Web archives as a data resource for digital scholars

Vlassenroot, Evelyne; Chambers, Sally ; Di Pretoro, Emmanuel; Geeraert, Friedel; Haesendonck, Gerald; Michel, Alejandra; Mechant, Peter

Published in:
International journal of digital humanities

Publication date:
2019

Document Version
Publisher's PDF, also known as Version of record

International Journal of Digital Humanities
https://doi.org/10.1007/s42803-019-00007-7

RESEARCH ARTICLE

Web archives as a data resource for digital scholars

Eveline Vlassenroot[1] · Sally Chambers[2] · Emmanuel Di Pretoro[3] · Friedel Geeraert[4] · Gerald Haesendonck[5] · Alejandra Michel[6] · Peter Mechant[1]

Published online: 08 March 2019

Abstract

The aim of this article is to provide an exploratory analysis of the landscape of web archiving activities in Europe. Our contribution, based on desk research, and complemented with data from interviews with representatives of European heritage institutions, provides a descriptive overview of the state-of-the-art of national web archiving in Europe. It is written for a broad interdisciplinary audience, including cultural heritage professionals, IT specialists and managers, and humanities and social science researchers. The legal, technical and operational aspects of web archiving and the value of web archives as born-digital primary research resources are both explored. In addition to investigating the organisations involved and the scope of their web archiving programmes, the curatorial aspects of the web archiving process, such as selection of web content, the tools used and the provision of access and discovery services are also considered. Furthermore, general policies related to web archiving programmes are analysed. The article concludes by offering four important issues that digital scholars should consider when using web archives as a historical data source. Whilst recognising that this study was limited to a sample of only nine web archives, this article can nevertheless offer some useful insights into the technical, legal, curatorial and policy-related aspects of web archiving. Finally, this paper could function as a stepping stone for more extensive and qualitative research.

Keywords Web archives · Digital scholarship · Curation of digital collections · Copyright · Technology for web archiving

✉ Eveline Vlassenroot
Eveline.Vlassenroot@UGent.be

✉ Sally Chambers
Sally.Chambers@UGent.be

Extended author information available on the last page of the article

1 Setting the scene: Archiving the web as a historical source

The history of web archiving goes back more than 20 years, with the first initiatives launched in 1996 by the Internet Archive, the National Library of Australia and Sweden (Schroeder and Brügger 2017). France was also a pioneer in the field with the National Library of France (BnF) undertaking its first web archiving experiments in 1999 (BnF 2014). However, web archiving has roots in a wider digital preservation movement, which emerged in the 1980s–1990s. Led by memory institutions, the aim of this movement was to develop strategies to respond to the rise of digital technologies and in particular address their ability to capture and preserve digital artefacts as 'records of social phenomena' (Schneider and Foot 2008). As web archiving is still a nascent field, clear definitions are sometimes difficult to find. For this reason, the phrase 'web archiving' is often used interchangeably with 'web preservation,' without any clarification or distinction between the two. For example, the International Internet Preservation Consortium (IIPC)'s definition of web archiving includes both terms: 'Web archiving is the process of collection portions of the World Wide Web, preserving the collections in an archival format, and then serving the archives for access and use' (IIPC 2017). 'Web archiving', therefore, refers to the whole process, whereas 'web preservation' is one of the steps in the process of archiving the web. Web preservation is a crucial step as, in the words of Reyes Ayala, it is 'the process of maintaining internet resources in a condition suitable for use' (2013: 1). A website can be captured and stored, but the preservation of this content ensures it will still be accessible over time. Given this long-term perspective, web archiving requires a strategic approach as much is required in terms of technologies, systems, policies, procedures and resources to make web archiving more than merely harvesting and storing online content.

