Critical reflections on unlocking web archives for humanities research

Friedel Geeraert, Royal Library and State Archives of Belgium
Alejandra Michel, NADI-CRIDS University of Namur
Eveline Vlassenroot, imec-mict-Ghent University

The web is fraught with contradiction. On the one hand, the web has become a central means of communication in everyday life which increasingly holds the traces of our history created by everyday people[1]. Yet, much less importance is attached to its preservation, meaning that potentially interesting sources for future humanities research are lost. In Belgium, no web archiving initiative exists on the federal level, a problem which the PROMISE (PReserving Online Multiple Information: towards a Belgian StratEgy) project seeks to remedy by developing a pilot web archive for Belgium[2].

This paper explores the role that web archives can play in digital humanities research from three different perspectives: content access, selection and the surrounding legal framework. Web archiving is a direct result of the digital turn. Web archives have a role to play in knowledge production and dissemination as demonstrated by a number of recent publications[3] and research initiatives related to the research use of web archives such as the BUDDAH project[4] and RESAW[5]. However, different policy decisions and legislation shape web archives and therefore also influence both the source material that is put at the disposal of the researchers and the way in which this material is made accessible. A critical reflection is therefore necessary about how web archives can be unlocked for digital humanities research but also about the potential limitations related to using web archives.

Web content is particularly ephemeral. It has been estimated that forty percent of web content disappears after a year and that an additional forty percent is changed[6]. Web archives therefore hold a lot of potential for researchers. Using data in web archives however, comes with a number of challenges. Several of the hurdles that were mentioned by web archiving specialists during in-depth interviews which were conducted to

[1] Milligan, I. (2016). Lost in the infinite archive: the promise and pitfalls of web archives. *International Journal of Humanities and Arts Computing,* 10(1), 78-94. doi: 10.3366/ijhac.2016.0161.

[2] The PROMISE project was initiated in 2017 by the Royal Library and the State Archives of Belgium. The universities of Ghent and Namur and the university college Bruxelles-Brabant are partners in the project. For more information see: https://www.kbr.be/en/promise-project.

[3] An excellent example is Brügger, N. & Schroeder, R. (Eds.). (2017). *The web as history: Using web archives to understand the past and present*. London: UCL Press.

[4] In the BUDDAH (Big UK Domain Data for the Arts and Humanities) project, a number of bursaries were awarded to researchers for carrying out research in their subject area using the UK web archive. (BUDDAH, 2014).

[5] RESAW stands for Research Infrastructure for the Study of Archived Web Material and has been established ‘with a view to promoting the establishing of a collaborative European research infrastructure for the study of archived web material.’ (RESAW, 2012).

[6] Brügger, N. (2005) *Archiving websites. General considerations and strategies.* (The Centre for Internet Research). Retrieved from: http://cfi.au.dk/fileadmin/www.cfi.au.dk/publikationer/archiving_underside/archiving.pdf.

establish a state of the art of web archiving practices abroad are substantiated in literature[7]. Web users, for example, have become used to search engines that automatically rank search results according to their relevance. As this is not the case in web archives, users are confronted with data overload[8]. Selection is another important limiting factor. As it is impossible to capture everything on the web, decisions need to be made with regards to which URLs are captured as well as to the depth and frequency of this capture. Add to this the technical difficulties involved in crawling the web (missing dynamic content such as Twitter feeds, embedded videos, Flash, ...) and one realises that incompleteness is an inherent trait of content stored in web archives[9]. It is therefore very important to document the context within which the web archiving and quality control has taken place for researchers[10]. An additional hurdle are the access restrictions to web archives which can vary greatly between institutions: from very restricted access, only being accessible to researchers or only available on the premises of the web archiving institution, to freely accessible to all[11].

