Exploring the 20-year evolution of a research community: web-archives as essential sources for historical research

Niels Brügger[1]
Professor of Media Studies at the School of Communication and Culture, Head of the Centre for Internet Studies and of NetLab, Aarhus University, Denmark.

Valérie Schafer[2]
Professor of Contemporary European History at C²DH (Luxembourg Centre for Contemporary and Digital History), University of Luxembourg, Luxembourg.

Interview edited by:

Friedel Geeraert, KBR,
Web-archiving scientific assistant

Nadège Isbergue, KBR,
Periodicals manager

Sally Chambers, Ghent Centre for Digital Humanities,
Digital Humanities Research Coordinator

Summary
The *PROMISE* project could not be conceived without tackling the question of access to Belgian web archives. Indeed, if it is important to save them, it is because of their importance as sources of information for research. Therefore, during the conference "Saving the Web: the Promise of a Belgian Web Archive", the access and the use of this specific type of archives were widely discussed. Two of the speakers at this conference, Valérie Schafer and Niels Brügger, agreed to share their experiences regarding the use of web archives, some of which, go back to the early 2000s.

Samenvatting
Het *PROMISE*-project kon niet worden bedacht zonder de kwestie van toegang tot Belgische webarchieven aan te pakken. Het belangrijk is om de Web te bewaren, omdat het belangrijk is als informatiebron voor onderzoek. Tijdens de conferentie "Saving the Web: the Promise of a Belgian Web Archive" werd daarom veel aandacht besteed aan de toegang en het gebruik van deze specifieke archieven. Twee van de sprekers op deze conferentie, Valérie Schafer en Niels Brügger, waren bereid om hun ervaringen te delen met betrekking tot het gebruik van webarchieven, die soms teruggaan tot begin jaren 2000.

Résumé
Le projet *PROMISE* ne pouvait se concevoir sans aborder la question de l'accès aux archives du web belge. En effet, s'il est important de les sauvegarder, c'est en raison de leur importance comme sources d'information pour la recherche. Par conséquent, lors du colloque "Saving the Web: the Promise of a Belgian Web Archive", l'accès et l'utilisation de ces archives spécifiques ont été largement abordés. Deux des intervenants lors de ce colloque,

[1] https://www.netlab.dk/netlab/people/niels/
[2] https://www.c2dh.uni.lu/people/valerie-schafer

Valérie Schafer et Niels Brügger, ont accepté de partager leurs expériences respectives quant à l'utilisation d'archives web, celles-ci remontant parfois au début des années 2000.

Introduction

The history of the World Wide Web already spans more than a quarter of a century[3]. The Internet Archive, as well as several National Libraries and Archives, have already been archiving the web for over 20 years[4]. Yet, the use of the archived web as an object of research remains at the fringes of (digital) humanities research[5]. Although many researchers in the humanities and social sciences still need to explore the potential of these archives, the research community around web-archives has been steadily evolving alongside these developments. Particular examples include the Big UK Domain Data for the Arts and Humanities (BUDDAH) project[6], the research being undertaken by members of RESAW, the Research Infrastructure for the Study of Archived Web Materials[7] and most recently WARCNet: Web ARChive studies NETwork researching web domains and events[8].

Within the context of the Belgium web-archiving project *PROMISE*[9], piloting access to the Belgian web archive for scientific research has been considered a key task from the outset. During the project, a survey with almost 400 respondents was conducted which aimed to understand "What are the requirements and needs of potential users of web archives?" and "How do they, or do they want to, access, use and consult web archives?". The results from this survey helped the project team to understand to what extent web archives can be seen as a data resource for digital scholars[10]. A particular challenge in this area is providing access to archived web-resources. Many web-archives remain solely accessible through dedicated computers inside (national) libraries due to legal restrictions. The article "Tour d'horizon sur les aspects légaux de l'archivage du web" by Alejandra Michel in this volume explores this issue in depth.

To conclude the *PROMISE* project, on 18 October 2019, KBR organised the colloquium "Saving the Web: the Promise of a Belgian Web Archive"[11]. This colloquium was the opportunity for the researchers of *PROMISE*'s project to present their work regarding web archiving. In addition to Neil's Brügger's keynote: "National web archives: the land of

---

[3] See for example, Brügger, N. (Ed.) (2017) *Web 25: Histories from the First 25 Years of the World Wide Web*, New York: Peter Lang Publishing.