For digital scholars in the social sciences and humanities, web archives are increasingly recognised as an essential source for studying cultural and social phenomena of recent decades (Schneider & Foot 2005). Some examples include: Brügger et al. (2017), who have been studying the evolution of national domains; Helmond et al. (2017), who used the Internet Archive Wayback Machine for empirically surveying the historical dynamics of social media industry partnerships and partner programmes; Chakraborty and Nanni (2017), who used archived websites as primary sources to examine activities of scientific institutions through the years, or Weber (2017), who traces the tumultuous history of news media on the web through an examination of archived news media content maintained within the Internet Archive. Furthermore, in the BUDDAH (Big UK Domain Data for the Arts and Humanities) project, a number of bursaries were awarded to researchers for carrying out research in their subject area using the UK web archive (BUDDAH 2014). At the European level, RESAW, the Research Infrastructure for the Study of Archived Web Materials, has been established 'with a view to promoting the establishing of a collaborative European research infrastructure for the study of archived web materials' (RESAW 2012).

Legal issues have implications for web archiving as they influence selection policies and users' access to archived online content:

1. Copyright legislation.[1]
2. Personal data protection as web archiving is likely to imply personal data processing. However, it is important to keep in mind that the General Data Protection Regulation (GDPR) authorises legal derogations from the rights of the data subjects when personal data are processed for historical or scientific purposes and for archiving purposes in the public interest.
3. The legal framework on authenticity and integrity of online content as web archives could be used before courts for probative reasons.
4. The issue of illegal contents violating public policy and their potential interest for researchers due to the automatic nature of web archiving tools.
5. Legally delimiting the national scope of competence in a web archiving context with unclear digital boundaries. Indeed, regarding potential overlap between legislation on legal deposit and on public records, it is important to have clear criteria to determine the country and the national heritage institution in charge of the archiving of a particular website.
6. Legislation concerning reuse of public sector information.

As the web has evolved from a publishing to a communication medium, it now presents a vast collection of primary sources for our past. This wealth of diverse information provides the necessary conditions for the emergence of web archiving as a truly interdisciplinary field, bringing together practitioners and scholars from different backgrounds: humanities, social sciences, computer and information sciences, libraries, archives, etc. (Ogden et al. 2017). However, the sheer quantity of information, and the constant evolution of the web, complicate its preservation and make diachronic study for researchers very challenging (Chakraborty and Nanni 2017). As Laursen notes: 'Curators do what they can to capture what they can, and their practices and opportunities change over time' (2017: 220).

The following sections report the findings of a review of web archiving activities in Europe. After a short description of our research methodology, we discuss the aspects of web archiving that affect the users of web archives. First, the web archiving selection process is analysed from an operational point of view, including an in-depth analysis of legal deposit legislation. The different ways in which the concept of a 'national web' is defined and the different selection strategies used by the studied web archiving institutions are explored. Second, the differences in policies regarding access to web archives are analysed, taking into account the legal framework with regard to copyright and the inclusion of illegal content in web archives. On an operational level, the user-friendliness of the studied web archives is explored based on an analysis of the available search functionalities. The role of metadata, and the importance of obtaining a thorough understanding of user needs and requirements are stressed. Third, the 'hands-on' or technical aspects of working with web archives are introduced and some of the challenges and main techniques to keep in mind when working with web archives are discussed. Our explorative analysis of European web archives ends with a discussion underlining four important considerations for digital scholars.

[1] For instance, obtaining prior authorisation of right holders, creating new exceptions for reproduction or communication to the public for archiving purposes and obtaining a fair balance between the public interest in preserving information of cultural or historical significance and the interests of rights holders.

2 Methodology

The research methodology consisted of three phases. In the first phase, a secondary research approach (also known as desk research) was taken. This involved summarising, collating and/or synthesising documentation related to existing web archiving projects. A number of web archiving initiatives were selected and analysed in depth. These included the National Library and National Archive of the Netherlands, the Royal Danish Library (Netarkivet), the National Library of Ireland, the National Library of France (BnF), the National Library of Luxembourg, the British Library, The National Archives UK and Arquivo.pt. in Portugal. With regard to the selection of our sample of web archiving initiatives, a number of characteristics were taken into account:

- Established web archiving initiatives
- Web archiving initiatives in countries where both the national library and the national archives are involved in web archiving (as the PROMISE project is a collaboration between the Belgian Royal Library and State Archives, useful lessons could be drawn from countries where both institutions engage in web archiving)
- Web archiving initiatives in countries with multiple official languages
- Web archiving initiatives in countries of different sizes
- Combination of web archiving initiatives relying on external service providers and initiatives that manage all aspects of the process in-house.

