Looking ahead: after web (archives)?

Helena Byrne, Beatrice Cannelli,
Carmen Noguera,
Michael Kurzmeier, Karin de Wild

WARCNET PAPERS

WARCnet
web archive studies

Looking ahead: after web (archives)?

Helena Byrne (the British Library), Beatrice Cannelli (University of London), Carmen Noguera (University of Luxembourg), Michael Kurzmeier (University College Cork), Karin de Wild (Leiden University)

helena.byrne@bl.uk

beatrice.cannelli@postgrad.sas.ac.uk

carmen.noguera@uni.lu

mkurzmeier@ucc.ie

k.de.wild@hum.leidenuniv.nl

WARCnet Papers

Aarhus, Denmark 2023

WARCnet Papers ISSN 2597-0615.

Helena Byrne, Beatrice Cannelli, Carmen Noguera, Michael Kurzmeier, Karin de Wild: *Looking ahead: after web (archives)?*

Published by the research network WARCnet, Aarhus, 2023.

Cover design: Julie Brøndum
ISBN: 978-87-94108-13-3

WARCnet
Department of Media and Journalism Studies
School of Communication and Culture
Aarhus University
Helsingforsgade 14
8200 Aarhus N
Denmark
warcnet.eu

The WARCnet network is funded by the Independent Research Fund Denmark | Humanities (grant no 9055-00005B).

WARCnet Papers

Niels Brügger: *Welcome to WARCnet* (May 2020)

Ian Milligan: *You shouldn't Need to be a Web Historian to Use Web Archives* (Aug 2020)

Valérie Schafer and Ben Els: *Exploring special web archive collections related to COVID-19: The case of the BnL* (Sep 2020)

Niels Brügger, Anders Klindt Myrvoll, Sabine Schostag, and Stephen Hunt: *Exploring special web archive collections related to COVID-19: The case Netarkivet* (Oct 2020)

Friedel Geeraert and Nicola Bingham: *Exploring special web archives collections related to COVID-19: The case of the UK Web Archive* (Nov 2020)

Friedel Geeraert and Barbara Signori: *Exploring special web archives collections related to COVID-19: The case of the Swiss National Library* (Nov 2020)

Matthew S. Weber: *Web Archives: A Critical Method for the Future of Digital Research* (Nov 2020)

Niels Brügger: *The WARCnet network: The first year* (Jan 2021)

Susan Aasman, Nicola Bingham, Niels Brügger, Karin de Wild, Sophie Gebeil and Valérie Schafer: *Chicken and Egg: Reporting from a Datathon Exploring Datasets of the COVID-19 Special Collections* (Dec 2021)

Olga Holownia, Friedel Geeraert, Abbie Grotke, Jennifer Harbster and Gulnar Nagashybayeva: *Exploring special web archives collections related to COVID-19: The case of the Library of Congress* (Feb 2022)

Niels Brügger: *The WARCnet network: The second year* (Dec 2022)

Michael Kurzmeier: *Using a national web archive for the study of web defacements? A case-study approach* (Aug 2023)

Valérie Schafer, Jérôme Thièvre and Boris Blanckemane: *Exploring special web archives collections related to COVID-19: The case of INA* (Aug 2020)

Niels Brügger, Valérie Schafer, Jane Winters (Eds.): *Perspectives on web archive studies: Taking stock, new ideas, next steps* (Sep 2020)

Friedel Geeraert and Márton Németh: *Exploring special web archives collections related to COVID-19: The case of the National Széchényi Library in Hungary* (Oct 2020)

Friedel Geeraert and Nicola Bingham: *Exploring special web archives collections related to COVID-19: The case of the IIPC Collaborative collection* (Nov 2020)

Caroline Nyvang and Kristjana Mjöll Jónsdóttir Hjörvar: *Exploring special web archives collections related to COVID-19: The Case of the Icelandic web archive* (Nov 2020)

Sophie Gebeil, Valérie Schafer, David Benoist, Alexandre Faye and Pascal Tanesie: *Exploring special web archive collections related to COVID-19: The case of the French National Library (BnF)* (Dec 2020)

Karin de Wild, Ismini Kyritsis, Kees Teszelszky and Peter de Bode: *Exploring special web archive collections related to COVID-19: The Dutch Web archive (KB)* (Nov 2021)

Michael Kurzmeier, Joanna Finegan and Maria Ryan: *Exploring special web archives collections related to COVID-19: The National Library of Ireland* (Feb 2022)

Emily Maemura: *Towards an Infrastructural Description of Archived Web Data* (May 2022)

Olga Holownia, Friedel Geeraert and Paul Koerbin: *Exploring special web archives collections related to COVID-19: The case of the National Library of Australia* (Dec 2022)

Helena Byrne, Beatrice Cannelli, Carmen Noguera, Michael Kurzmeier, Karin de Wild: *Looking ahead: after web (archives)?* (Aug 2023)

WARCnet Special Reports

Sharon Healy, Helena Byrne, Katharina Schmid, Nicola Bingham, Olga Holownia, Michael Kurzmeier, Robert Jansma: *Skills, Tools, and Knowledge Ecologies in Web Archive Research* (August 2022)

All WARCnet Papers and WARCnet Special Reports can be downloaded for free from the project website warcnet.eu.

Looking ahead: after web (archives)?

Helena Byrne (the British Library), Beatrice Cannelli (University of London), Carmen Noguera (University of Luxembourg), Michael Kurzmeier (University College Cork), Karin de Wild (Leiden University)

EDITORIAL NOTE

The five contributions in this WARCnet Paper are based on the authors' presentations in the keynote panel "Looking ahead: after web (archives)?" at the WARCnet Closing Conference in Aarhus, Denmark, 17-18 October 2022.

The panel took place Tuesday 18 October 10:00-11:30, and two months before the session the Organising Group of the conference had sent five questions to the panelists of which each panelist was asked to address at least two, at their own choosing:

(1) What do you consider the greatest challenges to having web archive studies advance?
(2) How to develop public and societal engagement with web archives?
(3) Is AI the future of web archiving and web archives studies?
(4) How can web archives be used alongside other sources, both analogue and digital?
(5) What new skills do web archivists and researchers of web archives of the future need (which they don't possess now)?

The panel presentations were followed by a discussion between panelists and by questions from the audience. However, these parts of the session are not included in the present publication.

The editors of the WARCnet Papers

WARCnet Keynote

Helena Byrne

We were tasked with responding to at least two of five questions but I chose to respond to all five questions.

WHAT DO YOU CONSIDER THE GREATEST CHALLENGES TO HAVING WEB ARCHIVE STUDIES ADVANCE?

The recent WARST report that I was involved in has highlighted that both the Library/Archive and Researcher communities found that the top three challenges they faced when working with web archive data were:

- Inconsistencies and incompleteness
- Legalities for acquisition/access
- Challenges with learning new skills

The Report also highlighted that these challenges persisted regardless of the experience of the respondee. In the report, we highlight how both communities would benefit from the provision of collaborative communal training across the full range of activities in the web archiving lifecycle. The study offers an overview of the types of skills and knowledge web archive practitioners and web archive users had prior to working with web archives, the skills they developed while working with web archives and the challenges they faced working with this type of resource.

We proposed that this might be used as a starting point to foster discussions in developing effective training materials for the necessary skills and tools for working with web archives across the spectrum of creator, curator, technician, or user/researcher. We further suggest that such training will also need to be benchmarked in a skills matrix, as it is very hard to develop and provide adequate training without a benchmark to measure against.

We also found that the challenges experienced by the participants in the study do not become less with increasing experience and highlight the need for training across all levels of experience. We suggest that, in order to develop targeted resources for both introductory and more advanced training, further research would be required to see how challenges shift with increasing experience across communities.

Web Archiving is a field that does not stand still. The ways in which people use and publish on the web as well as the technology they use to do this is always evolving. This means that challenges are always evolving and there is a need to keep evaluating the skills, tools, and knowledge in web archive research across the different communities involved (Healy et. Al., 2022).

HOW TO DEVELOP PUBLIC AND SOCIETAL ENGAGEMENT WITH WEB ARCHIVES?

This is a tough question because there is lots that can be done but not always the resources to do it. There are some good examples to follow within the web archiving community but one of the key challenges is consistency. Often promotion of web archives is around a single campaign or is for a short period of time due to project funding constraints.

Because of resource constraints, web archives generally do more targeted engagement amongst information professionals, researchers or specific communities that they want to represent in their collection.

The BESOCIAL project in Belgium had a crowd-sourcing campaign for nominations from the public to archive Belgian social media content that was also covered by the mainstream media (Messens et. Al., 2022). This campaign helped raise awareness of web archives in an audience that is not generally targeted by web archive organizations. Similarly, the UK Web Archive collaboration with the official UEFA Women's Euro England 2022 Heritage and Arts project saw the call for nominations to the UK Web Archive on the outdoor exhibition monoliths as well as promotional postcards distributed at fan events and in the UK Legal Deposit Libraries (UK Web Archive, 2022a, James, 2022, UK Web Archive, 2022b).