[4] Vlassenroot, E., Chambers, S., Di Pretoro, E., Geeraert, F., Haesendonck, G., Michel, A., & Mechant, P. (2019). Web archives as a data resource for digital scholars. *International Journal of Digital Humanities*, *1*(1), 85-111. https://doi.org/10.1007/s42803-019-00007-7

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

[6] https://buddah.projects.history.ac.uk

[7] http://resaw.eu/about/

[8] https://cc.au.dk/en/warcnet/

[9] https://www.kbr.be/en/projects/promise-project/

[10] Vlassenroot, E., Chambers, S., Di Pretoro, E., Geeraert, F., Haesendonck, G., Michel, A., & Mechant, P. (2019). Web archives as a data resource for digital scholars. *International Journal of Digital Humanities*, *1*(1), 85-111. https://doi.org/10.1007/s42803-019-00007-7

[11] https://www.kbr.be/en/colloquium-saving-the-web/

promise for researchers"[12], which is included as an article in this volume, the final session of the colloquium was dedicated to research use of web-archives. Presentations during this session included Eveline Vlassenroot: "User Requirements for a Web Archive"[13], based on the experiences of the *PROMISE* project; "Coordination of web archiving in the Netherlands" [14] by Jesse de Vos from the Netherlands Institute for Sound and Vision. Furthermore, Patricia Blanco, a Masters Student in Digital Humanities at KU Leuven and an intern on the *PROMISE* project described her experiences of using web archives for research in "Saving the Belgian Web: Web archiving practices, research opportunities and limitations"[15], which is also elaborated as an article in this special issue.

This event provided the opportunity to welcome international names in the field: Niels Brügger[16], Professor of Media Studies at the School of Communication and Culture, Aarhus University, Head of the Centre for Internet Studies and of NetLab and Valérie Schafer[17], Professor of Contemporary European History at C²DH (Luxembourg Centre for Contemporary and Digital History), University of Luxembourg. As a result, the interview that follows, provided the occasion not only to delve into a deeper discussion around the key themes, ideas and questions evoked during the *PROMISE* colloquium, but also to reflect on how these issues can help pave the way for future research in this area, both in Belgium and beyond.

An interview with Niels Brügger and Valérie Schafer

1. How did you become involved in studying the archived web?

Niels Brügger (NB): I have an academic background in French Language and Culture, and the History of Ideas, but in 1997 I moved from the French Department to Media Studies. At the time only one of my new colleagues studied the internet, but without having a clear historical perspective. Thus, there was an uncharted territory to go into for someone like me who was interested in the web and its very short history. But I soon found out that my object of study disappeared for my very eyes — the online web was changed or deleted in a way that was not familiar to anyone within media studies. So, my first step as a wannabe web historian was to ensure that I had a stable object to study. I therefore started to tinker with web archiving myself, but quickly realised that a professional organisation was needed here to make this happen at scale. Around the same time — in the Autumn of 2000 — I co-founded the Centre for Internet Studies at Aarhus University, and in our mission statement we mentioned that one of our aims was to have a national web archive established, without having any idea about how feasible that was and if that could be done at all. We sent out a press release about the new centre and it was well received by staff at the two national Danish libraries: the State Library in Aarhus and the Royal Library in Copenhagen, and together with

[12] https://www.kbr.be/wp-content/uploads/2019/11/2019-10-18_Saving-the-Web_Brugger.pdf
[13] https://www.kbr.be/wp-content/uploads/2019/11/2019-10-18_Saving-the-Web_Using-web-archives-for-researchers.pdf
[14] https://www.kbr.be/wp-content/uploads/2019/11/2019-10-18_Saving-the-Web_deVos.pdf
[15] https://www.kbr.be/wp-content/uploads/2019/11/2019-10-18_Saving-the-Web_Using-web-archives-for-researchers.pdf
[16] https://www.netlab.dk/netlab/people/niels/
[17] https://www.c2dh.uni.lu/people/valerie-schafer

them we got funding for a one-year pilot project aiming at investigating how a national Danish web archive could be established. To make a long story very short: the final report[18] from the project published in 2002 served as the basis for the discussion in the Ministry of Culture which later led to the revision of the Legal Deposit law passed in December 2004 where 'computer networks' were included, and in June 2005 the Danish web archive Netarkivet became a reality.