Moreover, the legal implications for digital humanists working with web archives are discussed. We will especially focus on the implications of two legal instruments. Firstly, the new General Data Protection Regulation[12] (hereafter "the GDPR") - which shall apply from 25th May 2018 - contains some interesting provisions for researchers. Indeed, the GDPR gives Member States the opportunity to create an exception from certain of its provisions when personal data are processed for the following purposes: historical or scientific research, statistical, as well as archiving in the public interest. Secondly, the

[7] Vlassenroot, E., Chambers, S., Di Pretoro, E. , et al. (2018). Web archives as a data resource for digital scholars. *Digital Scholar, 1(1).* (forthcoming)

[8] Webber, J. [British Library] (2017, November 16), *Interview with Jason Webber/Interviewers: Sally Chambers, Gerald Haesendonck, Alejandra Michel and Eveline Vlassenroot.* [M4A file]. Cowls, J., (2017). Cultures of the UK web. In N. Brügger and R. Schroeder (Eds.), *The web as history. Using web archives to understand the past and present*, (pp. 220-237). London: UCL Press. p. 235. ; Deswarte, R. (2015). *Revealing British Euroscepticism in the UK Web Domain and Archive Case Study.* Retrieved from http:// sas- space.sas.ac.uk/ 6103/.

[9] Ryan, M. [National Library of Ireland] (2017, November 16), *Interview with Maria Ryan/Interviewers: Gerald Haesendonck, Alejandra Michel and Eveline Vlassenroot.* [M4A file]. ; Winters, J. (2017). Coda: web archives for humanities research – some reflections. In N. Brügger and R. Schroeder (Eds.), *The web as history. Using web archives to understand the past and present*, (pp. 238-248). London: UCL Press. ; Brugger, N., & Finnemann, N. O. (2013). The web and digital humanities: theoretical and methodological concerns. *Journal of broadcasting & electronic media*, 57(1), 66-80. doi:10.1080/08838151.2012.761699.

[10] Tanésie, P., Aubry, S., Wendland, B. [Bibliothèque nationale de France] (2017, December 12), *Interview with Pascal Tanésie, Sara Aubry & Bert Wendland/Interviewers: Sally Chambers, Rolande Depoortere, Friedel Geeraert, Alejandra Michel, and Eveline Vlassenroot* [mp3 file]. ; Schroeder R., & Brügger, N. (2017). Introduction: the web as history. In N. Brügger and R. Schroeder (Eds.), *The web as history. Using web archives to understand the past and present*, (pp. 1-19). London: UCL Press, p. 11. ; Webster, P. (2017). Users, technologies, organisations. Towards a cultural history of web archiving. In N. Brügger (Ed.), *Web 25. Histories from the first 25 years of the world wide web*, (pp. 175-190). New York: Peter Lang.

[11] The Danish web archive (Netarkivet) for example is only open to researchers, the web archive of the Dutch national library can only be consulted in situ, while the Portuguese web archive (Arquivo.pt) is freely accessible online.

[12] Regulation (EU) 2016/679 of the European Parliament and of the Council on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation), 27 April 2016, L119/1.

Proposal Directive on copyright in the Digital Single Market[13] introduces a new mandatory exception that will facilitate the activities of research organisations: the text and data mining for the purpose of scientific research. The introduction of this new exception will bring multiple benefits. For example, it will provide researchers with more legal certainty and it will mean that researchers do not need to request the prior authorisation of right holders to perform reproductions or extractions for automated analysis of texts and data in digital form, if they fulfill the legal conditions[14].

Furthermore, the initial research results of the analysis of user requirements for a Belgian web archive will be presented. This analysis will focus specifically on user requirements with regards to the selection of, and access, to the content included in web archives. An online survey was launched in April 2018 and will run until the end of May 2018. A survey approach was taken in order to reach a wide audience of potential web archive users. The survey consisted of three different branches each aimed at a different target group: 1) people working as researchers, academics, students or who are in any way involved in research, 2) people working in archives, libraries or government institutions and 3) the general public (defined as 'active in another field/domain'). Each branch also differentiated between people experienced in working with web archives, and those who are not. The survey requested basic demographic information (e.g. gender, nationality, highest level of education, current job role, area of study), provided standardized questions to measure the respondent's digital literacy skills (operational and information navigation), challenges for (existing) web archives and preferences regarding various functionalities.

Other questions included in the survey were for example: which sources or information about which subjects would researchers most like to consult in web archives; the collaboration with heritage institutions for selection and curation of content; which questions would researchers like to answer using web archives; which search options and functionalities are considered indispensable; which methods for analysis are important for users; what kind of descriptive information (metadata) is perceived to be useful; what is the level of satisfaction of users who are working with existing web archives; how is data from social media platforms being used and in what format (e.g. .csv, .json) and how would users like to use these web archives (e.g. APIs).

