Introduction: digital humanities and the use of web archives

Kees Teszelszky¹

Published online: 24 November 2021

© The Author(s), under exclusive licence to Springer Nature Switzerland AG 2021

In this special issue of the International Journal of Digital Humanities, the focus is on the use of web archiving and web archives for digital humanities research. The importance of web archives in digital humanities research is steadily growing, as our daily life becomes more and more digital, and most of the primary sources of the future history of our time are created on the web. Also, an ever-increasing amount of data is selected, archived, and presented worldwide in web collections by organizations like the Internet Archive, Common Crawl, national libraries, and other academic research and digital heritage institutions. The question is how to do research on this enormous amount of data, what the best tools and methods are at this moment, and how this data can be effectively presented to interested users. Ian Milligan and his co-authors argue in their article in this volume: ‘historians are not ready to use web archives.’ We can add to this: humans are not ready to use web archives and fully understand and make use of the possibilities the data in these web collections offer.

The current special issue of International Journal of Digital Humanities contains articles by academic scholars and international researchers from heritage institutions about the use of web archives and web archive data. The focus is on the challenges for researchers to analyse data and the possibilities of doing research beyond the Wayback Machine. Special attention is also given to lesser-known web archives and archived parts of the web like the archive section of Wikipedia, and the possibilities of social media archiving for digital humanities research.

The Wikipedia archive, which is a part of Wikipedia, is an understudied web archived part of the web. **Sandra Folie** shows in her ‘Recovering contemporary genre histories – the development of chick lit as seen through the Internet Archive’s Wayback Machine and Wikipedia’s history page’ how this archived data can be used to write the history of literary genres by tracking early online definitions of contemporary genres and their development. The author used the chick lit genre as a case study to show the value of including web archives in the recovery of contemporary

✉ Kees Teszelszky
Kees.Teszelszky@kb.nl

¹ KB, National Library of the Netherlands, Den Haag, The Netherlands

genre histories. The chick lit genre emerged in the second half of the 1990s with bestsellers such as Helen Fielding's *Bridget Jones's Diary* and Candace Bushnell's *Sex and the City* (both 1996), at the same time as the online web sphere devoted to literature and literary history developed. The definitions of chick lit on the websites under examination proved to be balanced and were also regularly updated to reflect new developments, such as the emergence of new subgenres. In this way they are excellent sources on the development of the genre, especially since academic sources tend to look down upon chick lit and offer little more than cursory glances on it. According to the author, it could be very productive in the future if the relationship between academia, nonacademic web sources, and web archives was to be strengthened, and if not only academic material found its way into digital formats and archives – as has been the case for many years in Wikipedia.

Jessica Ogden and Emily Maemura examine the challenges for researchers to analyse data from web archives in their article ‘Go fish’: Conceptualising the challenges of engaging national web archives for digital research’. There is a broad gap between a research question and the answers provided by available data from the web archives and tools for digital humanities research. The authors describe several sociotechnical and organisational limitations of web archiving in order to understand how these factors and contingencies influence research engagement with national web collections. They have studied the sociotechnical and organisational restraints of two national archives in detail: those of the UK Web Archive at the British Library and the Netarchive at the Royal Danish Library. According to the authors, it is important to enrol digital humanities scholars in web archiving research via research programmes, ‘datathons’ and workshops. The existence and promotion of these can also provide introductory or general use tools for analysing web archives which can be used by a broader public of researchers. A focus on early research activities of orientating, auditing, and constructing can promote greater investment in resources devoted to negotiating the (potentially) broad gap between a research question and the answers provided by available data and tools. The value of curators engaging with researchers directly at the earliest stages of research should not be underestimated. In this way, ‘go fish’ becomes less an impossible demand than an act of optimism—a prompt to explore the space between the ‘future promise’ of web archives and the contemporary challenges they pose for researcher use.

Valérie Schafer and Jane Winters’s paper ‘The values of web archives’ can be viewed as complementary to the article of Ogden and Maemura. This study lays the groundwork for the development, promotion, and adoption of a set of core values for web archives, linked to principles of “good governance”, that will help these collections to tackle the challenges of sustainability, accountability, and inclusiveness. All these values are central to the long-term societal and cultural worth of cultural heritage institutions. Web collections cannot be the sole property of experts, like academic researchers or curators. Web archives will have to respond to new forms of engagement, publication, and public history. According to the authors, we have to take into account and value the creation of tools, data publications, and other activities that are sometimes considered as trivial and ancillary but remain time-consuming and highly valuable for academic research and, more broadly, for the entire community interested in web archives. The authors state that the development of

committees composed of multistakeholders to assess “good governance”, the sharing of curricula dedicated to web archives, increased open and regular dialogue with representatives of all stakeholders, and the development of engagement by design, which means lowering the barriers to access could be important to open web archives for society.

