http://doi.org/10.7207/twgn20-01

How Researchers Use the Archived Web

Peter Webster

DPC Technology Watch Guidance Note

April 2020

DigitalPreservationCoalition

© Digital Preservation Coalition 2020 and Peter Webster 2020

ISSN: 2048-7916

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, without prior permission in writing from the publisher. The moral rights of the author have been asserted.

First published in Great Britain in 2020 by the Digital Preservation Coalition.

1 Introduction

Opinions differ on the precise date of the birth of the World Wide Web, but we can now look back on a quarter century of Web history. Attempts to make a record of the Web began in the mid-1990s and became an established part of the work of libraries and archives (and national libraries in particular) in the first decade of the new century (Webster, 2017, 2019a). There now exists a vast but underused scholarly resource for the study of almost every possible aspect of the last two decades.

A new academic subdiscipline has taken a definite shape in the last five years, under the name of Web history. Despite the name, it is not only (or even mainly) historians who inhabit this new space, but humanists and social scientists of every kind, along with scholars of media and communication, library and information science, and computer science. It is now served by a dedicated journal, *Internet Histories*, along with reference works, monograph-length treatments of its scope and methods, and volumes of collected essays (listed in the References). There are also welcome signs that studies based on the archived Web are beginning to appear in the mainstream journals of individual disciplines.

There is a primary division within this diverse field between the study of the Web itself as a technical system, and the study of aspects of society and culture as they manifest themselves on it: a distinction between the history *of* the Web and history *on* the Web. The former tends to preoccupy those in computer science; the latter is most often the primary concern of humanities scholars. But the division is artificial to a degree, as the work of media and communication studies scholars shows; medium and message shape each other at every turn.

In this context of both novelty and diversity, this Guidance Note is designed to orient DPC member organisations, and others engaged in Web archiving (or intending to be), as to the kinds of uses researchers might expect to make of the content they collect. It is hoped that it may support the development of programmes of user research and engagement, and (in turn) inform collection development policies and the design of discovery and access services.

The scope of this Guidance Note is limited to the use of Web archives managed in one of the standard file formats (mainly WARC) and does not deal with archiving by screen capture. It also deals only with the open Web; the archives of social media platforms such as Facebook or Twitter present their own challenges and have a scholarly literature of their own. The studies cited are offered as examples of certain approaches rather than as endorsements of their substantive claims.

2 Five analytical strata

Niels Brügger set out a five-part scheme for thinking about the Web as an object of analysis, dividing it into five ‘strata’ (Brügger, 2018, pp.31-35). The first and smallest of these is the Web element: any coherent and identifiable single thing on the Web, be it an image, a portion of text, a navigation bar, a hyperlink. The second is the individual Web page: the array of individual elements that appears framed by a browser. The fifth and largest of the strata is the Web as a whole. In truth, few studies have focused solely on one Web element as it appears on a single page, or on single pages in general. Similarly, the task of studying the whole Web in terms of its content (rather than its technologies) has so far proved too vast and so has seldom been attempted. However, there are several studies which illuminate the nature of the whole Web by an analysis of elements that recur across it. Examples include studies of the development of the hyperlink (Helmond, 2019) or dealing with memes (McGrath, 2019), or with the evolution of the idea of the homepage (Mallapragada, 2019).

The two most commonly studied strata are the third and fourth, the third being the individual website. Individual sites that have been examined as a whole include those for universities (Nanni, 2017), media organisations (Nolan, 2017), and museums (Kahn, 2019). Ian Milligan's work on the now defunct GeoCities is of a single 'website' that was nonetheless written by a great many individual users (Milligan, 2017, 2019, pp.171-212).

Brügger's fourth stratum was what he called the 'web sphere': any grouping of elements, pages and/or sites that relate to some theme, be it an event, a kind of activity, or a geographic area. The individual web spheres so far studied are many and various. Just for the UK, there have been studies on organisations involved in public health (Gorsky, 2015), higher education (Meyer, Yasseri, Hale, Cowls, Schroeder and Margetts, 2017), private business (Musso and Merletti, 2016), and the military (Raffal 2018). A particular focus has been on individual national Web spheres. A recent volume of essays contains studies on Belgium, Canada, Denmark, France, Kosovo, the Netherlands, the UK and Ireland, and the European Union (Brügger and Laursen, eds., 2019).