Not all of these features are applicable to each initiative; the main aim was to study a representative mix of web archiving initiatives, based on the above characteristics.

The main research question for this study is: how are other European national libraries and national archives engaging in web archiving and how are the web archiving processes organised? The web archives were studied from a legal, technical and operational point of view. The aim was to create an overview of the web archiving processes in place in each of the institutions covering a) the selection (selection policy, legal framework), b) the web archiving process itself (crawling, quality control, indexation, preservation and storage) and c) access to, and use, of the web archive (policies, search functionalities and legal framework). Operational questions such as the composition of the web archiving teams in terms of professional profiles or the storage requirements in terabytes (TB) or petabytes (PB), were also included in the mix.

In the second research phase, interviews were conducted with representatives from the aforementioned institutions. The aim of the interviews was to fill in the gaps that remained on the specific initiatives following the literature review so that a complete overview of the web archiving activities was obtained for each of the institutions. All participants were interviewed either in face-to-face meetings or by conference call. The interviews were semi-structured, using both closed and open questions. Some interviewees already provided written replies to (some) of these questions beforehand, in which case the interview consisted mainly of follow-up questions. Interviewees included a mix of archivists, librarians, IT specialists, managers, digital curators and researchers (see Appendix A).

The third and final research phase encompassed further validation and synthesis. The answers to the questions that were obtained during the literature review, and in the interviews, were integrated. On the basis of which, comparisons were drawn, thereby obtaining an answer to the research questions and creating an overarching view of the selected web archiving initiatives. This allowed to us to distil the relevant aspects that are important for digital scholars.

3 Selection of content for web archives

3.1 How is web archiving framed by the law?

In all of the countries where our selected European web archiving institutions are based, the National Library is legally responsible for preserving and opening up cultural and historical heritage to the public, even if they have no legal deposit law (e.g. The Netherlands). There is a lot of information available online; thus, institutions believe that the preservation of online cultural heritage is naturally part of their legal mandate. In addition to the mandate to preserve a nation's heritage, there are two legal ways to enable web archiving: on the one hand, legal deposit legislation; on the other hand, legislation on public archives.

The majority of countries have gradually modified their national legal deposit legislation in order to widen it to the Internet and thus allow the collection and preservation of online information.[2] In Ireland, this process is ongoing as the legal deposit legislation is now under review to broaden its scope to include online contents (Ryan 2017). As Maria Ryan (2017) stated: 'The Irish situation is difficult because Irish Legal Deposit legislation does not extend to digital or online publications. The legislation is under review at the moment'.

The scope of this legislation is often very broad in regard to determining which websites should be archived. However, national legislation generally excludes personal correspondence and private spaces available on intranets, for privacy reasons.

Still, a minority of countries do not have any legal texts relating to legal deposit (at least, to the web legal deposit).[3] In these countries, the deposit of websites of cultural and/or historical significance to the National Library is in principle done on a voluntary basis (Beunen and Schiphof 2006, p. 18; Kunze and Power n.d., p. 2). Indeed, in the absence of a legal obligation to deposit publications, the consent of website owners is necessary.[4] These right holders are, therefore, able to refuse web archiving.

In the Web 2.0 world, obtaining the prior consent of each right holder is impracticable, especially since their identification can be very difficult. Therefore, heritage

[2] It is the case for France with the DADVSI Law (see « Loi n° 2006–961 du 1er août 2006 relative au droit d'auteur et aux droits voisins dans la société de l'information »), for Luxembourg (see « Loi luxembourgeoise du 25 juin 2004 portant réorganisation des instituts culturels de l'Etat »), for United Kingdom (see « Legal Deposit Libraries (Non-Print Works) Regulations of 5th April 2013 »), For Denmark (see "Danish Act n° 1439 on Legal Deposit of Published Material of 22nd December 2004").

[3] For instance, The Netherlands, Portugal and Switzerland (at the federal level).

