Web Archaeology: An Introduction

Aasman, Susan; De Haan, Tjarda; Teszelszky, Kees

Published in:
Journal of Media History

Susan Aasman, Tjarda de Haan and Kees Teszelszky

Web Archaeology: An Introduction

At the end of the twentieth century, when the scale and scope of the web was still limited compared to now, Roy Rosenzweig urged historians to acknowledge the increasing importance of the internet 'as a standard feature of everyday life'.[1] Roughly two decades later, historian Jane Winters concluded that, since the historical distance concerning the web 'is beginning to look like a reasonable chronological span', a more active embracing of doing web history might be appropriate.[2] That call is taken up more and more by scholars, recognising not just the web as a phenomenon with a past of which the history needs to be written, but also as a medium, that produces highly relevant historical sources about almost every aspect of society. Last decade, several books were published, a variety of specialised blogs were written, new peer reviewed journals started and dedicated conferences were held.[3] Together, these research activities demonstrate the potential of this field, varying from theoretical and methodological explorations to concrete case studies on the history of online pornography, of spam, of memes, or of the home page. Other histories address the emergence of webcam cultures in the nineties, the disappearance of the once very popular website GeoCities, but also the steady rise of online white nationalism online.

Many of these web histories deal with challenges that are related to issues of web archiving. They address questions such as: What are web archives, what do they collect or what have they collected so far and how do they differ from traditional (media) archives? While archiving the web is gradually acknowledged as a very important act in saving our cultural memory, the implications and critical understanding of what this act means for the type of collections or the type of sources produced, still needs further analysis. As web historian Ian Milligan observes in his recent book *History in the Age of Abundance. How the Web is Transforming Historical Research*: 'Web archives are not traditional archives not in content, form or conception.'[4] In addition, web archives are historical entities themselves, shaped over the years by changing conceptions of the web. Richard Rogers has described how specific historiographical points of view are built into web archives.[5] For instance, early web archives tried to collect and preserve single websites, while recently the focus has shifted to saving national web domains. An urgent issue that Rogers addresses is the prospects for archiving social media. As he admits, the first responsibility is the user herself who posts her vlog, selfie or tweet, but there are also institutional responsibilities for safeguarding our shared online presence. In any case, crawling and archiving social media through platforms like Facebook or Instagram will raise many issues related to size and scope, including technical problems, privacy issues and deliberate infrastructural hurdles. As Ian Milligan explains, these platforms should be seen as 'walled ecosystems' that might resist being

crawled. As such, we run the danger that a large part of our daily life online 'will largely be left of our history record', since neither Facebook, nor Google feel any obligation to actively pursue long-term preservation.[6]

In recent debates, there has been a growing attention for 'ad hoc' archives, as also Annet Dekker observed: 'It could be stated that today everything is archive and everyone an archivist.'[7] According to Dekker, we might want to look at these non-institutional examples as 'living archives', in the sense that they are 'open, collaborative and creative'.[8] As a consequence of this openness, these so-called living archives do not aim per se for a more conventional sustainable long-term storage strategy, as these archives are part of practices of active circulation and re-use of user-generated content. Who then should, and how then can institutional archives, together with individuals or communities of dedicated informal archivists, keep our collective presence online safe for the future? Abigail de Kosnik shows in her book *Rogue Archives*, how various groups of fans, political activists or other groups of amateurs, 'pirates' and hackers produce informal archival practices, creating interesting collections of new materials. Seen in this way, informal archival practices add to the safeguarding of our digital cultural heritage when standard archival procedures fail.

In addition to the archival issue, there is the growing body of research aimed at exploring the theoretical and methodological questions about the particularities of archived web records. In his groundbreaking work on web history and web archives, Niels Brügger shows how rethinking historical source criticism is crucial. For instance, one cannot expect a web archive to be 'an identical copy on 1:1 scale of what was actually on the web at a given time.'[9] The consequence is, according to Brügger, that archived web documents are by nature recreated, re-assembled and re-combined sources. 'Thus, the archived web document is the result of an active process and it does not exist prior to the act of archiving.'[10] Interaction between historians and web archivists becomes a necessary condition for doing web history.[11]

In this special issue of *TMG – Journal for Media History*, the focus is on the web history and especially on practices of web archiving. It showcases examples on how to actively engage as media scholars or as librarians and archivists with web archives and archived web-related born-digital sources. While it explores methods that deal with existing web archives, it will also address questions such as how to trace material retrospectively. In that sense, it is taking a 'web archaeological' approach, meaning that the focus is on actively uncovering the history of the web in its early days, emphasising the role of 'digging' and 'reconstructing' as central methods in tracing material objects (software, hardware, terminals, hard drives, cables, et cetera) and born-digital objects (websites, web elements like banners or avatars, blogs and vlogs, and many other forms of user-generated content).