The examples outlined were only for a particular project and more needs to be done to integrate web archives into the general experience of the public, especially when they engage with arts and heritage.

IS AI THE FUTURE OF WEB ARCHIVING AND WEB ARCHIVES STUDIES?

I think AI has an important role to play in web archiving and web archive studies but I do not think it is the future. AI can be very useful for working with data at scale and will certainly aid with improving automation of certain processes but some of the reasons why it cannot solve all the problems in web archiving or web archive studies include:

- There are so many biases already built into algorithms that these biases would then also be reflected in web archive collections and research.
- The field of web archiving is in constant flux and there may not be the data available to develop new algorithms to adapt to the evolving challenges.
- Manual input from multiple sources for curating web archive collections is still important to ensure there is diversity within the collections and that content that is not possible to scope in automatically is included in collections.

- As web publishing knows, no boundaries and many web archives are restricted to collecting only in their legal jurisdiction it is hard to see how AI could be effectively used to scope in content in scope for collection under the UK NPLD but published on a .com server hosted outside the UK. Not all websites publish contact details or about information, that says which country they are publishing from.

HOW CAN WEB ARCHIVES BE USED ALONGSIDE OTHER SOURCES, BOTH ANALOGUE AND DIGITAL?

Researchers have often highlighted anecdotally to web archivists about the challenges of citing archived web content. This was also a challenge highlighted in the WARST report (Healy et. Al., 2022, p. 78). Not all citation guidelines encourage the use of content on the live web let alone refer to the archived web. At the IIPC WAC/RESAW conference in 2017 web archive citation was a theme that came up in a number of presentations (IIPC, 2017). The archived web as a reference resource is something that needs to be championed by the web archive community so that it is reflected in the citation guidelines. This is a task that would be well suited to WARCnet to lobby for as it is a balanced mix of academics and practitioners. I think this would help open up the possibility of researchers using web archives alongside other sources.

WHAT NEW SKILLS DO WEB ARCHIVISTS AND RESEARCHERS OF WEB ARCHIVES OF THE FUTURE NEED (WHICH THEY DON'T POSSESS NOW)?

Web archivists require such a varied skill set that it is very difficult for just one person to have all these skills. It would be far more sustainable for the variety of skills to be shared across a team.

Coding skills such as python would be very useful for working with web archive data to help analyze the data. I personally do not have any coding skills but some web archivists and researchers working with web archives do have these skills.

One key skill that is required for the future but we as a community are still in the early stages of developing is training skills. The majority of us have learnt how to work with web archives on the job, which makes recruitment and retention challenging. The IIPC Training Working Group has started to make some headway on this challenge (IIPC). By making, more general training programmes available for web archivists and researchers of web archives they will be able to adjust the changes in the field more easily and it would help bring more people on board with using this resource (Healy et. Al., 2022, p. 115).

BIBLIOGRAPHY

Healy, S.; Byrne, H.; Schmid, K.; Bingham, N.; Holownia, O.; Kurzmeier, M.; Jansma, R. (2022) *Skills, Tools, and Knowledge Ecologies in Web Archive Research*. WARCnet Special Reports, September 2022. Aarhus, Denmark: WARCnet.

International Internet Preservation Consortium (IIPC). (2017, June 14-16). Web_Archiving_Programme-14-16June.pdf. https://netpreserve.org/wp-content/uploads/Web_Archiving_Programme-14-16June.pdf

International Internet Preservation Consortium (IIPC). (Date unknown). Training materials. https://netpreserve.org/web-archiving/training-materials/

James, G. [@GaryJamesWriter]. (2022, July 7). @HBee2015 Your Web Archive postcard went down well on the @TraffordArchive stall in the #WEUROS2022 fan park yesterday. Together with our postcards & booklet it became a bit of a collector's piece. [Tweet]. Twitter. https://twitter.com/GaryJamesWriter/status/1544964499387650048

Messens, F, Denis, L-A, Heyvaert, P, Vlassenroot, E, Rolin, E, Wartin, P, Vandepontseele, S. (2022, September 15). Presentation of the BESOCIAL project results. [Video]. YouTube. https://www.youtube.com/watch?v=YkblD_F_ExU

UK Web Archivea [@UKWebArchive]. (2022, July 31). Check out these brilliant monoliths across the #WEURO2022 host cities. Can you help us preserve #FootballHistory? Nominate your favourite #WEURO2022 online content: https://blogs.bl.uk/webarchive/2022/06/what-content-should-i-nominate-uefa-womens-euro-to-ukwa.html #WebArchiving #NPLD #ENG #GER @WEURO2022 @WEUROTicketing [Tweet]. Twitter. https://twitter.com/UKWebArchive/status/1553713307407769600

UK Web Archiveb [@UKWebArchive]. (2022, July 19). If you happen to pass the @britishlibrary St. Pancras or Boston Spa lobbies, pick up one of our limited edition #WEURO2022 collection postcards. #WebArchiving #NPLD [Tweet]. Twitter. https://twitter.com/UKWebArchive/status/1549405771573739520

Looking ahead: After web (archives)

Beatrice Cannelli

INTRODUCTION

In the past couple of decades, the field of Web archive studies has certainly matured. As highlighted by Weber (2020), a rapid analysis of citations and abstracts comprising the expression "web archiving" demonstrated a clear growth in the use of web archives as research method in fields that go beyond media studies and history.

The interest toward web archive studies and the recognized cultural and historical value associated to an ever-growing number of digital traces that constitute our collective memory on the web has indeed increased, particularly in the last couple of years in response to recent political and health crisis. Moreover, the attention toward research utilizing the archived web has moved alongside the effort and dedication of many archiving institutions across the globe to develop national web collections, or transnational as in the case of the joint work done by members of the International Internet Preservation Consortium (IIPC).[1]

Web archive studies are still developing, following the rapid evolution of the web itself and the emerging of methods to approach and study the archived web. In order to enhance and find new paths for research using archived material from the web, it is important to identify and discuss some of the challenges to the advancement of web archive studies.

THE NEED FOR MORE AWARENESS AND THE PROBLEM OF ACCESS

Among the proposed themes for this panel, the first question I would like to address is indeed the one concerning barriers:

(1) What do you consider the greatest challenges to having web archive studies advance?

1. See the full list of collaborative collections built by members of the IIPC based on themes and critical events here: https://netpreserve.org/projects/collaborative-collections/

One of the main obstacles that seems to be holding back the development of web archive research is that it remains relatively unknown within academia and, in particular, in the field of digital humanities.

Web archive studies have considerably grown in the last couple of decades supported by, and supporting in return, the improvement of technologies and practices related to web archiving (Ben-David, 2021). Moreover, the fine-tuning and the reproducibility of specific research frameworks have been supporting research among scholars new to web archive studies, providing them with solid starting points (Brügger, 2021). Proof of the interest raised by web archive studies is also the considerable number of related conferences contributing to the sharing of knowledge and the consolidation of networks, such as those organized by the Web ARChives Research Network (WARCnet) itself, the IIPC[2] or the Research Infrastructure for the Study of Archived Web Materials (RESAW).[3]

Recent events such as the COVID19 crisis and the war in Ukraine have further reinforced the central role played by the web and social media in our daily communication and lives, leading many researchers from different disciplines to study content archived from the web (Aasman et al., 2021, 2022; Fritz et al., 2020; Nyvang & Hjörvar, 2020; Schafer et al., 2020). However, the number of researchers working with the archived web is still quite small. Often the cause can be traced back to the marginalized space dedicated to web archive studies in academia, making this one of the greatest challenges to the advancement of web archive studies.

Drawing from my personal experience, when looking back, for instance, at what I studied in my master's degree which focused on digital archiving and preservation, I am still surprised of how little of the courses I attended was dedicated to web archives or included the development of skills specifically related to the use of the archived web for research purposes. Certainly, not everything can be thoroughly covered in the span of one year or so but offering the opportunity to learn more about how to study the archived web or use it as a resource for a wide range of disciplines, ideally offering practical examples, would have opened new research prospects to a larger number of students and future researchers.

An additional obstacle to the use of web archives and consequently for the development of web archive studies is represented by the restrictions on access to web collections curated by national archives or libraries under non-print legal deposit legislation.