[4] Prior authorization of the right holders is not necessary for websites that have fallen into the public domain or that were made available under the system of Creative Commons License (Beunen and Schiphof 2006, p. 16).

institutions acting in countries that do not have legislation for the web legal deposit do not always ask the permission from the websites owners before proceeding to collect their website, preferring to take a pragmatic approach. On the one hand, they either notify the website owner of their intention to archive their website and if he/she does not object, they consider that the website owner implicitly consents to the archiving.[5] As Kees Teszelszky states:

> The biggest problem for web archiving in the Netherlands and for our national library is that we do not have a legal deposit like you have in Belgium. [...] So then we decided [...] to [use] the opt out method. So if we want to archive the website, [...] we do not ask permission, we say we are going to archive and if people are not reacting on our wish, then we are archiving. (Teszelszky 2017a)

On the other hand, they either choose to archive all websites included in their selection policy without prior notification, but allow the website owner to object to the archiving by using Robot Exclusion Protocols.[6] In any case, these heritage institutions are generally very cautious. In this way, they develop a very effective takedown policy in the event of subsequent objections by website owners, through the removal of the archived content from their database.

There are a number of advantages for heritage institutions of relying on legal provisions that enable them to frame their web archiving activities in order to solve the aforementioned difficulties. Firstly, legislation on web legal deposit has the advantage of offering greater legal certainty and facilitating the web archiving by forcing the website owners to comply with the legal deposit obligation. Indeed, such legislation means that heritage institutions are not required to ask for prior permission from website owners (Beunen and Schiphof 2006). Without that legislation, the owners' consent would be required, because the archiving of a website composed of various protected contents[7] necessarily triggers an act of reproduction,[8] likely to infringe copyright. Alongside a web legal deposit obligation, some countries created some copyright exceptions covering activities intrinsically linked to web archiving and access.[9] It is, in fact, technically impossible to archive a website without reproducing it. In this way, these kind of exceptions have proved unavoidable in order to permit acts of reproduction (Graff and Sepetjan 2011: 179–180). Secondly, some countries have a legal provision allowing the heritage institution responsible for web archiving to require

[5] This approach is the one of the National Library of The Netherlands (KB Nederland, n.d.-b and n.d.-d).

[6] This approach is the one of Arquivo.pt. in Portugal (Arquivo.pt, n.d.-c).

[7] Let us indicate that websites are composed of a set of elements that can be each protected by copyright (original texts, images, search engine, database, etc.) and may each have a different right holder (KB Nederland n.d.-e). We also have to underline the fact that websites can also be composed of elements protected by other rights such as trademark law, database right, neighboring rights and image right (KB Nederland n.d.-b).

[8] Act for which the consent of the right holders is in principle required.

[9] In France, the DADVSI Law has introduced an exception allowing acts of reproduction and communication related to the web legal deposit (see French Heritage Code, art. L132–4 to L132–6). In the United Kingdom, Sections 19 to 31 of the Legal Deposit Libraries (Non-Print Works) Regulations of 5th April 2013 and Section 44A of the Copyright, Designs and Patents Act of 15th November 1988 allow the realization of certain activities related to web legal deposit without that they violate copyright.

domain names management bodies to help them identify website owners.[10] Thirdly, some legislations go even further by allowing heritage institutions to require website owners to give the passwords and access keys necessary for collecting their website.[11] This makes it considerably easier for the heritage institutions to obtain the web material covered by legal deposit.

Concerning the criteria for deciding the scope of the national web archive at the national level, we noticed some similarities in the choices made by the studied countries with legal deposit legislation. Considering that online information falls within the scope of competence of one state or another, there are three main principles to be followed. Firstly, a state considers itself competent to archive online contents published within its national domain name. Secondly, a state also considers itself competent to archive online contents published on other domain names if one of these additional conditions is met: if the website was registered to the national body responsible for managing domain names or by a citizen of the state; if the content of the website is related to the state (i.e. concerns the general affairs of the state); if the content of the website was drafted by a citizen of the state or in the national territory. Luxembourg also has an additional criterion which was not found elsewhere: if the production of the publication has been supported by the state. Thirdly, the language of the content is an additional criterion. However, this criterion only applies to countries with a single national language but does not work for countries with multiple national languages that are also national languages of other countries.

