of Web Archives*. NetLab.

Dougherty, M., & Meyer, E. T. (2014). Community, Tools, and Practices in Web Archiving: The State-of-the-Art in Relation to Social Science and Humanities Research Needs. *Journal of the Association for Information Science and Technology*, *65*(11), 2195–2209.

Duff, W.M., Yakel, E., & Tibbo, H. (2013). Archival Reference Knowledge. *The American Archivist,* *76*(1), 68–94.

Fernando, Z. T., Marenzi, I., & Nejdl, W. (2018). ArchiveWeb: collaboratively extending and exploring web archive collections—How would you like to work with your collections? *International Journal on Digital Libraries*, *19*(1), 39–55.

Gebeil, S. (2016). Quand l'historien rencontre les archives du Web. *Revue de La BNF*, (2), 185–191.

Hockx-Yu, H. (2014). Access and scholarly use of web archives. *Alexandria*, *25*(1-2), 113-127.

Stirling, P., Chevallier, P., & Illien, G. (2012). Web archives for researchers: Representations, expectations and potential uses. *D-Lib Magazine*, *18*(3/4).

Van Deursen, A. J. A. M., Helsper, E. J., & Eynon, R. (2014). *Measuring digital skills. From digital skills to tangible outcomes project report*.

Winters, J. (2017). Breaking in to the mainstream: demonstrating the value of internet (and web) histories. *Internet Histories*, *1*(1–2), 173–179.

Yakel, E., & Torres, D. (2003). AI: archival intelligence and user expertise. *The American Archivist*, *66*(1), 51–78.