Due to often rigid national legal frameworks, data protection and copyright regulations, web archives implement rather restrictive policies in relation to access to the content preserved. Although some institutions provide access to the material at multiple selected locations across national territories, limitations on onsite access only still affect researchers who do not live close to these buildings. This represents an even greater burden for Early Career Researchers who might not have the necessary fundings or the budget to travel to specific institutions. The socio-economic factor is indeed among the

2. Further information about the annual IIPC Web Archiving Conference (WAC) can be found here: https://netpreserve.org/general-assembly/

3. More on the RESAW community and the conferences they host here: http://resaw.eu/events/

obstacles that affect access and engagement with web archives, as emerged from a recent study carried out by members of the WARCnet to investigate skills required and challenges related to participating in web archive research within different communities (Healy et al., 2022).

Onsite access only can be a further barrier for researchers seeking to use such collections to study transnational events, requiring them to request access and visit multiple, scattered institutions, for example, where remote access is not an option. Moreover, Healy et al. (2022) highlighted in their report that restrictions on how data can be accessed and used, often determined by having to work across different national legal frameworks, represent an additional layer of complexity for researchers collaborating on transnational projects involving web archives.

While legal and ethical matters will still limit the way researchers will be granted access to these collections in the foreseeable future, or at least until the material archived (especially from social media) will involve individuals who are still alive and whose lives might still be affected by the use of such content, it is essential to continue to work towards improving access to web archives.

Increasing awareness among researchers and the use of web archives across different disciplines would certainly sustain this process, as a high demand for consulting these collections may lead to easier ways to access them. Furthermore, building strong international collaborations between institutions - not only within the GLAM sector but also with universities - is an added point to consider when looking into how to improve both national and transnational usage of web archives and having web archives studies advance.

CROSSING BOUNDARIES: ENHANCING PUBLIC ENGAGEMENT WITH WEB ARCHIVES BEYOND ACADEMIA

The second question I would like to reflect upon is the following:

(2) How to develop public and societal engagement with web archives?

Raising awareness of an important resource such as web archives is vital not only among researchers, and across disciplines. As the web continues to change and often disappear causing irreplaceable loss, it is fundamental to also foster engagement with web archives beyond the boundaries of academia.

In fact, stimulating the engagement of the wider public could prompt a deeper appreciation of this kind of resources, illuminating the importance of preserving web-based content now rather than later. Moreover, public and societal engagement, especially when sought on a community level, represents a unique opportunity to adding further value to existing web archives, helping eventually to bridge gaps related to representativeness of collections and opening potential new paths for researchers to explore.

In this sense, good examples of attempts at seeking engagement with the wider public are all those crowdsourcing experiences organized by many memory institutions around the world, such as national libraries and archives, during the first few months of the COVID-19 pandemic (Zumthurm, 2021). In particular, all those campaigns inviting members of the public to suggest or submit testimonies, memories, and digital records shared on the web regarding the COVID-19 outbreak. Such experiences demonstrated how making the public become the protagonist of similar archiving efforts benefit public awareness and increase engagement with web archive collections.

Among the approaches that could help improve awareness of the public toward web collections and the cultural value that content shared on the web may hold, an interesting example comes from an experimental exhibition held at the Museum of London in the summer of 2022, titled “Into the Twitterverse”.

In this exhibition, curators from the Museum displayed content collected during the first lockdown from members of the public living in London, trying to capture the first-hand experience of the COVID19 crisis from different communities across the British capital. The content included in the exhibition spanned from tweets that had gone ‘viral’, to WhatsApp chats created by the curators with the purpose of collecting thoughts and comments from a selected group of Londoners during lockdown.

This experimental exhibition showed to the public how tweets posted by common people can assume a specific cultural and historical value, participating in the creation of our collective memory and supporting the study of extraordinary events, so much so to being displayed as any other work of art in a museum. In the exhibition, the collected born-digital material was showcased in creative ways, for instance, as “digital rain”, taking inspiration from the green computer code of a famous movie franchise; positioned over an interactive map of London which the user could explore through a touchscreen; or as printed and framed pictures, allowing members of the public to start perceiving in a different way what they often tend to treat as *just tweets*.

In order to increase public awareness of web archives and their value, both for research and society, similar experiences to those described above would certainly be of help.

Working more closely with minority groups and community archives, local museums, or libraries, would also go a long way towards increasing awareness of web archives, encouraging the wider public to engage with and explore these collections, at least those that are publicly available online.

REFERENCES

Aasman, S., Bingham, N., Brügger, N., De Wild, K., Gebeil, S., & Schafer, V. (2021). *Chicken and Egg: Reporting from a Datathon Exploring Datasets of the COVID-19 Special Collections*. WARCnet Paper, Aarhus.

Aasman, S., Clavert, F., de Wild, K., Gebeil, S., Brügger, N., Schafer, V., & Sirajzade, J. (2022). Studying Women and the COVID-19 Crisis through the IIPC Coronavirus Collection.

Ben-David, A. (2021). Critical web archive research. In *The Past Web: Exploring Web Archives* (pp. 181–188). Springer.

Brügger, N. (2021). The Need for Research Infrastructures for the Study of Web Archives. In *The Past Web: Exploring Web Archives* (pp. 217–224). Springer.

Fritz, S., Milligan, I., Ruest, N., & Lin, J. (2020). Building community at distance: A datathon during COVID-19. *Digital Library Perspectives*, 36(4), 415–428.

Healy, S., Byrne, H., Schmid, K., Bingham, N., Holownia, O., Kurzmeier, M., & Jansma, R. (2022). *Skills, Tools, and Knowledge Ecologies in Web Archive Research*.

Nyvang, C., & Hjörvar, K. M. J. (2020). *Exploring special web archives collections related to COVID-19: The Case of the Icelandic web archive*. WARCnet Papers. WARCnet.

Schafer, V., Thièvre, J., & Banckemane, B. (2020). *Exploring special web archives collections related to COVID-19: The case of INA*. WARCnet Papers. WARCnet.

Weber, M. S. (n.d.). Web Archives: A Critical Method for the Future of Digital Research. 17.

Zumthurm, T. (2021). *Crowdsourced COVID-19 Collections: A Brief Overview*. 4(1), 77–83. https://doi.org/10.1515/iph-2021-2021

Looking ahead: after web (archives)?

Carmen Noguera

Abstract: This text is based on an oral presentation at the panel Looking ahead: after web (archives) during the WARCnet Closing Conference in Aarhus on 17 and 18 October. I reply to two questions based on my experience as a Doctoral Researcher working on a PhD thesis on Digital Cultures and their development in Luxembourg (the 1990s- present day): How can web archives be used alongside other sources, both analogue and digital? What do you consider the greatest challenges to having web archive studies advance? In addition, I will briefly address the question: What new skills do web archivists and researchers of web archives of the future need (which they don't possess now)?

Keywords: web archives, web archivists, web archives challenges

HOW CAN WEB ARCHIVES BE USED ALONGSIDE OTHER SOURCES, BOTH ANALOGUE AND DIGITAL?

As many authors have argued, one can't write the recent history of the 1990s without using web archives as a historical source (Brügger, 2012, 2017). Using web archives doesn't mean we can't use other sources; they can be complementary, as we have seen in several studies. For instance, we could mention the research conducted by Schafer (2019) on the history of the Internet and the Web in France in the 1990s. She combined web archives (retrieved from the Wayback Machine and French BnF and INA web archives), newsgroups, oral interviews, site reports, press, and audiovisual archives. We could also add as an example the research of the Danish public service broadcaster DR's website by Brügger (2010), where he used web archives — a combination of his own archiving with Internet Archive — together with minutes of meetings, strategy papers, correspondence, etc. from the company archives.

In my concrete experience, web archives are an essential source without which I will miss a vital part of the history of the web in Luxembourg. I use web archives as a source (Brügger, 2009) and an object of study (Brügger, 2012). For example, web archives have been of great value to access the list of the first websites of the 90s, thanks to the repositories archived by Internet Archive in the early years of the Restena (Réseau

Téléinformatique de l'Education Nationale et de la Recherche) website. This gave me access to the web landscape of the 1990s. At the same time, web archives are a primary tool for defining my diachronic analyses of the selected case studies. For example, the Wayback Machine enabled me to analyze the evolution of the first participatory sites in Luxembourg in the 1990s, such as party.lu and luxusbuerg.lu, and see how they evolved throughout the years.

Nevertheless, web archives can't be my only source. I can highlight, for instance, the importance of oral interviews for my research. In these concrete mentioned case studies, although web archives are an essential source, there are many elements that one can't approach by using only web archives, such as the context behind, the goals, the ideas and motivations of the owners, their challenges and achievements, their relationship with stakeholders, their role in the Luxembourg ecosystem, their relationship with the audience, among others. On the importance and limitations of oral interviews for disseminating the history of the Internet and the web, we can mention the book *Oral Histories of the Internet and the Web* edited by Niels Brügger and Gerard Goggin, which brings together several interviews with key players on the history of the Internet and the web.