In countries without legal deposit legislation for online content, the scope is defined in a similar way. In Ireland, in addition to the national top level domain, web material that is of Irish interest, has heritage value and that treats a subject of interest, is also considered to be within scope (National Library of Ireland 2017a). In Portugal, the top level domain of all Portuguese speaking domains, except for the Brazilian domain, are included, as are websites on other domain names that are of broad interest to the Portuguese community. While in The Netherlands, websites about Dutch language, history and culture on both the national domain, and other domain names, are within the scope of the project (Arquivo.pt n.d.-c; Sierman and Teszelszky 2017).

There is a marked difference between public record legislation that regulates the activities of national archives and legal deposit legislation that frames the missions of national libraries. Where web legal deposit legislation exists (UK, France, Denmark, …) numerous detailed provisions are included to frame web archiving activities. For instance, the text of the UK legal deposit legislation[12] comprises of more than 20 legal provisions, specifically related to web legal deposit. However, in public records legislation, the same text that applies to classic public records also applies to online public records, meaning that there are no specific legal provisions where web archiving is concerned, except for in the Library and Archives of Canada Act. The legal text on

[10] For instance, in France, Article L132–2-1 of the French Heritage Code authorize the "Bibliothèque Nationale de France" to turn to domain names management bodies or to the Higher Audiovisual Council to identify the publishers and producers of websites. There is also a similar legal provision in Denmark (See Danish Act n° 1439 on Legal Deposit of Published Material of 22nd December 2004, §11).

[11] It is the case in France (see French Heritage Code, art. R132–23-1, II), United Kingdom (see Legal Deposit Libraries (Non-Print Works) Regulations of 5th April 2013, Section 16 (4)) and Denmark (see Danish Act n° 1439 on Legal Deposit of Published Material of 22nd December 2004, §10).

[12] The Legal Deposit Libraries (Non-Print Works) Regulations 2013

public records therefore applies to websites of public institutions only because the notion of "records" is broadly defined (for instance, as 'all types of medium').

3.2 How is web archived content selected?

Our analysis shows a great deal of variation when it comes to selection strategies and criteria. Furthermore, the terminology for describing the web archiving approach, differs between web archiving initiatives. As can be seen in Table 1, in the case of Arquivo.pt., two main strategies can be distinguished: broad crawls (covering top-level domain crawls (e.g. .be, .fr) and relevant content outside of the national domain(s)) and selective crawls (thematic or events-based collections, for example). The selection policy of national archives with regard to web archiving differs in the sense that it is mostly limited to the public records of governmental organisations. For national libraries, the scope of collection is broader as web archiving is seen as part of the legal deposit legislation or as a complement to the more traditional electronic or paper collections of publications in countries without legal deposit legislation.

All national libraries and Arquivo.pt. in Portugal combine broad crawls with selective crawls, except for the National Library of France (BnF) where a representative sample of the web is taken instead of a complete top level domain crawl and the National Library of The Netherlands, where only a selective approach is taken Table 1.[13]

Different methods are used to identify the content that does not reside under URLs of the national domain. The British Library, for example, uses Geo-IP localisation to locate information on servers in the UK or make use of UK postal addresses (Hockx-Yu 2014). At the Royal Danish Library a specific system has been developed to identify this content. As Jakob Moesgaard explained:

> We've built a system that basically looks at everything we harvest. It looks at all the links that point out [...] and then it analyses the content on all of those pages. [...] It scans for regular expressions that cover Danish phone numbers and [...] we try to have this sort of validation process ranking [...] to see if [...] it looks Danish enough for us to trust that we should automatically add it to the archive. (Moesgaard and Larsen 2017a, b)

In the case of selective crawls, there are different ways of determining these collections. Some institutions have defined overarching selection criteria for these collections. The British Library, for example, focuses on websites that publish research, reflect the diversity of lives, interests and activities in the UK and demonstrate web innovation for the UK Web Archive (UK Web Archive n.d.-a). In general, websites deemed of interest to the nation are included in the selection, meaning websites that are representative of the diverse society, or that are linked to the history and culture of a nation. It is interesting to note that the popularity, uniqueness or the degree of innovation of websites is sometimes also taken into account, as well as websites that publish research (KB Nederland n.d.-a; National Library of Ireland 2017a; Maurer and Els 2017a; Gomes 2017b).