In the peer-reviewed part, Anne Helmond and Fernando van der Vlist take this approach to a new and innovative level as they explore the challenges and opportunities of social media archiving. They show a specific interest in understanding the historicity of the social media platforms, which evolved in just a decade from systems on top of the open web's infrastructure to the very closed platform-based ecosystems of today. Earlier, we mentioned Ian Milligan's warning that because these platforms have become 'walled ecosystems', they might be left

out of history. Nonetheless, the authors introduce a methodology that makes it possible to perform historical research using social media platforms. By taking a multi-layered and multi-sided approach, researchers will avoid focusing solely on end-users, user-generated content or user interactions. Rather, the emphasis should be on the ways the platforms operated and how this changed over time. Helmond and Van der Vlist demonstrate how there are many material traces archived related to platform architectures, data strategies, infrastructural presences, et cetera, and that can be used as entry points for specific platform historiographies.

Whereas Helmond and Van der Vlist develop a strategy for historical platform studies, Susan Aasman focuses in her contribution on the archival issues related to just one particular platform, namely YouTube. Aasman is interested in the front-end of the early history of this platform and all the visible forms that YouTube produced: videos, comments, user profiles, tags, and so on. The author explores the ambiguous status of YouTube as an archival media space that contains millions of records while at the same time the platform is extremely volatile as video uploads disappear just as easily without leaving any trace. In order to understand the challenges and opportunities for reconstructing the early years of YouTube, Aasman evaluates the usability of both formal and informal web archival practices. These informal archival practices concern users who curate early YouTube videos, user profiles and comments or even old layout versions of the website. Developing methods that critically acknowledge the value of user-generated archival practices can lead to additional insights knowledge of that early community.

The second part of this special issue contains a dossier that showcases the craft of doing web history, or more precisely: web archaeology of the Dutch web domain. The history of internet in the Netherlands has shown an early and widespread active engagement of individual users and online communities who eagerly explored the new opportunities to publish content, develop new forms and formats or built active participatory communities. Many of these early users and usages have disappeared or become untraceable. As digital librarian Kees Teszelszky explains in his contribution, there is a serious gap in the way that cultural heritage institutions archived Dutch websites that were produced before 2007. His article narrates the search for archival traces of *De Opkamer*, probably the first Dutch online literary journal, by his team at the Dutch National Library (KB). It is an entertaining story of librarians locating – via different routes – material objects like floppy disks and assessing those as valuable artefacts containing unique data about the early history of the website. However, this contribution is also an interesting reflection on the overlapping temporalities when the archival practices of web archivists themselves becomes part of the – to be – archived object. The article serves as a reminder of Niels Brügger's warning explaining that web archiving is very much a process of document creating. In that process much more happens than preserving an original object.

The idea that archiving early web initiatives is more than locating and preserving visible web elements (e.g. web pages) or invisible elements (e.g. code), becomes quite clear in Tjarda de Haan and Erwin Verbruggen's article on the 'excavation' of the oldest Dutch online community De Digitale Stad (The Digital City), which existed from 1994 until 2001. This virtual city was the oldest Dutch online community and played a key role in the internet history of both Amsterdam and the Netherlands. In the user manual, the authors explain their web

archaeological approach, emphasising the role of 'digging' and 'reconstructing' as central methods in tracing and preserving material and born-digital objects. Web archaeology aims to dig in the various layers in the web's history, excavating (born-digital) objects, interpreting the traces and trying to reconstruct the original status of digital objects. According to De Haan en Verbruggen, the Digitale Stad was an online community that experimented with online forms of communication, creating along the way new objects, ideas and practices in new digital hybrid forms, like web pages, news groups, chats etc. To trace that constellation, material elements such as terminals, hard drives, cables and modems but also oral histories of the participants and their software programs, interfaces, emails and screenshots of avatars were collected. The richly illustrated article carefully instructs the reader about the different stages of not only the digging part of the process but also the set of choices you face once you want to preserve the material and make it accessible for researchers.

The issue concludes with a detailed exploration of the value of born-digital sources produced in the early years of the Dutch blogosphere. Iris Geldermans undertook a critical source analysis of archived weblogs. She points out the difficulties for historians who have to deal with the dynamic and fluid nature of this particular web genre, containing references to other blogs which might not have been archived, or containing only fragments of internal large archives of earlier blogposts. Geldermans analysed several archived versions of early Dutch bloggers as preserved in the Dutch National Library (KB). She noticed that because archived weblogs are static, they might lack both content and context. These kinds of case studies of weblogs should help make researchers aware of the challenges of web historical research.