One practical example of the importance of oral interviews for my research has been applied to the analyses of multiculturalism and multilingualism on Luxusbuerg, the first chat platform in Luxembourg. From the study of the Wayback Machine, it could be seen that Luxembourgish was prioritized over other languages. English was used on the site menu, but Luxembourgish was the language for the interaction. What was the reasoning behind this decision? We could only know by the oral interviews and explore it further through press articles.

One of the owners of Luxusbuerg, Raoul Mulheims, explained the reasons for the English-Luxembourgish use in an oral interview I conducted. He stated that the use of English was partly because one of the partners was British and didn't speak Luxembourgish and because English was the predominant language of the Internet at the time. But the fact that the chat language was mainly Luxembourgish didn't allow much participation from the international community and, in a way, revealed some existing tensions among users, especially when writing in French. He further explained that they opened a channel called #francophone around 2003 to avoid this tension, and the French speakers could use their language without getting unpleasant comments.

In addition, we could see from the analysis of the channel rules of the most successful channel, #flirt, via the Wayback Machine that the Luxembourgish language gained more protagonism as Luxusbuerg evolved. For instance, in 2000, the rules recommended using Luxembourgish as the primary language for the chat.

> We will try to keep Luxembourgish the main language in the channel but French, English and German will be tolerated as long as the channel stays a "Luxembourgish one". We will NOT tolerate other languages under ANY circumstances, because that would result in an ultimate chaos with all this people. They may discuss as well in private. (Luxusbuerg, 2000)

In the same channel in 2005, Luxembourgish became the mandatory language from 6 am to 12 am.

> Wei den Numm et schon seet, ass Luxusbuerg een letzebuergeschen Chat. Aus deem Gronnd wellen mier am #flirt dass am Channel och just letzebuergesch geschwriwen get vun 6h-24h. Duerno kennt der dann och englesch, franseisch an deitsch schreiwen, mee eeben an Moossen. Waat mer op allenfall wellen vermeiden, ass dass am Channel 4 verschidden Sprochen geschuaat gin, oder dass eng aaner Sprooch wei letzebuergesch d'Iwerhand hellt. Daat get einfach zevill Chaos an Gedeesems. Fier dei wou franseisch wellen schwetzen, get et den #francophone, oder den Privat. An fier all aaner Sprooch gelt dann och: am Privat diskuteieren. (Luxusbuerg, 2005)
>
> As the name implies, Luxusbuerg is a Luxembourgish chat. For that reason, we want in the #flirt that only Luxembourgish is written on the channel from 6h-24h . Then you can write English, French, and German, but even in measure. What we want to avoid in any case is that several languages are spoken on Channel 4, or that a language other than Luxembourgish prevails. That's just too much chaos and confusion. For those who want to speak French, there is the #francophone, or the private. And for all other languages the same applies: discuss in private. (Luxusbuerg, 2005)

In addition, we complemented web archives and oral interviews with an analysis of the press articles at the time, which contributed to contextualizing further the case study. For example, in 2002, we could find the article "Internet: Gefahr oder Hilfe für das Luxemburgische?" ("Internet: Danger or help for the Luxembourgish?") in which the authors analyzed whether the Internet was helping to widespread the Luxembourgish or killing off minority languages. The authors concluded that the Internet and the web contributed to the rise of the Luxembourgish written language, mainly thanks to sites such as Luxusbuerg.

> Zunächst einmal kann man ganz eindeutig feststellen, dass das Internet als neues Kommunikationsmedium bereits sehr viel zum in der letzten Zeit zu beobachtenden Aufblühen der luxemburgischen Schriftsprache beigetragen hat, sei es im Web oder in der elektronischen Post. Noch nie wurde soviel auf Luxemburgisch geschrieben wie in den letzten zwei bis drei Jahren. Als Paradebeispiel gilt hier wohl die Web-Seite Luxusbuerg: http:/'/www. luxusbuerg. lu/ deren Gründer mit ihrem „Chatportal" eine „Plattform für die luxemburgische Onlinegesellschaft" schaffen wollten. Mit durchschnittlich 3 400 Benutzern am Tag' kann das Unterfangen jetzt schon als Erfolg bezeichnet werden. Man kann hier mühelos mit verschiedenen Leuten über die unterschiedlichsten Themen auf Luxemburgisch „plaudern". (D'Lëtzebuerger Land, 2002)
>
> It is quite clear that the Internet as a new communication medium has already contributed greatly to the recent rise in the Luxembourgish written language, whether in the web or electronic mail. Never before has so much been written in Luxembourgish as in the last two to three years. A prime example is the website Luxusbuerg http://www. luxusbuerg lu, whose founders wanted to create a "platform for the Luxembourg online society" with their "chat portal." With an average of 3,400 users daily, the reception can already be described as a success. You can easily "chat" with different people about the most diverse topics in Luxembourgish. (D'Lëtzebuerger Land, 2002)

But what makes web archives and their study different from other branches of historical studies? Bachimont, 2017; Brügger, 2008, 2012; Musiani et al., 2019 have discussed the

inherent characteristics of web archives and if they mark a rupture or continuity with respect to the traditional ones.

The authors agree in highlighting that web archives are a re-construction, which "involves a number of subjective choices where a lot of and unpredictable coincidences with regard to software, strategy, and purpose, integration in an archive and so on." (Brügger, 2008:.157), making it harder to evaluate the authenticity of the sources. The archived website does not exist before the act of archiving, but it is created using elements that were online at a given point in time (Brügger, 2012: 321). In other words, the web archive is not the perfect mirror of the live web, challenging the question of temporality as understood in traditional archives:

> La raison essentielle tient à la nature même des contenus et des procédures de collecte : en particulier, la durée de captation étant supérieure au rythme de mise à jour du site, l'archive résultant de la collecte rassemble en fait des parties de site renvoyant à des temps ou époques diférents du site. (Bachimont, 2017)
>
> The essential reason is the very nature of the contents and the collection procedures: in particular, the duration of capture being greater than the rate of updating of the site, the archive resulting from the collection, in fact, gathers parts of the site referring to different times or periods of the site. (Bachimont, 2017)

How can the researcher face this challenge? By developing new hermeneutics to study archived websites while connecting with the classical philology traditions, creating a "philology of the digital trace," as suggested by Bachimont.

The intrinsic nature of web archives or "documents of the web," a term coined by Brügger, 2012, leads us to the importance of developing new specific tools, capabilities and skillsets of web archivists and researchers to advance web studies.

WHAT DO YOU CONSIDER THE GREATEST CHALLENGES TO HAVING WEB ARCHIVE STUDIES ADVANCE?

Firstly, one of the main challenges I would mention is how to use all these digital, digitized, and born-digital sources and combine them and their metadata.

Secondly, I could highlight the challenges coming from the very nature of the web archives themselves. In general, the archived web has the following characteristics, as stated by Brügger (2018):

- an original is lacking, and I can't compare my archive with the original since it doesn't exist anymore,
- it is incomplete. I faced that on many occasions, for example, in analyzing party.lu, as several years are missing in the Wayback Machine,
- it is a unique version, not a copy of the online web, meaning it might have changed. For example, the archived version of party. lu on Wayback Machine in 1999 mixes the site party.lu with the company that integrated it, zap.lu, around 2007,

- there is a temporal and spatial inconsistency between the archived fragments. The previous idea is even related to that. One can perceive this when navigating through the hyperlinks in the archived website, experiencing temporal jumps.

Thirdly, I would emphasize as a challenge the lack of standard practices, documentation, and guidance framework**.** Web archives require exhaustive documentation, guidance, and a shared conceptual framework to systematize practices, advance the field, and engage with more users (Vlassenroot et al., 2019). The WARCnet platform is an example of knowledge and best practices sharing that helps scholars enter the field of web archives. Thanks to the WARCncet network, we will prepare a book chapter on the evolution of the .lu domain through web archives, together with Janne Nielsen, who has already studied the domain names of the Danish web (Brügger, Laursen and Nielsen, 2017), and which will contribute to the reproducibility of methodology in other countries.

Getting access to web archives is another challenge worth to be mentioned. For example, access to the web archives of the National Library of Luxembourg is only possible on their premises. And this is not only a specificity of Luxembourg. In various countries, there are different levels of restrictions according to national regulations and legal constraints (Brügger and Schafer, 2020), (Winters, 2017), (Vlassenroot et al., 2019), and there is a lack of international law on web archives access. Nevertheless, in the concrete case of Luxembourg, I should highlight as an asset the proximity to the library and the fact that I am working closely with the web archivists, for example, on the mentioned book chapter. From my experience, working with web archivists on metadata and distant reading is more fruitful for the research on both sides.

Another challenge I would mention is the limitations of the study of web archives compared to the online web. As stated by scholars, the technologies studying the live web have progressed quickly, but we can not use the same tools to analyze web archives.