[13] Sierman and Teszelszky 2017; BnF 2017a, b; Maurer and Els 2017b; UK Web Archive (n.d.-a); Hockx-Yu 2014; Brügger et al. 2017; Arquivo.pt n.d.-c; Ryan 2017; National Library of Ireland 2017a, b.

Table 1 Overview of general selection strategies for web content

Country	Institution	Broad crawl	Selective crawl
The Netherlands	National Library	No	Yes
France	National Library	No (Representative sample)	Yes
Luxembourg	National Library	Yes	Yes
UK	British Library	Yes (non-print legal deposit)	Yes (open UK web archive)
Denmark	Royal Danish Library	Yes	Yes
Portugal	Foundation for Science and Technology (FCT) (Arquivo.pt)	Yes	Yes
Ireland	National Library	Yes	Yes

Another way to create selective collections is to build them based on specific themes, events or even emergencies (mostly focusing on natural disasters or other unforeseeable events). There is a large variety in how thematic collections are defined. They can, for example, be centred around the different collection departments in the institution, as is the case in the National Library of France (BnF 2017a) or focus on other themes such as literary collections or health and social issues amongst others, which is the case in the National Library of Ireland (National Library of Ireland (n.d.-d)). Event-based collections on the other hand are more coherent between institutions. Most often they are about events such as elections (national or local), commemorations, referendums or sporting events such as the Olympics.

With regard to social media, a number of web archiving initiatives include them in their collections. From a technical point of view, archiving social media is challenging (e.g. due to the vast amount of data generated or changing access policies), which explains why increasingly sophisticated proprietary and open source software and services are available to support social media archiving. The policies with regard to social media differ widely between institutions. Table 2 provides an overview of which institution preserves which social media.[14] The most popular social media platforms captured by the studied web archiving initiatives are Twitter, YouTube and Facebook. The social media accounts that are captured, in general focus on important people, organisations and events such as political parties, politicians, newspapers, journalists, athletes, other celebrities, etc. In the case of Arquivo.pt. no special efforts are made to harvest social media, although their web archive does contain some material stemming from Facebook and Twitter (Gomes 2017b). The National Library of the Netherlands is also not currently harvesting social media, but they have it included in their 10-year plan. At the National Archive of The Netherlands social media are not yet included in their collection either, but tests have been scheduled in 2018 to archive social media (Teszelszky 2017a; Posthumus and van Luin 2017a).

[14] Tanésie et al. 2017; Maurer and Els 2017b; British Library 2017a; British Library (n.d.-b); National Archives (n.d.-a); Netarkivet.dk 2017; Moesgaard and Larsen 2017a

Table 2 Overview of social media included in web archives

Country	Institution	Facebook	Twitter	YouTube	Instagram	Flickr
France	National Library	(used to, not anymore)	Yes	No	No	No
Luxembourg	National Library	Yes	Yes	Yes	Yes	No
UK	British Library	Yes	Yes	No	No	No
UK	National Archives	No	Yes	Yes	No	No
Denmark	Royal Danish Library	Yes	Yes	Yes	Yes	No
Ireland	National Library	No	Yes	Yes	No	(starting in 2018)

Some institutions also make use of certain exclusion criteria, some of which concern the legality of the content. The national legislations are unanimous on what constitutes illegal content: child pornography, hate, xenophobic or racist speech, speech inciting to violence, etc. Some institutions take specific measures to exclude this content automatically. The National Library of France, for example, makes use of a filtering tool (Tanésie et al. 2017). Additional exclusion criteria are sometimes in place, for instance, excluding content that is already included in other web archives or material that cannot be captured for technical reasons (KB Nederland (n.d.-d); Moesgaard and Larsen 2017a). In the case of The National Archives UK, additional selection criteria have been developed for Twitter content, for example, tweets written by the selected government organisations are included, but retweets or tweets sent from non-governmental accounts to government accounts are excluded (National Archives (n.d.-a)).