Valérie Schafer (VS): Thanks to Niels Brügger! In 2011 I edited an issue of the French journal *Le Temps des Médias* dedicated to the history of the Internet[19] and Niels proposed an article to us dedicated to Web archives. I started to extend my research to the history of the Web about a year later, after work more devoted to the history of the Internet itself, its infrastructure, protocols, etc. My first contact with the *Wayback Machine* was a revelation! I was immediately fascinated by Web archives and it hasn't stopped since. It was also the beginning of stimulating collaborations with Niels Brügger and colleagues who were working on these issues like him and whom he was able to bring together. The date of 2011 may seem very late as the *Wayback Machine* has been available online since 2001, but I can assure you that researchers who were already interested in Web archives such as Louise Merzeau, Fabienne Greffet, Dana Diminescu and a few others were still very rare in France at the time. Archivists and librarians were ahead of the game in this field.

2. What research projects related to the archived web are you currently involved in?

NB: I am currently heading two large projects. The first started some 3-4 years ago, and it aims at mapping the entire Danish web domain .dk and its development from 2005-2015. This is only possible since we have access to the copies of the Danish domain in *Netarkivet* where the entire web domain is archived four times per year. In addition, we have had access to *Netarkivet*'s High Performance Computer 'The Cultural Heritage Cluster'[20] which has made it possible to run large scale quantitative analyses. Among other things we have investigated the number of specific file types, and we now know that the ratio between text and image files only evolves slowly: we get more pictures, but we also get more text. However, over 10 years, the number of pictures grows a bit more than the amount of text. Also, we investigated the billions of hyperlinks on the Danish web and could map how many links to social media were on the Danish web in the period (these results can be found in a recent article "Big data experiments with the archived Web"[21]). The second project is a network, called *WARCnet*, 'Web ARChive studies network researching web domains and events', the research aim of the project is to investigate how national web domains and events on the web have developed over time. The project includes researchers and web archives from Denmark, France, Luxembourg, the Netherlands, Germany, the UK, and Belgium. The events we will focus on could be terrorist attacks, sport events, or elections, and definitely the COVID-19 pandemic will be studied (read more about *WARCnet* on the project website[22]).

[18] http://netarkivet.dk/publikationer/webark-final-rapport-2003.pdf
[19] https://www.cairn.info/revue-le-temps-des-medias-2012-1.htm
[20] http://en.statsbiblioteket.dk/kulturarvscluster/
[21] https://firstmonday.org/ojs/index.php/fm/article/view/10384
[22] https://cc.au.dk/warcnet/

VS: From 2014 to 2018 I was involved in two research projects that came to an end with my transfer from the CNRS[23] (France) to the University of Luxembourg. One of these projects was dedicated to the Web of the 1990s in France (*Web90*[24]) and the other to the digital traces and born-digital heritage of the 2015 attacks (*ASAP*) in partnership with the Bibliothèque nationale de France (BnF) and the Institut national de l'audiovisuel (Ina). This second project had a strong focus on digital social networks. Since then, I have dedicated myself in particular to developing in the field of teaching, at master's level. This included a training course on Web archives, in the framework of a week-long winter school involving Web researchers and archivists. I have also had the pleasure of publishing the book; *Qu'est-ce qu'une archive du web?*[25] with three colleagues from the *Web90* project, mentioned above. Additionally, I broadened my reflection on the sustainability of Digital studies and Digital Humanities, by crossing several facets of digital heritage, whether digitised or born-digital. Finally, I'm involved in the *WARCnet* project, launched in 2020, as previously mentioned by Niels.

3. Why is archiving the (national) web indispensable for research and for society at large?

NB: A society that does not preserve sources to document its own history is a poor society. History continues to play a pivotal role in our understanding of the present and for our ways of anticipating the future, and if we want to base our history writing on scholarly ground. We need to have as many sources as possible. Today this also includes the online web, just as we have previously collected handwritten documents, print media, film, radio and television.

VS: Web archiving is included in the legal deposit framework in several European countries where Web contents are considered as publications. While newspapers are kept in national libraries, how can we imagine not keeping online news as well? The same applies to the audio-visual sector. But more broadly, the Web is today a medium of expression and communication for political, administrative, academic, economic and social life, for our personal and professional lives. It would no longer make much sense to keep only paper documents. Moreover, this archiving is a treasure trove of billions of pages in all sectors of collective intelligence, but also a reflection of the controversies, crises and challenges of our societies. National archives are not redundant as a result of the Internet Archive, they complement it, refine it nationally and keep track of "national heritage". National repositories also guarantee the sustainability, the legal nature of the framework for access and citation of these sources, and their durability. Libraries and archiving institutions also do a remarkable job related to digital literacy and training. They develop specific tools and are in direct contact with researchers' requirements as a result of collaborative projects.

4. What are the main research approaches that you use to study the archived web?

NB: I have used methods in both ends of the continuum of close and distant reading, to use the terms coined by Franco Moretti[26]. My first web history study focused on one website

[23] Centre national de la recherche scientifique, the French National Centre for Scientific Research.