3 Data and access

The archived Web, then, has in the last five years come into focus as an object of study for a great variety of scholars. As well as focusing their enquiries on different strata of the archived Web, scholars have also approached the archives themselves in different ways. The earliest – and still by far the most common – means by which users access the archived Web is through a browser-based playback mechanism, aided by some form of URL, metadata and/or full-text search. This playback method is typically how journalists use deleted materials to hold those in power to account. Where the archived Web is used in legal proceedings, it is also usually in this form: a visual representation of something on the Web at some point in the past, delivered to the researcher in their browser in real time.

Academic users have also made rich use of small collections of archived pages and/or sites in just this way, such as many of the essays in the seminal collection on *Web History* (Brügger, ed., 2010). More recently, studies on matters as diverse as the Islamic punk scene (Dougherty, 2017) or the history of the MMR vaccine crisis (Millward, 2019) have been built on a close reading of collections of pages and sites as they would have been visible to the user.

The archived Web can also be analysed at the less visible level of the source code that underlies it. Web archives provide the pre-eminent resource for understanding the technical development of the languages in which the Web has been written. More widely than that, the work of Anne Helmond (2017) points towards a history of the Web using its code – and third-party tracking code in particular – as a way of siting websites within a wider ecology of information flow and exchange with third parties.

Of course, source code can be read as closely as the visible webpage. But it may also be read programmatically at a much larger scale, and recent studies have done so, using links, images and (for the UK) postcodes. Webster (2019b) used a dataset of link relations between sites in collections from the UK to understand the nature of a particular web sphere in Northern Ireland. Milligan (2018) demonstrated the potential of large-scale image analysis across whole national domains and in smaller more cohesive web spheres. Most recently, geographers have begun to exploit the potential of UK postcode data in large collections as a means of understanding the Web's spatial nature (Tranos and Stich, 2020). Both this study and that by Webster used secondary datasets derived from primary WARC data, whereas Milligan worked directly with WARC files themselves. All three show the demand for access to data alongside web-based discovery and access services.

4 References

Brügger, N. (ed.) (2010) *Web History*. New York: Peter Lang.

Brügger, N. (2018) *The archived Web. Doing history in the digital age.* Cambridge, MA: MIT Press.

Brügger, N. and Laursen D. (eds) (2019) *The historical Web and digital humanities: the case of national Web domains*. London: Routledge.

Dougherty, M. (2017) '"Taqwacore is dead. Long live Taqwacore" or punk's not dead? Studying the online evolution of the Islamic punk scene'. In: Brügger, N. and Schroeder, R. (eds) *The Web as History*. London: UCL Press, pp.204-19. Available at https://www.uclpress.co.uk/products/84067 (Open Access). DOI: 10.14324/111.9781911307563

Gorsky, M. (2015) 'Into the Dark Domain: The UK Web Archive as a Source for the Contemporary History of Public Health'. *Social history of medicine*, 28(3), pp.596-616. Available at https://academic.oup.com/shm/article/28/3/596/1670384. DOI:10.1093/shm/hkv028 (Open Access).

Helmond, A. (2017) 'Historical Website Ecology. Analyzing Past States of the Web Using Archived Source Code'. In: Brügger, N. (ed.) *Web 25: Histories from the First 25 Years of the World Wide Web*. New York: Peter Lang, pp.139–155.

Helmond, A. (2019) 'A historiography of the hyperlink: periodizing the Web through the changing role of the hyperlink'. In: Brügger, N. and Milligan, I. (eds) (2019) *The SAGE Handbook of Web history*. London: SAGE, pp.227-41.

Kahn, R. (2019) 'The nation is in the network: locating a national museum online'. In: Brügger, N. and Laursen D. (eds) (2019) *The historical Web and digital humanities: the case of national Web domains*. London: Routledge, pp.161-77.