When digital scholars use web archives for their research, it is important that they take into account how the archived web content is selected and who is responsible for making that selection. In some institutions specific collection specialists are responsible for making the selection, while in other cases, selection is a responsibility that is shared between a large number of people, each devoting only a limited amount of time to selecting the content. This is, for instance, the case at the National Library of France (BnF 2016) where the selection is done transversally, meaning that each department contributes to the web archiving by entering URLs into the system (Tanésie et al. 2017).

Furthermore, some institutions collaborate with external partners. The National Library of Ireland sometimes contacts specialists in the field. For their collection on the Irish elections, they contacted political analysts, lecturers and journalists in order to obtain their feedback on what should be included in the collection (Ryan 2017).

The role of digital scholars, along with the general public, in the selection of content for web archives is a topic worthy of consideration. For example, engagement from the digital scholars as well as the general public is already being sought: the national libraries of France, The Netherlands, Luxembourg, Denmark and Ireland, and Arquivo.pt., all provide a way for people to make suggestions for websites to be included in the selection (BnF 2017c; KB (n.d.-d); BnL n.d.; Netarkivet.dk 2016a; Ryan 2017; Arquivo.pt (n.d.-e)).

As 'all web archives to a greater or lesser degree can only suggest comprehensiveness', (Koerbin 2017: 194) web archiving institutions have a very important role to play as facilitator. They should ensure that sufficient information about the web archiving

context is made available so that researchers can find the answers to the questions evoked by Webster (2017: 175–176): 'Why has this content been archived, by whom and on whose behalf?' There is a clear demand for this information. Sara Aubry of the National Library of France (BnF) stated:

> This is information researchers increasingly request meaning that they wish to understand the context of the production of the archive in order to gain insight into whether [a resource] was archived as part of a selective crawl or of a broad crawl, if it was part of a specific project, how long the crawl lasted, [...], so really everything about the context of the capture. (Tanésie et al. 2017, translated from French)

However, even though the importance of this contextual information is understood, it is sometimes not made available. From a research perspective, this lack of contextual information is problematic.

Finally, the web archiving process itself has an impact on what digital scholars can do with the material:

> The purpose, strategies and technology of an archive affect what is archived and the manner in which it can be accessed, and in this way influence the possibility of constructing a research object on the basis of the material in the archive. (Nielsen 2016)

It is important that digital scholars keep these various aspects in mind, when they undertake their research using data from web archives.

4 Consultation, access and ease of use of web archives

4.1 How to consult and access web archived content?

It is essential to underline that access conditions differ widely between web archives as can be seen in Table 3. Some of the web archives are freely accessible online such as Arquivo.pt. in Portugal or the web archive developed by the National Library of Ireland[15] (Arquivo.pt n.d.-a; National Library of Ireland n.d.-a). For the national libraries, this mission of making national heritage accessible to the public is complementary to their national heritage preservation mandate. However, granting such access to the public must comply with the legal provisions related to copyright. Indeed, the vast majority of archived online content is protected by copyright and, while it is clear that their mere archiving is not likely to cause too much damage to right holders, this is not the case when making this content available to the public (Beunen and Schiphof 2006).

As a result, in a number of web archives, only specific parts of the collections are freely accessible. In the case of the British Library, the Open UK web archive and the

[15] In the case of the National Library of Ireland, this only counts for the web archive collections that were based on a selective policy. Access conditions to the web material collected during the top-level domain crawl that started in 2017 were not yet defined at the time of the interview.