[24] https://web90.hypotheses.org

[25] https://books.openedition.org/oep/8713

[26] See for example: Moretti, F. (2005) Graphs, maps, trees: abstract models for a literary history. Verso and Moretti, F. (2013) Distant reading. Verso Books.

only, namely the website of the national Danish Public Service Broadcaster (DR), and I based the study on a large number of internal documents, but also archived versions of the website. I used classical historiographical methods such as document analysis and source criticism. In my latest projects, like the one mentioned above about the entire Danish web domain, I have used more quantitative based methods, simply because of the scale of the material. In some sub-projects of that study I have also used network analysis to map the hyperlink network of the Danish web. As is always the case one has to adopt the approaches, methods and tools that are the best fit to help answer the research question one pursues, and to enable the opening up of the available sources.

VS: I have used the web archives both as a research object and as a source for research. With Francesca Musiani, for example, we approached the politics of web archiving and explicitly the question 'Do Web Archives have Politics?' through an STS (Science and Technology Studies) approach, studying web archives as boundary objects[27], their modes of governance, and the multiple agencies and stakeholders that contribute to them.

I also used web archives as a source, for example for my habilitation to direct research and the book *Under Construction. La fabrique française d'Internet et du Web dans les années 1990* (Ina Éditions, 2019), and the *Web90* project already mentioned. There my approach was globally rather qualitative. This did not prevent me from conducting experiments in distant reading, for example on the plethora of born-digital sources of the 2015 attacks in France. I am convinced that a scalable and "medium" reading (a notion I developed, that incorporates both qualitative and quantitative approaches, that is sensitive to the medium (Web, platforms, multimedia and transmedia) and to the context of production) is necessary. As noted by Jane Winters (2017: 239), *"For most humanities scholars it will be a very long time before they transition to using solely digital sources, let alone solely born-digital sources, and for many this will never be the case. They will continue to mix and match, to compare and contrast, and to work with overlapping sets of material which contain subtly different information and are designed for subtly different audiences."*

5. What do you consider to be the most exciting aspects of studying the historical web?

NB: I think one of the driving forces in my academic career has been to enter uncharted territory. In a way this is always what researchers do — if they knew the answers to their questions there was basically no need to do the research — but to me there is a big difference between, on the one hand, doing yet another study of social media use, and, on the other hand, venturing into a field of study and a type of sources where no one has been before. I am definitely intrigued by the latter, despite the fact that you literally have to start from scratch with not much to guide you and also you may feel a bit alone until more people start to see that what you are doing may be interesting. When I started studying the history of the web around 2000, one could count the researchers in the field on one hand, and it was not until 2010-11 that a researcher interest in the archived web emerged and I had someone to play with. But for now, things have improved a lot, and the field of web archive based studies is maturing more and more, in particular with many early career scholars entering the field.

[27] See for example: Nicolini, D., Mengis, J., & Swan, J. (2012). Understanding the role of objects in cross-disciplinary collaboration. Organization science, 23(3), 612-629.

VS: I first started by focusing my thesis on the history of Internet infrastructures[28]. My approach was essentially that of a historian of innovation, engineering cultures and standards. With the history of the Web, it's a more bottom-up approach, a history of digital cultures for the general public. This is how I went from "pipes" to content. I also like the possibility of integrating visual studies and maybe in the future also sound studies. And then, I found with Web archives, a research community, initially modest but growing, full of sharing and friendly spirit, as evidenced, for example, by the biennial RESAW conferences[29], mixing researchers and archivists. I am also passionate about returning to the question of heritage, of the archive, as in the collective work *Qu'est-ce qu'une archive du Web?* and about working in constant interaction with librarians and archivists. The digital hermeneutics dimension, which is one of the core research areas of my laboratory, the C2DH[30], at the University of Luxembourg, also remains for me a field of reflection that is constantly being renewed and of which I never tire.

6. What are currently the most significant challenges for research use of web archives?

NB: There are several, but the need for documentation of what is in web archives, and the need for having content from web archives extracted would be high on my list. As to the first because of the scale and the way web archiving works web archives tend to be black boxes where no one, not even the web archives themselves, know exactly what is in their collection and why. As to the latter we are facing the problem of Research Data Management, that is: who should manage our research data, should it be the web archives or the research institutions? In the first case researchers then have to comply with the research tools offered by the web archives, in the latter case researchers can use the tools and methods that are actually the best fit for their study. Therefore, I think better and more available documentation, and the possibility of being able to extract content from web archives are key to taking the next steps in web archive studies.