In addition, it is challenging to have the user perspective from the archived web (Meyer, Thomas & Schroeder, 2011). For example, for the analysis of the websites of the 1990s, I can't access the past registered audience to understand behavior and traffic patterns and compare them with competitors.

Last but not least are the skillset challenges and technical knowledge, which might act as barriers. In the WARCnet paper "You shouldn't Need to be a Web Historian to Use Web Archives: Lowering Barriers to Access Through Community and Infrastructure," Milligan (2020) discussed the importance of lowering barriers to researchers willing to work with web archives. But still, scholars should be comfortable working with data at scale, and more specific training should be provided to them. This leads us to the question of the skillset needed from web archivists and researchers.

WHAT NEW SKILLS DO WEB ARCHIVISTS AND RESEARCHERS OF WEB ARCHIVES OF THE FUTURE NEED (WHICH THEY DON'T POSSESS NOW)?

I will reply to this question more briefly than the others, mostly because I am still learning what I can achieve with web archives. Milligan (2020) and Healy et al. (2022) provided

detailed information about the necessary skill sets for anyone working with web archives at scale. I highlight some of the ones that, from my experience, are very important:

- software and tools,
- programming, scripting languages,
- natural language processing,
- basic statistical knowledge,
- data analysis skills.

At the same time, the researcher needs to have basic knowledge of crawling to understand how the collecting software might have changed over time, and the differences within a website crawled at different times. Design and internet-related skills and generic information science skills are also essential for scholars working with web archives.

Additionally, researchers need a general knowledge of web archives, web archiving, and curation. This will enable them to understand the context in which web archives have been created and ask themselves why and how this data was collected.

The ability to perform source criticism and to address issues of sources reliability is essential due to the specificities of the reborn digital material. We can't be entirely sure that the source we have in the archive ever existed on the Internet in the same form. For example, evaluating the reliability of a source on the archived web implies comparing different versions to get as close as possible to the original one, leading us to the idea of developing new hermeneutics while returning to the classical philology tradition to better approach the study of web archives, as mentioned in the first part.

REFERENCES

Bachimont, B. (2017) Bruno, 2017, *L'archive du Web : une nouvelle herméneutique des traces ?*, Web Corpora, 21 juin 2017. [https://webcorpora.hypotheses.org/288]

Brügger, N. (2008). *The archived website and website philology: A new type of historical document?* Nordicom Review, 29, 155–175.

Brügger, N. (2009). *Website history and the website as an object of study*. New Media & Society 11(1–2): 115 – 32

Brügger N. (2010) Website history: an analytical grid. In: Brügger N (ed.) *Web History*. (pp. 29–59). New York: Peter Lang.

Brügger, N. (2012). *Web History and the Web as a Historical Source*. Zeithistorische Forschungen/Studies in Contemporary History, Online-Ausgabe, 9, p .316-325. https://doi.org/10.14765/ZZF.DOK-1588

Brügger, N. (2017). Webraries and Web Archives – The Web Between Public and Private, in D. Baker, & W. Ewans (Eds.), *The End of Wisdom?: The Future of Libraries in a Digital Age* (pp. 185–190) Oxford: Chandos Publisher. https://www.academia.edu/30729119/Webraries_and_Web_Archives_The_Web_between_public_and_private

Brügger, N., Laursen, D., & Nielsen, J. (2017). Exploring the domain names of the Danish web. In N. Brügger & R. Schroeder (Eds.), *The web as history: Using web archives to understand the past and the present* (pp. 62–80). London: UCL Press.

Brügger, N. (2018). *The Archived Web : Doing History in the Digital Age*. Cambridge, MA: MIT Press.

Brügger N. & Goggin, G. (Eds.) (2022). *Oral Histories of the Internet and the Web*. Taylor & Francis Ltd

Healy. S., Byrne H., Schmid, K., Bingham, N., Holownia, O., Kurzmeier, M., Jansma, R. (2022): *Skills, Tools, and Knowledge Ecologies in Web Archive Research*. WARCnet, Aarhus, Denmark.

Meyer, E., Thomas A., & Schroeder, R. (June 30, 2011) *Web Archives: The Future(s)*. SSRN Scholarly Paper. Rochester, NY. https://doi.org/10.2139/ssrn.1830025.

Milligan, I. (2020) *You Shouldn't Need to Be a Web Historian to Use Web Archives: Lowering Barriers to Access Through Community and Infrastructure*. Newark New Jersey USA: WARCnet. https://doi.org/10.1145/2910896.2910913

Musiani, F., Paloque-Berges, C., Schafer, V. & Thierry, B. *Qu'est-ce qu'une archive du Web?* OpenEdition Press, 2019. https://doi.org/10.4000/books.oep.8713

Mousel, P. & Lulling J. (2002, December 20) "Internet: Gefahr Oder Hilfe Für Das Luxemburgische?" D'Lëtzebuerger Land 49, no. no 51/52, p.16-17. Retrieved from https://www.land.lu/page/article/441/1441/FRE/index.html

Schafer, V. (2019) Exploring the "French web" of the 1990s in Brügger, N., and Laursen, D. *The Historical Web and Digital Humanities: The Case of National Web Domains* (pp.145-160). Milton, UNITED KINGDOM: Taylor & Francis Group. http://ebookcentral.proquest.com/lib/unilu-ebooks/detail.action?docID=5725928

Winters, J. (2017) Coda: Web archives for humanities research: some reflections', in Brügger & Schroeder (ed.) *The Web as History: Using Web Archives to Understand the Past and Present (*pp. 238-248) London: UCL Press.

Vlassenroot, E., Chambers S., Di Pretoro E., Geeraert, F., Haesendonck G., Michel A., & Mechant, P. (April 1, 2019). *Web Archives as a Data Resource for Digital Scholars*. International Journal of Digital Humanities 1, no. 1: 85–111. *https://doi.org/10.1007/s42803-019-00007-7*

Luxusbuerg channel rules. #Flirt. (2000) retrieved from Internet Archive. https://web.archive.org/web/20001208100700/http://www.luxusbuerg.lu/index.cgi?&site=flirt&language=eng&display=gl

Luxusbuerg channel rules. #Flirt. (2005) retrieved from Internet Archive. https://web.archive.org/web/20051029213044/http://www.luxusbuerg.lu/index.php?tab=content&channel=flirt&ContentID=67&PHPSESSID=5eb605fad3fe29547b649dcc410a8776

The Golden Age of Web Archiving

Michael Kurzmeier

Abstract: The following is an extended version of a talk given at the WARCnet closing conference in 2022. The text includes a figure which was not originally part of the talk, but helps to illustrate the argument.

This talk is my response to first two questions:

> *(1) What do you consider the greatest challenges to having web archive studies advance?*
>
> *(2) How to develop public and societal engagement with web archives?*

I started working with web archives in 2018, and joined WARCnet in 2020. For this talk, I want to look back at how my own engagement with web archives changed and evolved, and propose a discussion I think the experts in the field need to have. Back in 2020 I would have answered that the relative lack of tools and the relative lack of use cases are the main challenges. This would have been a somewhat evasive chicken and egg answer: The lack of tools makes it hard to produce use cases, and the lack of use cases makes it unattractive to develop further tools. Now that I got to work with the people at WARCnet and have seen the impressive research that comes out of this network so far, alongside significantly lower barriers of access to the field of web archive studies, I need to rethink that.

Instead of remaining uncommitted, I want to propose that we, in this current moment, are living at the end of the golden age of web archiving. With this, I do not mean to give the impression that everything related to archiving is great and all problems solved, nor do I want to sound overly gloomy. With this expression — that we are looking back on a golden age of web archiving — I rather want to describe changes I believe are already affecting the field and will only continue to do so in the future. At the same time, I want to warn you not to see the future of web archiving exclusively through the lens of past conditions, because these conditions have changed and will continue to change.[1] To illustrate these points, I have added a conceptual graph:

1. Borrowed from McLuhan's Rear-View Mirror concept (1967), web archiving efforts are at risk of seeing future archives through past archives.