Brenda Reyes Ayala explores the difference between the expectations of scholars and the realities of the web archiving process in her article ‘When expectations meet reality: Common misconceptions about web archives and challenges for scholars.’ She examines the misconceptions users of the Archive-It services have about the content of web archives, its use and how the web actually works. Archive-It is a web archiving subscription service of the Internet Archive. It allows institutions and individuals to collect, preserve, and access born-digital content from the web, and create digital archives without having to install or maintain software or infrastructure. Reyes Ayala examines the nature and type of misconceptions about web archives and challenges for scholars. Current web technology creates ‘black box’ websites which are not easy to understand and are therefore hard to archive and study. Web archivists have to be aware that not all can be preserved from a website, and, therefore, researchers must understand that not all can be studied. It is important that web archivists create proper documentation and researchers study the working of the web.

Niels Brügger discusses in his article ‘Digital humanities and web archives: Possible new paths for combining datasets’ the importance of web archives making their collections available as data, as opposed to the page per page viewing possibilities of the Wayback Machine. Web archives must be made ready to use for genuine digital research methods, which is not the case now. Most web archives only offer very limited access to their collections as data, such as extracted words, hyperlinks, or metadata in log files from the web crawling process. The author shows the importance of the possibility of combining metadata and hyperlink data from a web archive, and demonstrates that such a venture can provide valuable knowledge about the overall structure of a web domain. Data from the web archive must be made available as such, and information about the provenance of the collections, including documentation on curatorial choices and technical decisions regarding collecting, preserving, and extracting the data, should become transparent, so the result of analysis of a web archive can be better understood.

Janne Nielsen presents a historic study of the use of a specific tracking technology in her article ‘Using mixed methods to study the historical use of web beacons in web tracking’. She studies the use of the web beacon on the Danish web between 2006 and 2015 using data from the national Danish web archive. The author combines a large-scale quantitative mapping of the use of these web beacons on the Danish web with a qualitative study of specific websites. Using this mixed-method design, the author identifies the prevalent third-party domains setting web beacons and the different purposes for beacon use. The author shows that the number of Danish domains in the archived data decrease over time, both in the top of the list and generally, which indicates that international players play an increasing role in the tracking using web beacons compared to Danish players. This study shows us also how web history can be written by analysing archived data.

Nick Ruest, Samantha Fritz, Ryan Deschamps, Jimmy Lin and Ian Milligan introduce the ‘Archives Unleashed Cloud’ in their article ‘From Archive to Analysis: Accessing Web Archives at Scale through a Cloud-Based interface.’ Most researchers of web archives only have the Wayback Machine as an option to study the content of web archives or to get access to the archived websites. The Archives Unleashed Cloud is a web-based interface for working with web archives at a large scale, for instance the already mentioned Archive-It services. In their article, the authors introduce the Archives Unleashed Cloud as a response to the challenges which researchers face when dealing with web archives, like the amount of data, text, images or pages, messy metadata, or the difficulty to gain access. They argue that the Archives Unleashed Cloud presents a (much needed) specific solution both to the problems of web archive analysis and to more general problems of working with data at scale.

Although digital humanities research on social media from the live web has been done for quite a long time, social media archiving and research on archived social media content is still in its infancy. The challenges of preserving social media content are huge. The social media landscape has changed significantly in the last few years, with new platforms coming up and others going offline, an explosion in use and misuse, and a continuous change of function and technique. **Eveline Vlassenroot, Sally Chambers, Sven Lieber, Alejandra Michel, Friedel Geeraert, Jessica Pranger, Julie Birkholz and Peter Mechant** wrote an exploratory analysis of web archiving and social media. Their study is a continuation of their description of the state-of-the-art of national web archiving (published in the first issue of *International Journal of Digital Humanities*), which is now extended to social media archiving (Vlassenroot et al., 2019). They describe the current legal, technical, and operational aspects of preserving social media content, like the selection and preservation policy. Their analysis is complemented with results from an online survey to which 15 heritage institutions (national libraries and national archives) responded. The authors discuss and reflect on important challenges in and flaws of social media archiving which are very relevant for future researchers of this data.

To conclude, we can state that the authors of this issue show us that there are much more possibilities to study web archives and web data beyond the mere use of the Wayback Machine. We first need to study the humans behind the web archives and understand how web archives were created and by whom. We also have to look for other kinds of the archived web and for hidden data in the web archive.

The articles in this issue are thus a first step to understand and make use of the great possibilities of archived web data which web collections can offer for future research.

References

- Vlassenroot, E., Chambers, S., Di Pretoro, E., et al. (2019). Web archives as a data resource for digital scholars. *International Journal of Digital Humanities*, 1, 85–111. <https://doi.org/10.1007/s42803-019-00007-7>