VS: Undoubtedly, one of the most important challenges is to enable transnational studies to be conducted smoothly. While the Internet Archive allows access to its collections online, this is not the case for many national collections, especially in Europe, due to the limitations imposed by copyright, legal deposit, etc. Legal deposit, which was, for example, introduced for the Web Archive in 2006 in France, is an opportunity that has made it possible to formalise and define the missions of institutions that collect the Web, but it limits access to the Web Archive. It is necessary to travel physically to consult these archives, for example at the Bibliothèque nationale de France (BnF). Allowing remote transnational access to European Web archives or at least the metadata for example, would be a step forward. The interoperability of data and metadata is obviously also a major issue here. Another challenge is the uneven progress in archiving between European countries; some have an experience of a decade or more, while other countries are only now beginning to address this issue. Finally, there is a major educational challenge, to enable students but also their teachers to fully grasp the potential of the archived Web.

---

[28] Schafer, V. (2007). Des réseaux et des hommes. Les réseaux à communications de paquets, un enjeu pour le monde des télécommunications et de l'informatique françaises (des années 1960 au début des années 1980) (Doctoral dissertation, Paris 4).

[29] https://resaw.eu/events/

[30] Luxembourg Centre for Contemporary and Digital History (C²DH): https://www.c2dh.uni.lu

7. What future challenges in web archiving and/or in studying web archives do you foresee?

NB: In continuation of the above: if web archives are not well-documented and if researchers are not able to get material out of the web archives this may be a challenge for pushing web archive studies further. But also raising awareness about the existence of web archives is important, and that goes for fellow researchers and the wider public. It would be sad if web archives were to experience cutbacks because no one could see their relevance.

VS: Transnational studies remain a major challenge, as I have already said. There is also the capacity to capture audiences on the web and to conduct diachronic studies on different online platforms by being able to relate constantly changing indicators of participation. Another challenge is to develop research uses of web archives from the bachelor's level onwards, or even before. Finally, I also believe that there is also a big effort to be made towards public engagement with web archives. Web archives should not be reserved for researchers only, they are a fertile ground for a much wider public!

8. What future developments on the web do you anticipate and how do you think web archiving institutions can prepare for these changes?

NB: This is a very difficult question to answer. If someone had asked me this question in, say, 2004 I could not have answered *Facebook* or *Twitter*. In other words, it is hard to imagine what will come next, and we therefore tend to see the potential future as a mirror of our present +10%, that is: more of the same. However, what can be said is, first, that the development of the web is very rapid, changes happen all the time, and some of them are very fundamental game changers. Secondly, web archives are always lagging behind when it comes to the archiving of all the new web developments. It took years until web archives were able to collect and preserve *Facebook* and *Twitter* — if this is solved at all in a useful way even today — and there is no reason to believe that this tendency will stop or change.

VS: The historian is not always the best equipped for foresight. But let's give it a try! First of all, I imagine, and I hope, a broadening of audiences and research projects based on Web archives: the privileged period experienced by the pioneers, who were able to benefit from a tailor-made follow-up of their projects and the personalised help of librarians and archivists, will have to evolve. Also, there will surely be new literacies and pedagogical tools to be developed, and other new forms of support. There are lab projects under consideration in some institutions already. Another trend to follow is of course that of closed communities, private gardens, paywalls, and other obstacles to archiving which are constantly developing on part of the Web, and the emergence and rapid disappearance (think of *Vine*[31]) of new digital social networks. This calls for rapid adaptations. In the same way, to be able to seamlessly combine digitised and born-digital sources between institutions seems to me to be an important trend, in addition to the need for transnational studies as well as gateways, ad hoc infrastructures, etc.

9. What are the main ethical and legal issues related to web archiving and the study of the archived web?

[31] https://en.wikipedia.org/wiki/Vine_(service)

NB: In my view there is nothing new under the sun in terms of ethical and legal issues, compared to other media forms. It is also important to keep the two separate. On the one hand, for countries with a web archive based on Legal Deposit legislation they are entitled to archive whatever has been made publicly available on the online web. I still find this a reasonable and good approach to ensuring that we have preserved an important part of our cultural heritage. But, second, the next question is whether one can use the archived material, and here copyright and privacy legal frameworks have to be taken into account, as they should with all other source types as well. But even if we can legally use material from web archives, we may not want to do this, and this is where research ethics comes in. So, the legal frameworks guide what web archives can archive and what researchers can use, whereas ethical considerations guide what we should do, which in some cases may be less than what we can do.