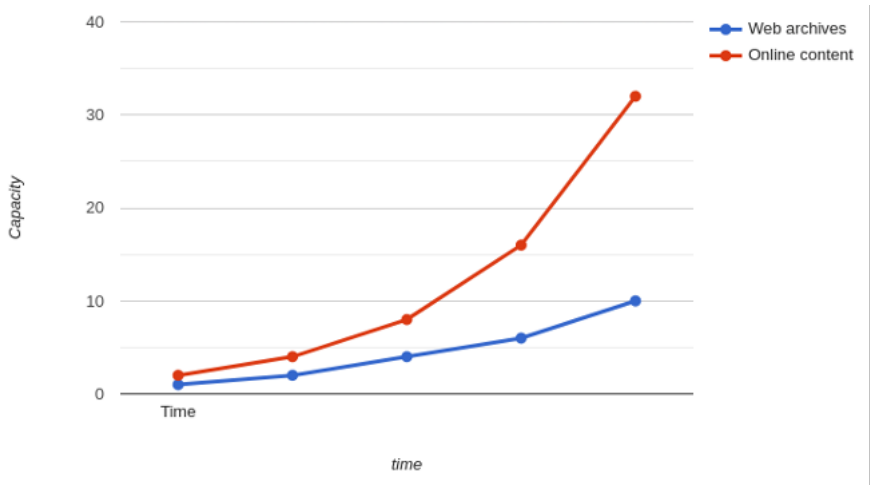

Figure 1: Web archiving capacity/Overall online content

This conceptual graph shows the growth of overall public online content and the overall web archiving capacity. The numbers are for illustrative purposes only and the timeline is deliberately vague. What the graph shows is that total web content grows at a much faster rate than archived content. It shows that while of course web archives increase their capture rate and storage capacity over time, they can not keep up with the overall growth in online content. What this graph does not describe, but what is another very important factor in understanding the changing role of web archives is the accessibility of online content over time. Generally, data has become a resource which is produced, exploited and retained within closed platform ecosystems.[2] To truly understand the situation, we would need to add another axis to this graph showing how much less accessible digital content has become over time.[3]

In this accessibility issue also lies the reason why — even through excessive upscaling of our capacity — we can not easily revert the situation. What I described as the golden age is the time period up to about the middle of the graph, where archiving capacity and online output were closer to each other than they are now. From there on, we have to acknowledge that accessibility and scale are ever growing issues. While archiving the front-end of a social media platform is “only” complicated by questions of scale and potentially infinite networks of hypertext, archiving usually has no access to the back-end of a social media platform. These challenges are exacerbated when we consider that most social media content is delivered through apps, where interface design is very different from HTML-based interfaces. Further, most web archiving workflows focus on text (both the machine-readable HTML and human-readable text as content), and have great difficulties dealing with video-based social media platforms.

2. For an introduction to this topic, see Zuboff (2019).

3. Part of the inspiration for this talk came from a presentation and discussion with Anat Ben-David based on her work on counter-archiving Facebook (2020). The other part came from a discussion with Vladimir Tybin. I owe thanks to both.

These developments create a paradoxical situation for the field of web archiving. While we might increase our absolute capture rate every year, our relative capture rate decreases. While we have more content in our archives than ever before, we find it harder from that content to extrapolate anything close to a representation of the historic web or the history on the web. If we do not engage with these new conditions web archiving is operating under, web archivists risk unwittingly specialising in text-based web content that is no longer representative of the majority of web content. Peter Webster reminds us that scale has always been a problem for web archives (2019, 34), and we can assume that the issue of scale — that is, how to handle and present vast amounts of data — will continue to be an issue in web archiving.

This change has a range of implications for web archiving, of which I am going to briefly outline two which are likely going to affect all aspects of web archiving.

- The first is that small and focussed collections will become more important than large, general collections. This is because smaller collections can to some extent mitigate the unavailability of content. This focus might be in terms of the time period covered (anything pre-2000, for example), or in terms of the research object itself (such as one specific hashtag on one specific platform).
- Secondly, and this point follows from the first, web archives will have to increasingly become political in the sense that they will have to engage and negotiate with platforms over access rights. Even in countries with relevant legal deposit legislation, the de facto power of platforms and the inherently unarchiveable structure of their content through algorithmic curation and backend-dependent services will force web archives to the table.

Having briefly outlined these changes and the consequent challenges for web archiving, it is time to reconsider what the goal of web archiving is and from that goal develop strategies in order to react to a change in the conditions web archives operate in.

And to do that, we should ask ourselves "What is the goal of web archive studies?" In a general sense, what is it that web archive studies want to achieve? If we understand Web Archive Studies merely as applied data analytics, we are stuck arguing for more and better data and we are dependent on whoever may provide that data to us. If we make the claim that our web archiving efforts produce a somewhat comprehensive view of history on the web or the history of the web, we have to be prepared for that claim to lose significance as the relative capture rate decreases.

Instead, if we understand Web Archive Studies as a potential gateway to public histories written from possibly one of the most comprehensive sources available, we change our perspective on the data we are working with. If we understand web archive studies as not only working with available (or rather, accessible) data, but as shaping the research data for the future, we can acknowledge the value this data holds and advocate for it to be used in line with public interest before private interest. We are beginning to see efforts in this field — for example, the combination of web archives with other data sources to “unlock the full potential of the treasure trove that web archives constitute”. (Brügger 2021).

This is the first and main challenge. We, as the web archiving community, must continue establishing Web Archive Studies not as an isolated, but a highly connected field of research. We must continue to show and advocate for the value of public access to the data from which our cultural records are written. If we break down this huge challenge into projects and objectives, we can begin to identify steps. To again just outline two, on a political level web archiving is in need of EU-wide legal deposit legislation, but also actual funding for institutions to carry out this mandate. This will be a long road, but I believe it is a worthwhile goal for web archivists and Internet historians.

On a practical level, we need more research into user requirements and delivery of tools to meet these requirements. But we also need critical reflection of the data we are working with, its origin and collection practice.

So if my 2019 answer would have been that we are in a "chicken and egg" situation, my 2022 answer is that we have to communicate the value of a web archive's underlying resource — data — to a wider public. Because through the shared understanding of the value of archived data held by national libraries and web archives, we (the web archiving community) can make a strong case.

The question is how to actually do that? How can we develop public and societal engagement with web archives? From my own experience, and Helena Byrne can weigh in on this — there is considerable interest in the topic of web archives as we saw through two Engaging with Web Archives conferences and the WARST survey project (Healy et al. 2022). Projects like these help introduce new researchers to the topic, they produce knowledge on how to use web archives and who uses them, and they also help form networks of researchers.

Also, as we are no longer in the described chicken and egg situation, we have greatly increased *showcase examples*. The Journal of Digital History and the WARCnet papers are just two examples of output formats that are very suitable for web archive research.

My short answer is that through continuation of these outreach events and publication formats, we can more effectively develop our main argument, which is based around the value of comprehensive data sets for humanities research. My longer answer is again referring back to the problem of diminishing accessibility. This problem already affects individuals and organisations who — often too late — find that commercial data collection platforms are not archives. This is not a new phenomenon, as cases such as Geocities and Myspace show, but one that forms a public who might not yet know why they should engage with web archiving tools and existing archives. In other words, developing public and societal engagement with web archives is another aspect of communicating the value of public data access. Tools and showcase examples are important, but they must be embedded in this advocacy context. I could not find a better way to put it than Jane Winter did:

> Web archives, and other kinds of born-digital data, do bring the possibility of, and perhaps even necessitate, a radical reframing of digital history – through their scale, their heterogeneity, their complexity, their fragility. [Historians] might, by combining big data approaches with humanistic understandings, at last begin to develop genuinely new research questions and generate new knowledge. (2019)

These new research questions and outputs which we now — four years later — are beginning to see, help to increase and communicate the value of web archives.

In conclusion, web archives will in future be less able to claim comprehensiveness and representation despite increased intake capacity. The relative share of captured content to content produced is likely to diminish, and so is the overall accessibility of content. This makes it necessary for web archives to engage in both platform politics (which I think they should avoid) and legislation (which I think they should seek). Communicating the value of data and the right to access data from which our historical record will be written is what I see as the key for web archives to mitigate the described changes and to develop public and societal engagement with the archived web. Web archivists should be aware that we are "past the peak" and that our understanding of web archives and web archiving practices may be too focussed on the text-based web. To deal with this problem, we need to engage in discussion about the changing nature of web archives and be open to new ideas, methods and technologies that can be used to capture and preserve the ever-evolving ecosystem of the web.

REFERENCES

Ben-David, Anat. 2020. "Counter-Archiving Facebook'." *European Journal of Communication, May* 026732312092206. https://doi.org/10.1177/0267323120922069.

Brügger, Niels. 2021. "Digital Humanities and Web Archives: Possible New Paths for Combining Datasets." *International Journal of Digital Humanities* 2 (1–3): 145–68. https://doi.org/10.1007/s42803-021-00038-z.

Healy, Sharon, Helena Byrne, Katharina Schmid, Nicola Bingham, Olga Holownia, Michael Kurzmeier, and Robert Jansma. 2022. "Skills, Tools, and Knowledge Ecologies in Web Archive Research." WARCnet Papers. https://cc.au.dk/fileadmin/dac/Projekter/WARCnet/Healy_et_al_Skills_Tools_and_Knowledge_Ecologies.pdf.

McLuhan, Marshall. 1967. *The Medium Is The Message*. http://archive.org/details/pdfy-vNiFct6b-L5ucJEa.

Webster, Peter. 2019. "Existing Web Archives." In *The SAGE Handbook of Web History*, edited by Niels Brügger and Ian Milligan, 30–41. London ; Thousand Oaks, California: SAGE Publications.