In summary, this paper will discuss a) the potential of web archives for digital humanities researchers, b) introduce the legal framework surrounding the use of web archives and c) discuss the initial results of the user requirements survey with regards to the selection of and access to content in web archives.

[13] Proposal for a Directive 2016/0280 of the European Parliament and of the Council on copyright in the Digital Single Market, 14 September 2016, COM(2016) 593 final.

[14] Proposal for a Directive 2016/0280 of the European Parliament and of the Council on copyright in the Digital Single Market, 14 September 2016, COM(2016) 593 final, cons. n° 8.

Keywords: web archiving, user requirements analysis, digital humanities research, digital object curation, digital analysis of web archives, born-digital heritage collections

References

Brügger, N. (2005) *Archiving websites. General considerations and strategies.* (The Centre for Internet Research). Retrieved from: http://cfi.au.dk/fileadmin/www.cfi.au.dk/publikationer/archiving_underside/archiving.pdf .

Brügger, N. & Schroeder, R. (Eds.). (2017). *The web as history: Using web archives to understand the past and present.* London: UCL Press.

Brügger, N., & Finnemann, N. O. (2013). The web and digital humanities: theoretical and methodological concerns. *Journal of broadcasting & electronic media*, 57(1), 66-80. doi:10.1080/08838151.2012.761699.

BUDDAH, Big UK Domain Data for the Arts and Humanities. *Bursaries*. Retrieved from https://buddah.projects.history.ac.uk/news/bursaries/ . Last accessed on 04/02/2018.

Cowls, J., (2017). Cultures of the UK web. In N. Brügger and R. Schroeder (Eds.), *The web as history. Using web archives to understand the past and present*, (pp. 220-237). London: UCL Press.

Deswarte, R. (2015). *Revealing British Euroscepticism in the UK Web Domain and Archive Case Study.* Retrieved from http:// sas- space.sas.ac.uk/ 6103/.

Milligan, I. (2016). Lost in the infinite archive: the promise and pitfalls of web archives. *International Journal of Humanities and Arts Computing*, 10(1), 78-94. doi: 10.3366/ijhac.2016.0161

RESAW (Research Infrastructure for the Study of Archived Web Materials). (2012) *About RESAW.* Retrieved from http://resaw.eu/about/. Last accessed on 04/02/2018.

Ryan, M. [National Library of Ireland] (2017, November 16), *Interview with Maria Ryan/Interviewers: Gerald Haesendonck, Alejandra Michel and Eveline Vlassenroot.* [M4A file].

Schroeder R., & Brügger, N. (2017). Introduction: the web as history. In N. Brügger and R. Schroeder (Eds.), *The web as history. Using web archives to understand the past and present*, (pp. 1-19). London: UCL Press.

Tanésie, P., Aubry, S. & Wendland, B. [Bibliothèque nationale de France] (2017, December 12), *Interview with Pascal Tanésie, Sara Aubry & Bert*

Wendland/Interviewers: Sally Chambers, Rolande Depoortere, Friedel Geeraert, Alejandra Michel, and Eveline Vlassenroot [mp3 file].

Vlassenroot, E., Chambers, S., Di Pretoro, E. , et al. (2018). Web archives as a data resource for digital scholars. *Digital Scholar, 1(1).* (forthcoming)

Webber, J. [British Library] (2017, November 16), *Interview with Jason Webber/Interviewers: Sally Chambers, Gerald Haesendonck, Alejandra Michel and Eveline Vlassenroot.* [M4A file].

Webster, P. (2017). Users, technologies, organisations. Towards a cultural history of web archiving. In N. Brügger (Ed.), *Web 25. Histories from the first 25 years of the world wide web,* (pp. 175-190). New York: Peter Lang.

Winters, J. (2017). Coda: web archives for humanities research – some reflections. In N. Brügger and R. Schroeder (Eds.), *The web as history. Using web archives to understand the past and present,* (pp. 238-248). London: UCL Press.