VS: I find Ian Milligan's presentation[32] on this subject very inspiring. I particularly like this reflection: *"I feel similarly uncomfortable with leaving the voices of everyday people completely outside the historical record when there is ample opportunity to include them. Moving to a full opt-in process would likely lead to the historical record being dominated by corporations, celebrities and other powerful people, tech males, and those wanted their public face and history to be seen a particular way"*. There are many ethical and legal issues at stake, from copyright and the possibility of reproducing screenshots from web archives to the application of General Data Protection Regulation (GDPR) and the question of anonymisation. One of the challenges highlighted by the National Forum on Ethics and Archiving the Web and the *Documenting the Now*[33] project, launched in 2016 following the Ferguson protests and riots and the Black Lives Matter movement, is also inclusiveness, not to reproduce existing biases, to take into account issues of gender and cultural diversity for example.

10. Use of web archives is in most cases rather limited. How do you think that the (research) use of web archives could be boosted?

NB: It is important to have some convincing showcases that can illustrate that one could not study phenomenon X if one did not include the archived web as a source. There may be a difference who the users should be: researchers or the wider public? The showcases should probably be different depending on the audience. And also, I think web archives and researchers studying the archived web should try to reach more out to the groups, institutions, or whatever is studied and try to create a community around the study. Personally, I would very much like to investigate the inclusion of the public in the research a bit more, to venture into "citizen science".

VS: There is a growing interest in the research world for web archives. This is not to say that teachers and students necessarily take the next step towards practice. But in my opinion, this is just a matter of a few years. The efforts we make within bachelor and master's programmes to familiarise students with these types of sources, but also initiatives such as datathons for

---

[32] https://ianmilli.wordpress.com/2018/03/27/ethics-and-the-archived-web-presentation-the-ethics-of-studying-geocities/

[33] http://www.docnow.io/

more experienced audiences, seem to me to be a path to pursue. Putting tutorials online too. And of course, the availability of tools such as *The Archives Unleashed Toolkit*[34] developed at the University of Waterloo. The 'Saving the Web' colloquium[35] that KBR organised as part of the PROMISE project, which is the basis for this interview, and which involved a multitude of stakeholders, also seems to me to contribute fully to this need to make Web archives widely known.

11. Do any formal ties exist between the national web archive in your countries of residence and the research institutions you work for? Do your institutions have any (in)formal say in the way the selection policies are shaped? If yes, how is this organised on a practical level?

NB: The pilot project mentioned in the beginning of this interview, where two internet researchers collaborated with the two national libraries, has proven to be a blueprint of how such collaborations can take place. Since this first collaboration in 2000 the internet and web research community at Aarhus University has had a number of close collaborations with the national libraries and Netarkivet, which they run. All of my above-mentioned projects have built on this collaboration. In addition, it is stated in the Danish Legal Deposit law that an editorial committee is established with representatives from the web archive, the researcher communities, and the web content providers. I have been a member of this committee for a number of years, and we have discussed collection policies, access forms, and much more. Also in Denmark we have a national digital research infrastructure called Digital Humanities Lab (DIGHUMLAB, see more on dighumlab.org) in which four universities and the Royal Library participate. The research infrastructure for the study of the archived web, NetLab which I'm heading, is part of DIGHUMLAB. This organisational structure has played a pivotal role in sustaining the web archive and researcher contact over time, in contrast to establishing these contacts only based on time limited funding where a lot of knowledge is lost after the funding ends.

VS: The National Library of Luxembourg (BnL) and the team behind; Yves Maurer and Ben Els who are dedicated to Web archiving, are very dynamic (as their website[36] and the recent "Content at risk"[37] event they co-organised can testify) and focused on exchanges with the academic world. For the past two years (since I arrived at the University of Luxembourg), we have been able to work together on several occasions: Yves and Ben are involved in the Master's winter school that I coordinate to present Luxembourg's Web archives to students and help them to discover the content and tools developed at the BnL. The students also have the chance to have a presentation by Els Breedstraet and her team on the archiving of European institutional sites, which is also conducted in Luxembourg, within the framework of the European Publications Office[38]. In addition, in June 2021, we will organise in partnership between the University of Luxembourg and the BnL a "web archiving week" during which a datathon and two major conferences dedicated to web archives will be held;

[34] https://archivesunleashed.org/aut/
[35] https://www.kbr.be/en/colloquium-saving-the-web/
[36] https://www.webarchive.lu
[37] https://www.science.lu/fr/content-risk
[38] https://op.europa.eu/fr/home

the International Internet Preservation Consortium (IIPC[39]) at the BnL and RESAW[40], a European research group on web archives created and federated by Niels Brügger, at the university. We are therefore working actively together, without playing an active role in the selection policies. However, we feel that the selection policies are very well done, and we know that if we had a suggestion it would be listened to.