Winters, Jane. 2019. "Web Archives and (Digital) History: A Troubled Past and a Promising Future?" In *The SAGE Handbook of Web History*, edited by Niels Brügger and Ian Milligan, 30–41. London ; Thousand Oaks, California: SAGE Publications.

Zuboff, Shoshana. 2019. *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*. First. New York: PublicAffairs.

Looking ahead: after web (archives)

Karin de Wild

Abstract: What will be the future of web archival research? Web archives afford new opportunities for research. The aim of this WARCnet paper is to provide an exploratory overview of areas for future research. First, it will argue that one of the key challenges is access and advancing scholarly use of web archives. This will be followed by a reflection on public and societal engagement and in particular how to exhibit websites from the past. And it will conclude by reflecting on the potential of AI for computational studies and how this can make web archives more accessible.

Keywords: future, web archives, web history, challenges, public engagement, artificial intelligence

INTRODUCTION

In 2022, we were invited as early career scholars to join a keynote panel on the WARCnet Closing Conference at Aarhus University (Denmark), this to reflect upon the future of web archive studies. It was a period in which the Covid-19 pandemic had reminded us once again of the importance of preserving the web and why future historian will most likely increasingly use web archives. We are fortunate to enter the field of web archival studies while it is still emerging, but at the same time we can draw upon expertise of a growing group of scholars and archivists. Early attempts to archive the web started in the mid-nineties.[1] A pioneer within this field, the Internet archive, still holds the largest web archive collection with content going back as early as 1996. Although that they continue to function as an important forum for discussion and sharing expertise, today they are accompanied by a growing number of national web archives, each with their own collection policies and expertise. With the rise of web archives, there is also a growing body of literature on web history and how to make the web of the past come alive. Other important developments can be found in the field of digital history, which quickly gained momentum, and where methods and methodologies are developed for studying digital collections, like the web

1. Early attempts to archive material on the internet, including the web, were carried out in Canada in 1994–1995 (Brügger, 2011).

archives, and the computational turn towards big data. But although that the field of web archive studies is no longer 'new', there are still many opportunities for future research. The web is rapidly changing, both technically, as well as that the historical, social and cultural context in which it functions changes. And there are still many theoretical and methodologic challenges that may open up new directions in the field. It is my honor to briefly reflect on some possible area's for future research from my background in heritage studies, specialized in digital collections.

KEY CHALLENGES

The first question that I like to reflect upon is what could be some of the key challenges in web archive studies today? Now that we have these new historical records preserved in web archives and waiting to be further explored, which key challenges need to be tackled so that one can make full use of these resources? Web archives hold a wide variety of sources, from websites to blogs and tweets, millions of images and widgets. Also the content itself covers many different subjects, from early web publishing platforms like Geo-Cities, to social movements like climate activism and #metoo. These archives have a potential scholarly value in a broad range of disciplines that goes far beyond media studies, history and social sciences. Web archive studies is in its core a broad and interdisciplinary field that ideally advances through the development of a wide variety of approaches and expertise. However, there is still a barrier to make full use of these resources. First there are ethical and legal issues, that make us question how to access and use these resources responsible. Yet, also from a technical perspective there are still many questions. The large amount of (new types of) historical records, and incredible size of data, is not necessary easily accessible. Therefore, I like to start with highlighting this key challenge: How to make the sources in web archives more accessible?

The dominant access to web archival resources is currently through a page-page-page approach. Tools like the Wayback Machine and emulators (like oldweb.today) make it possible to replay an archived web page and get an impression of their look and feel. No longer do researchers need to access servers to study webpages from the past, but that does not mean that specialist knowledge is no longer required, especially digital source criticism becomes increasingly important. The archived web is, what Niels Brügger termed 're-born digital material', the original born-digital resource is changed while collecting and preserving it (Brügger, 2016). In other words, many of the web pages never existed as portrayed in the Wayback Machine. Web pages are, for example, quite complex sources that exist out of multiple files, like text, images, video, etc. An archived website as viewed within the Wayback Machine may involve patching together a number of files from different time periods. Through combining these files, it is possible to give an impression of how the web page used to look like in the past, but rearranging elements can obviously also change its original appearance. The archived web is in many ways different from other digital

sources, therefore it is essential to provide information on how to evaluate the reliability of the archived web.[2]

There is a relationship between the writing of web history and the possibilities and limitations of the tools available. It was pointed out by Ian Milligan that the scope of early scholarship in the history of the web was influenced by the wide use of the Wayback Machine (introduced in 2001) (Milligan, 2019, p. 233). This demonstrates the importance to expand methods and use a broader range of tools to be able to ask new questions. Besides the study of single web pages, scholars are increasingly interested in exploring (also) computational methods for studying the archived web. This could enable researchers to model the evolving state of the web over time in order to arrive at a better understanding of for example its influence on our daily lives, the way online events unfold or how a technology develops. The astonishing amount of data in web archives has obviously also huge implications for the methods and methodologies used.[3] However, the crucial first step for analysing web archival data, is that one first needs to gain access — and this is still a challenge.

Besides the legal and ethical concerns, there are also still practical issues at play, for example which routes do archives provide to search, request or extract derived datasets? There are best-practice examples, from dedicated Application Programming Interfaces (API's) to search engines (like the SolrWayback Machine) and other pathways can be found in the live web (e.g. SPARQL). Yet, not all web archives are already able to provide access. Plus, providing multiple routes for search and data extraction is not enough for providing access. Another essential component is improving the usability of the data. Is there a data dictionary that explains the data elements that are being used in the dataset or database? And not all historians are fluent in data processing, and also don't need to be. But how can we still overcome barriers through for example training, instructions, forming effective collaborations in interdisciplinary teams or by developing dashboards that can track, analyse and visually display the data? And is it possible to combine datasets, in other words is the data in web archives interoperable? To conclude, another essential component contributing to the accessibility of datasets is transparency about its quality and provenance. Historical data is inherently messy, there are many inconsistencies and gaps. Still it is essential to strive for transparency, as this could lead to better samples, as well as insights in how to combine and analyse the data.

Obviously, new problems will arise as soon as usable datasets are provided. Combining datasets into transnational corpora, for example, will open up many new challenges like dealing with ambiguity of different languages and words. Through the support of WARCnet, the field of web archival studies has organized itself in interdisciplinary working groups, each capable of solving certain problems. This offers a great opportunity for the field to further advance, yet regardless of the research direction, each analysis needs to start with concrete data. What is available will shape the field, making some studies more feasible than others. Therefore the key to advance the field (or maybe its current "Achilles heel"), is in my opinion improving accessibility.

---

2. See for example chapter 9 "Toward a Source Criticism of the Archived Web" (Brügger 2018).

3. As of January 1, 2023, the Internet Archive holds over 780 billion web pages in the Wayback Machine.

PUBLIC ENGAGEMENT

In this second part, I like to open up ideas for public and societal engagement of web archives and in particular the display of websites from the past. Some curators have taken up a pioneering role in exhibiting websites both in the online and offline realm. Personally, I am most familiar with the display of Internet art, art that is made on and for the internet. Originally the web was seen as the most appropriate environment for exhibiting these artworks. Websites are deeply shaped by the technical infrastructure that performs them and the infrastructure through which users access them. To put them in the immediate context of a physical gallery was seen as wrenching them from their original environment. The web also enabled artists to exhibit their work without the endorsements of institutions, giving them more artistic freedom and being able to defy the stability and canonisation, so typical of the museum. Although that the interrelationship with the environment stays obviously important for the meanings and values attributed to Internet art, today a wide range of display models is used: online, hybrid and offline; as well as both outside and inside institutions. I will briefly highlight some examples. Most of the exhibitions present artworks, but nevertheless may these insights be relevant for the display of websites more in general.

In the 1980s, the digital space was experienced as something outside of reality, a constructed world that can be entered, experienced and explored, but that is separated from our everyday lives. This idea of a separate virtual realm was mostly expressed within the term 'Cyberspace', popularized by science fiction writer William Gibson in his book *Neuromancer* (Gibson, 1984). He described it as a "consensual hallucination" created by millions of connected computers. "Cyberspace" was being positioned as a new and largely imaginary territory. The first galleries for Internet art were developed in line with this view, outside of our physical realm. For example, in 1999 the Guggenheim Virtual Museum was initiated by Senior Curator of Film and Media Arts John Hanhardt and Curator Matthew Drutt. The ambition was to become the first virtual museum dedicated to the display of Internet art and to display these artworks in a unique cultural destination on the web (Rashid, 2017). The gallery spaces were fluid, interactive and it transcended gravity so that the visitor was able to fly through the exhibitions. The online galleries aimed for a unique spatial experience, that would be un-imaginable in our real world. It offered an opportunity for fully exploring the new opportunities of the web, a new to discover territory.