Mallapragada, M. (2019) 'Cultural history of the "homepage"'. In: Brügger, N. and Milligan, I. (eds) (2019) *The SAGE Handbook of Web history*. London: SAGE, pp.387-99.

McGrath, J. (2019) 'Memes'. In: Brügger, N. and Milligan, I. (eds) (2019) *The SAGE Handbook of Web history*. London: SAGE, pp.505-19.

Meyer, E.T., Yasseri, T., Hale, S.A., Cowls, J., Schroeder, R., Margetts, H. (2017) 'Analysing the UK web domain and exploring 15 years of UK universities on the web'. In: Brügger, N. and Schroeder, R. (eds) *The Web as History*. London: UCL Press, pp.23-44. Available at https://www.uclpress.co.uk/products/84067 (Open Access). DOI: 10.14324/111.9781911307563

Milligan, I. (2017) 'Welcome to the Web: the online community of Geocities during the early years of the World Wide Web'. In: Brügger, N. and Schroeder, R. (eds) *The Web as History*. London: UCL Press, pp.137-58. Available at https://www.uclpress.co.uk/products/84067 (Open Access). DOI: 10.14324/111.9781911307563

Milligan, I. (2018) 'Learning to "See" the Past at Scale: Exploring Web Archives through Hundreds of Thousands of Images'. In: Kee, K. and Compeau, T. (eds), *Seeing the Past with Computers*. Ann Arbor: University of Michigan Press,, pp.116-36 . Available at https://www.fulcrum.org/concern/monographs/70795899c (Open Access).

Millligan, I. (2019) *History in the age of abundance?* Montreal: McGill-Queen's University Press.

Millward, G. (2019). ‘A history with Web archives, not a history of Web archives: a history of the British Measles-Mumps-Rubella vaccine crisis, 1998-2004’. In: Brügger, N. and Milligan, I. (eds) (2019) *The SAGE Handbook of Web history*. London: SAGE, pp.464-78.

Musso, M. and Merletti, F. (2016) ‘This is the future: A reconstruction of the UK business Web space (1996–2001),’ *New Media & Society*, 18(7), pp. 1120–42. Available at: https://journals.sagepub.com/doi/10.1177/1461444816643791. DOI: 10.1177/1461444816643791

Nanni, F. (2017). ‘Reconstructing a website’s lost past: methodological issues concerning the history of unibo.it’, *Digital Humanities Quarterly*, 11(2). Available at http://www.digitalhumanities.org/dhq/vol/11/2/000292/000292.html (Open Access)

Nolan, S. (2017). ‘Born outside the newsroom: the creation of *The Age Online*’. In: Brügger, N. (ed.) *Web 25: Histories from the First 25 Years of the World Wide Web*. New York: Peter Lang, pp.107-22.

Raffal, H. (2018) ‘Tracing the online development of the Ministry of Defence and Armed Forces through the UK web archive’, *Internet Histories*, 2(1-2), pp.156-178. Available at: https://www.tandfonline.com/doi/full/10.1080/24701475.2018.1456739. DOI: 10.1080/24701475.2018.1456739

Tranos, E. and Stich, C. (2020). ‘Individual internet usage and the availability of online content of local interest: a multilevel approach’. *Computers, Environment and Urban Systems,* 79 (101371). Available at https://www.sciencedirect.com/science/article/pii/S0198971519300808. DOI: 10.1016/j.compenvurbsys.2019.101371 (Open Access).

Webster, P. (2017). ‘Users, technologies, organisations: towards a cultural history of world Web archiving’. In: Brügger, N. (ed.) *Web 25: Histories from the First 25 Years of the World Wide Web*. New York: Peter Lang, pp.175-90.

Webster, P. (2019a) ‘Existing Web archives’. In: Brügger, N. and Milligan, I. (eds) (2019) *The SAGE Handbook of Web history*. London: SAGE, pp.30-41.

Webster, P. (2019b) ‘Lessons from cross-border religion in the Northern Irish web sphere: understanding the limitations of the ccTLD as a proxy for the national web’. In: Brügger, N. and Laursen D. (eds) (2019) *The historical Web and digital humanities: the case of national Web domains.* London: Routledge, pp.110-23.