12. What would be the research project of your dreams in the context of web archives if funding and organisational limitations were not an issue?

NB: I would very much like to replicate the studies we make of the Danish web domain on a European level. This is, in fact, the aim of the *WARCnet* network, but we only have two years and limited funding, but, who knows, maybe this network project can be a stepping-stone to something bigger.

VS: The project proposal that I submitted last year, but didn't make it! It deals with a diachronic study of viral content since the 1990s. It is likely to contribute fully to the study of European digital cultures, popular cultures on the Web and digital social networks. It will allow comparative studies between European countries, both in terms of their Web archiving strategies and the circulation of viral phenomena. But I continue to believe in this dream, a resubmission of the project has already been made. However, it is still difficult to make the specificities of archived Web research understandable: for example, it must be explained to the evaluators that we cannot provide them with a precise, turnkey corpus at the time of submission, because the creation of corpora is a challenge in Web archives that is fully part of the research process.

13. Do you have any particular advice for institutions wishing to start web archiving, for example which pitfalls to avoid at all cost? Do you have any specific lessons learned from your own experience?

NB: One of the most important lessons learned from the Danish case is that close collaborations between web archiving institutions and research communities are key, and that both parties benefit from this. Web archives will get invaluable information about researchers' interests and needs to help guide the archiving practices and make the collections relevant for future researchers. Researchers will benefit from having the best possible archival copies of important cultural heritage. However, there are also a couple of possible pitfalls to keep in mind. First, it is important that the feedback loop to researchers coversas many different research areas and approaches as possible with a view to making the web archive collections useful for a variety of research topics in the future. Secondly, do not forget to include the costs to these collaborations as a running cost in the budget of the web archive, just as important a budget item as buying hard- and software. If the researcher involvement in web archive activities is only temporary and unsystematic, for instance as part of a funded research project that ends after two years, both web archives and researchers tend to reinvent the wheel every time one research project is followed by a new project. As always when it comes to research infrastructure long term sustainability is pivotal.

---

[39] https://netpreserve.org
[40] https://resaw.eu

VS: I am amazed by what institutions such as BnF and Ina have achieved in France since 2006. They have of course evolved their archiving strategies and their collection and consultation tools over the years to take into account the evolution of the Web, the arrival of social-digital networks, etc. And they have also chosen to work in collaboration with the world of research. It is important to listen to the end users of Web archives. Another piece of advice would be not to get locked into a preconceived vision of what is considered interesting content, often modelled in this case on the world of paper. On the Web, it is not only the online press or institutional sites that count! The culture of Internet users, their forms of expression, in short, an inclusive and non-elitist vision of content is fundamental. Inclusiveness also becomes an important issue: we must not perpetuate gender bias in the Web archives, for example, and we must not reinforce the invisibility of social groups. Countries embarking on web archiving now have the opportunity to build on previous experiences, which can be shared, for example, within the IIPC (International Internet Preservation Consortium). It is an important asset to be able to capitalise and build on these previous experiences.

14. Our next project will investigate the development of a sustainable social media archiving strategy for Belgium. What do you think it is essential for us to consider, to ensure that researchers use this social media archive in the future?

NB: Everything that has been said above also applies for a social media archive: documentation, possibility of extracting material, creating interest and relevance in research communities, and close collaborations with researchers. Belgium is in a fortunate position in the sense that you can take the best from already existing initiatives, without having to make all the mistakes that others have previously made.

VS: Digital social networks are evolving very rapidly (*Vine* and other digital social networks have already disappeared) and with them the need for the responsiveness of archiving institutions. *Instagram* or *Periscope* are less well archived than *Twitte*r for example or *YouTube*. This is due to collection methods (public API, use of *Heritrix*, etc.) and collection choices, but there are also legal limitations on some digital social networks, such as *Facebook* for private content. An interesting point is that of retweets, which are frozen in time at the time of archiving. If the tweet has a lot of retweets after its archiving date, you can't see them. Another challenge is to collect good hashtags on *Twitter*. This requires monitoring, sometimes in real time, which is difficult to predict in advance, as shown by the real-time collections during the attacks of 2015 and 2016 in Europe, even if automated solutions are being tested to more effectively follow the major trends. The citability of these archives is obviously an important point to take into account as well. Another challenge is to be able to cross-reference these results with other collections ("classic" websites, circulation of content between the various digital social media, for example from *Twitter* to *Facebook* and vice versa, and with audio-visual sites that include content linked to digital social networks or, indeed, the comments on them). Finally, documenting these collections in order to point out the choices, limitations, accounts or hashtags included or excluded, etc. is obviously also very important, especially when researchers are conducting quantitative studies on the data for which it is important that they know the representativeness, the gaps and the biases.