While virtual galleries used to be designed as exceptional spaces, inherently different from everyday life, nowadays we see that there is more contiguous with "offline" exhibition spaces. An example is the exhibition "World on a Wire" (2021) organized by Hyundai and Rhizome. This exhibition transformed the gallery space into a hybrid-reality. The exhibition presents a selection of artist-made synthetic life forms that exists both in the physical and online realm. The works call into question the distinction between reality and representation, our relationships with technology and the natural environment. The exhibition title is drawn from a 1973 TV movie by German director Rainer Werner Fassbinder in which a massive computer simulation causes the protagonist to question whether his own reality is also a virtual construction, which ultimately drives him insane. This terrifying thought, that

simulations cannot be easily separated from the world it seeks to model, is taken as a point of departure in the exhibition. As simulations of life, some of the artworks were able to generate continuously new and unpredictable forms. Therefore, the works were presented simultaneously in physical galleries (at Hyundai Motorstudios worldwide), as well as in the online realm. After the closure of the physical galleries, the artworks remain accessible online, where it is possible for audiences to follow their evolving state at any moment in time. The exhibition space was neither separate from real space nor simply a continuation of it, instead it explored the complex interplay between the real and the virtual. And the exhibition was no longer visible in a unified, coherent space, but like the web it was distributed over several platforms at the same time.

Looking at the specific features of websites reveals some paradigmatic changes that challenges their display within institutional galleries. One of those key characteristic is that they are in a constant state of flux. Once collected, only a snapshot of the website is captured - a static and fixed version of what was once dynamic content. A more rigorous understanding of a website would have to acknowledge its instable character, but how to embrace this open-ended meaning-making and contrast it with the fixity of artworks when displayed within traditional galleries? Websites are constantly re-shaped under the changing pressures and perspectives of the present. In that sense, exhibitions are interesting as they show a similar process of forming and re-forming the past. They embed artworks within different (art) historical, social and political contexts. Exhibitions often give us a contemporary perspective on the past, even when they value displaying the 'original' state of the artworks. But can the artwork in an exhibition be treated as living things that transforms under the influence of new contexts? And what if an artwork exists in multiple versions and even continuously further evolves? Then which version to put on display? In 2015, I wrote a paper on the process of re-exhibiting online artworks and how this involves continously finding new display forms in terms of reinstalling a technical system, but also re-creating social contexts of which the form is far from uniform (Wild, 2016). This vision will be further explored in the upcoming exhibition 'Reboot' at the New Institute (Rotterdam), the Dutch museum for Architecture, Design and Digital culture. A team of curators, scholars and artists will rethink in the upcoming months the displays of a large group of digital (and various online) artworks with as a central question how to display these artworks now and in the future. It was also involve reconstructing some of the exhibition histories of these artworks to gain a more comprehensible understanding of their evolving lifespan, essential for finding ways for interpretation and relevant display forms.

ARTIFICIAL INTELLIGENCE

In this last section, I like to conclude with speculating on the influence of the rapid rise of Artificial Intelligence (AI) on web archive studies. Yet, my first thought is that all suggested below may alter. In the past we saw that the rise of new technological advancements, often goes hand in hand with utopian and dystopic ideas. In the end, the advancement of the technology itself is an experimental process of which the outcomes are hard to predict. Therefore, a simple answer to predict the influence of this technology is rather impossible,

but still it is important to think beyond the AI hype and instead try to reflect on the more sustainable values of its applications.

As discussed, web archival data is one of the largests heritage collections and to further understand its content, the support of computational methods will become increasingly important. Web archives are still (relatively) isolated from other cultural heritage datasets and data in the live web, yet there is the potential to make the (meta)data interoperable to develop accumulated bodies of knowledge that move away from institutional and disciplinary regimes. Yet the bigger the dataset, the more complex it is to make sense of them. It can quickly reach a point where the data becomes impossible to read and it even can become invisible to the human eye. Therefore, we are increasingly in need of smart algorithms for accessing, managing and organizing this data. No wonder then, that AI models are becoming a necessity to support us in finding where we are looking. Mining both textual, as well as visual data can help us give meaning to the flows of information on the web (and beyond) and transform it into useful knowledge.

So how will AI be used in web archival studies? One promising application is that AI can support researchers in finding what they are looking for in big data collections. A recent example is a study that we developed as part of WARCnet and in collaboration with the Archived Unleashed Project Team.[4] In an interdisciplinary team and through a series of datathons, we are exploring how deep mining large web archival collections can help find entry points for humanities research (Aasman S., 2023). Machine learning models can support disovering essential research areas (frequently occuring topics), hidden themes (associative topics) or trends (topics that are relevant over a longer period of time). Currently, three existing models are tested: Latent Dirichlet Allocation (LDA), Word2vec, and Doc2vec. Outcomes can be used as a starting point for a qualitative study, but it can also be used to identify and build topic-specific corpora to dive deeper into the research area through quantitative analysis. The final aim is to develop a methodology that involves various approaches, each with slightly different outcomes, that will allow researchers to select a route that fits their research purposes best.

Although that Artificial Intelligence offers many opportunities for new methods and methodologies and may open up exciting new directions in the field of web archival studies, there are also still many challenges to overcome. For example, how to deal with the uncertainty about the status of the archived material? What is the scope of a collection and which essential curatorial decisions (by humans and machines) were taken? Is there already documentation about how each archived website entered the collection? Another area, where web archival studies can contribute, is further understanding and eliminating human and societal biases, not only in the data itself, but also in the way AI systems are developed, deployed and used. The field is still developing processes and practices to test for and mitigate bias in AI systems. While the technology is further progressing, I hope that romantic or uncanny attitudes towards these technologies will not have too much effects on

4. This study is conducted by Joshgun Sirajzade, Karin de Wild and Frédéric Clavert, supported by WARCnet and the Archived Unleashed team, that is Ian Milligan, University of Waterloo (CA), Nick Ruest, University of Waterloo (CA), Valérie Schafer, University of Luxembourg (LU), Niels Brügger, Aarhus University (DK), Susan Aasman, University of Groningen (NL) and Sophie Gebeil, University of Aix-Marseille (FR). Archive-IT and the IIPC generously gave access to Web archival data at scale.

new directions. Instead it is essential to expand AI literacies, also within the field of web archival studies. Only then can we further explore how humans and machines can effectively (and ethically) analyse the web of the past.

REFERENCES

Aasman S., N. Brügger, F. Clavert, K. de Wild, S. Gebeil, V. Schäfer, J. Sirajzade. (2022, December 20). Studying Women and the COVID-19 Crisis through the IIPC Coronavirus Collection. IIPC netpreserve.org. https://netpreserveblog.wordpress.com/2022/12/20/studying-women-and-the-covid-19-crisis-through-the-iipc-coronavirus-collection/

Brügger, N. (2016). Digital Humanities in the 21st Century: Digital Material as a Driving Force. *Digital Humanities Quarterly*, 10(2). www.digitalhumanities.org/dhq/vol/10/3/000256/000256.html

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

Gibson, W. (1984). *Neuromancer* (1st ed.). New York: Ace Books.

Milligan, I. (2019). *History in the age of abundance? How the web is transforming historical research.* Montreal: McGill-Queen's University Press.

Rashid, H. (2017, 25 July). *Learning from the Virtual.* E-flux Architecture, Post-Internet Cities. http://www.e-flux.com/architecture/post-internet-cities/140714/learning-from-the-virtual/

Wild, K. de (2016). An interactive mnemonic space for Jodi.org: the process of re-exhibiting. *22nd International Symposium on Electronic Art (ISEA),* Hong Kong.

WARCnet Papers is a series of papers related to the activities of the WARCnet network. WARCnet Papers publishes keynotes, interviews, round table discussions, presentations, extended minutes, reports, white papers, status reports, and similar. To ensure the relevance of the publications, WARCnet Papers strives to publish with a rapid turnover. The WARCnet Papers series is edited by Niels Brügger, Jane Winters, Valérie Schafer, Kees Teszelszky, Peter Webster and Michael Kurzmeier. In cases where a WARCnet Paper has gone through a process of single blind review, this is mentioned in the individual publication.

The aim of the WARCnet network is to promote high-quality national and transnational research that will help us to understand the history of (trans)national web domains and of transnational events on the web, drawing on the increasingly important digital cultural heritage held in national web archives. The network activities run in 2020-22, hosted by the School of Communication and Culture at Aarhus University, and are funded by the Independent Research Fund Denmark | Humanities (grant no 9055-00005B).

WARCNET PAPERS

INDEPENDENT RESEARCH FUND DENMARK

warcnet.eu warcnet@cc.au.dk twitter: @WARC_net facebook: WARCnet
youtube: WARCnet Web Archive Studies slideshare: WARCnetWebArchiveStu