To conclude, authors of this issue are very much aware of the fact that much of their work is a 'race against time'.[12] A certain portion of what has been produced online will be archived, but no doubt, a much larger portion will eventually disappear. According to Milligan, we are somewhere 'between everything and nothing'.[13] In the meantime, the issue also shows that we must acquire knowledge, skills and a critical appreciation of what Anat-Ben David recently called 'memoryware' a combination of preservation techniques that involves 'software and hardware, but also crawlers, bots, curators and users – through which the web's history is both documented, and constructed.'[14]

Notes

1. Roy Rosenzweig, "Wizards, Bureaucrats, Warriors, and Hackers: Writing the History of the Internet," *The American Historical Review* 103, no.5 (1998): 1530. doi:10.2307/2649970.
2. Jane Winters, "Breaking in to the Mainstream: Demonstrating the Value of Internet (and Web) Histories," *Internet Histories* 1, no. 1-2 (2017): 173-179. doi:10.1080/24701475.2017.1305713.
3. See for instance: Niels Brügger ed., *Web History* (New York: Peter Lang, 2010); James Curran, "Rethinking Internet History," in *Misunderstanding the Internet*, ed. James Curran, Natalie Fenton and Des Freedman (London: Routledge, 2012), 34-67; Megan Sapnar Ankerson, "Read/Write the Digital Archive: Strategies for Historical Web Research," in *Digital Confidential*, ed. Eszter Hargittai and Christian Sandvig (Cambridge: MIT Press, 2015), 29-54; Niels Brügger and Ian Milligan ed., *The Sage Handbook of Web History* (London: Sage, 2019); Ian Milligan, *History in the Age of Abundance? How the Web is Transforming Historical Research* (Montreal & Kingston: McGill-Queen's University Press, 2019); Niels Brügger, *The Archived Web. Doing History in the Digital Age* (Cambridge: MIT Press, 2019). Since 2017, there is the journal *Internet Histories. Digital Technology, Culture and*

Society, and special issues such as 'Born-Digital Archives', *in International Journal of Digital Humanities* 1, no. 1 (2019). Relevant blogs are for instance: Webstory (https://peterwebster.me/) by British historian Peter Webster or the Web Science and Digital Libraries research group (https://ws-dl.blogspot.com/). Important research infrastructures that organise (bi)annual conferences on web archiving are: IIPC (International Internet Preservation Consortium) Web Archiving Conference and RESAW network – A Research Infrastructure for the Study of Archived Web Materials. In July 2019, the latest edition was held in Amsterdam, the Netherlands: 'The Web that Was: Archives, Traces, Reflections'. Safeguarding and preserving born-digital heritage is, however, not only a matter of urgency and growing concern for scholars and heritage professionals. Also creators and policy makers need to take actions. See: "FREEZE! A Manifesto For Safeguarding and Preserving Born-Digital Heritage" (1 November 2017), https://hart.amsterdam/image/2017/11/17/20171116_freeze_manifest.pdf.

4. Milligan, *History in the Age of Abundance*, 69.
5. Richard Rogers, "Periodizing Web Archiving: Biographical, Event-Based, National and Autobiographical Traditions," in *The Sage Handbook of Web History*, ed. Niels Brügger and Ian Milligan (London: Sage, 2019), 42-57.
6. Milligan, *History in the Age of Abundance*, 88.
7. Annet Dekker ed., *Lost and Living (in) Archives. Collectively Shaping New Memories* (Amsterdam: Valiz, 2017), 5.
8. Ibid., 17.
9. Brügger, *Web History*, 6.
10. Ibid., 7
11. See for more on this: Brügger, *The Archived Web*, 137-140.
12. Milligan, *History in the Age of Abundance*, 66.
13. Ibid., 105.
14. Anat Ben-David, "Web Archives as Memoryware: Critical Reflections on Sources and Methods for Web History", Keynote at the 2019 International Internet Preservation Consortium (IIPC) Web Archiving Conference (WAC), hosted by the National and University Library in Zagreb (NSK), June 2019, https://www.anatbendavid.info/single-post/2019/06/11/Web-Archives-as-Memoryware-Critical-Reflections-on-Sources-and-Methods-for-Web-History.

Biographies

Susan Aasman is Associate Professor at the Centre for Media and Journalism Studies and director of the Centre for Digital Humanities at the University of Groningen. She was the editor-in-chief of TMG – Journal for Media History from 2012-2019. Her field of expertise is in media history, with a particular interest in amateur film and documentaries, digital culture and digital archives. She was a senior researcher in the research project 'Changing Platforms of Ritualised Memory Practices: The Cultural Dynamics of Home Movie Making' (2012–2016). Together with Andreas Fickers and Jo Wachelder, she co-edited Materializing Memories: Dispositifs, Generations, Amateurs (Bloomsbury, 2018). She also is the co-author of Amateur Media and Participatory Culture: Film, Video and Digital Media (Routledge 2019). Her current research involves projects that address the possibilities of using digital tools for doing media historical research.

Tjarda de Haan works as a web archaeologist, digital heritage expert and web director for her own company Bits and Bytes United, and as Head of Collections at Atria, knowledge institute for emancipation and women's history. She studied Contemporary History in Amsterdam and Berlin and worked as a guest curator for e-culture at the Amsterdam Museum.

Kees Teszelszky studied political science in Leiden and Eastern European studies in Amsterdam and holds a PhD in cultural history from the University of Groningen (2006). Since 2012 he has been researching web archives. He is curator of digital collections at the Dutch Royal Library (KB). He is involved in the selection, storage and accessibility of born digital sources, in particular websites, online news and e-books. He is also an enthusiastic web archaeologist in projects such as mapping the Dutch National Web Domain, the Dutch blogosphere and Dutch web incunabula.