Conclusion

Even though the *PROMISE* project[41] formally came to an end in December 2019, the work does not stop there. The recommendations and scenarios to establish a federal strategy for the preservation of the Belgian web need to be implemented. An important step in this direction was the appointment in February 2020 of a permanent scientific assistant for web-archiving at KBR to ensure the continuity of this work. Already, metadata descriptions of over 2,400 websites that were harvested as part of the *PROMISE* web-archiving pilot have been provided as a first level of access to the Belgian web archive in the KBR Catalogue[42]. Belgium has much to learn from researchers such as Niels Brügger and Valérie Schafer, and the advice and guidance that they provided in this interview will be extremely valuable in the months to come.

Regarding future research, the work related to born-digital collections will continue in a follow-up project; *BESOCIAL*, the aim of which is to develop a sustainable social media archiving strategy for Belgium. This 24-month project will take a two-fold approach to social media archiving in Belgium; firstly, investigating the archiving and analysis of the social media channels used by Belgian newspapers included in KBR's collections, and secondly, the archiving of social media content related to events of national and historical importance (e.g. the fire at Notre-Dame in Paris or the terrorist attacks in Brussels). The results of *BESOCIAL* are intended to be a first major step towards implementing a long-term Social Media Archiving strategy for Belgium.

Furthermore, two additional initiatives at KBR will also contribute to the provision of research access to the Belgian web-archive. Firstly, February 2020 saw the start-up of the emerging Digital Research Lab at KBR[43] as part of a long-term collaboration with the Ghent Centre for Digital Humanities (GhentCDH). The aim of the KBR Digital Research Lab is to stimulate the study of Belgium's digitised and born-digital historical, literary and cultural heritage of the 19th - 21st centuries by: a) facilitating data-level access to KBR's digitised and born-digital collections, b) ensuring that the digitised and born-digital collections are embedded into the researcher's workflow in a user-friendly manner and c) optimising the digitised collections for using digital humanities research methods, such as text and data mining. Related to this, the *DATA-KBR-BE* project will facilitate data-level access to KBR Collections for digital humanities research through a new open data platform (data.kbr.be), which will make KBR's available as 'Collections as Data'[44] and ensure that they are compliant with the FAIR (Findable, Accessible, Interoperable and Reusable) principles of research data management.

For both of these initiatives, KBR can draw on international experience in this area, such as the international GLAM (Galleries, Libraries, Archives and Museums) Labs Community[45] and the Special Web Archive Collections at the National Library of Luxembourg[46], in the KB

[41] https://www.kbr.be/en/projects/promise-project/

[42] For example, here is the metadata description of archives website of "Brusselse Bibliotheken", the Dutch-language public libraries in Brussels: https://opac.kbr.be/Library/doc/SYRACUSE/20776949/brusselse-bibliotheken

[43] https://www.kbr.be/en/projects/digital-research-lab/

[44] https://collectionsasdata.github.io

[45] https://glamlabs.io

[46] https://www.webarchive.lu/what-we-have/

Lab at the National Library of the Netherlands[47] and a dataset containing the descriptive metadata of over 2 million websites archived by the Austrian National Library available in ÖNB Labs[48], both of which are particular sources of inspiration. Such 'library labs' are ideal incubators for both increasing access to archived-web resources alongside other digital collections, such as digitised newspapers, and therefore stimulating their take-up and use in the (digital) humanities and social sciences research communities and beyond.

Edited by:
Friedel Geeraert
Web-archiving scientific assistant
friedel.geeraert@kbr.be

Nadège Isbergue
Periodicals manager
nadege.isbergue@kbr.be

KBR
Boulevard de l'empereur 4 – 1000 Brussels
www.kbr.be

Sally Chambers,
Digital Humanities Research Coordinator
sally.chambers@ugent.be
GhentCDH
St. Pietersnieuwstraat 35 – 9000 Ghent
www.ghentcdh.ugent.be

May 2020

[47] https://lab.kb.nl/datasets?f%5B0%5D=field_product_type%3A1
[48] https://labs.onb.ac.at/en/datasets/