JISC UK web domain dataset for example are freely accessible, whereas the UK non-print legal deposit web archive is not (UK Web Archive (n.d.-b); British Library (n.d.-b)). At the National Archive of the Netherlands, a specific status for access is assigned to each archived website: open, restricted or offline (Posthumus and van Luin 2017b). Some web archives, which are not freely available, are only accessible on the premises of the library from specific workstations. In the case of the UK non-print legal deposit web archives, the law also specifies that only one user can access a certain piece of online content at any given time.[16] A reader card needs to be obtained in some cases to gain access to the reading rooms as is the case in the National Library of The Netherlands (KB Nederland (n.d.-e)). At the National Library of France (BnF)[17] however, the legislation is more flexible: accredited users are allowed to bring their own laptop to connect to the network. At the Royal Danish Library remote access is provided for PhD-level researchers (Moesgaard and Larsen 2017b). Some web archives are also only open to researchers and others are not accessible at all, as is the case for the web archive of the National Library of Luxembourg where the technical infrastructure is not yet in place (Maurer and Els 2017a). However, in most cases, the access restrictions are in place because of copyright reasons. As Webster states: 'A common feature of most web archiving backed by legal deposit legislation is some sort of restrictions on the access afforded to the end user of the archive' (Webster 2017: p. 180) Table 3.

The British Library found a way to avoid certain access restrictions with their interface SHINE of which the beta version was launched in December 2017 (UK Web Archive (n.d.-d)). Their archive is open to anyone, but for content that is not publicly available, only the metadata is shown (Webber 2017). Other web archiving specialists showed interest in the SHINE interface, Yves Maurer from the National Library of Luxembourg stated that:

> The SHINE interface of the UK British Library would be very useful for digital humanities researchers, for sociologists, political scientists maybe or even journalists. (Maurer and Els 2017a)

In the context of access to web archives, it is important to keep in mind the interests of rights holders. Table 4 provides an overview of how the studied institutions allow web archives to be used. Some countries are keen to put in place a fair balance between the interests of website owners and the interest of the public to access archived online content. Indeed, some heritage institutions respect a kind of 'embargo' on access (meaning that content can only be made accessible to the public at the end of a certain period) upon a duly justified request of right holders. For instance, in the United Kingdom right holders have the opportunity to submit a written request to the deposit library to prevent readers' access for a renewable period of three years in order to protect their commercial interests. The British Library grants this 'embargo' request if it considers that providing access to readers during the specified period would unreasonably prejudice the interests of right holders (see Legal Deposit Libraries (Non-Print Works) Regulations of 5th April 2013, Section 25). Arquivo.pt. in Portugal also makes use of an automatic 'embargo' for all online publications. They are attentive to the interests and rights of authors by respecting an access

[16] See Legal Deposit Libraries (Non-Print Works) Regulation of 5th April 2013, Section 23.
[17] See French Heritage Code, art. R132–23-2.

Table 3 Overview of access methods to the web archives

Country	Institution	Access method		Who has access?
		Open & freely accessible online	Physical access on location	
The Netherlands	National Library	No	Yes	Everyone with a paid library card. Big data researchers can gain access after a meeting and having signed a contract.
The Netherlands	National Archive	Yes (for websites with an 'open' status)	Yes (for websites with a 'restricted' or 'offline' status)	'Open' & 'offline' status websites: everybody. Some items are 'restricted', which means you need a special permission (a research proposal is required to obtain this permission or proof that the subject of the archived content is dead). Together with the special permission a signed form is needed stating you understand your own responsibilities under the privacy-law.
France	National Library	No	Yes (but also from within the 26 partner libraries)	Authorized users of the BnF (18 years or older and for university studies, professional or personal research. For the latter two categories, interviews are conducted before accreditation is given.)
Luxembourg	National Library	No	No	No public system yet.
UK	British Library	Yes (for the UK web archive)	Yes (for the legal deposit UK web archive and JISC domain dataset)	Everyone with a reader's pass.
UK	National Archives	Yes	No	Everyone
Denmark	Royal Danish Library	Yes (only for researchers conducting research on a Ph.D-level or above)	Yes (only for researchers)	Only for research purposes after filling an application form that needs to be evaluated.
Portugal	Foundation for Science and Technology	Yes	No	Everyone
Ireland	National Library	Yes	No	Everyone

The information included in this table can be found in: KB Nederland (n.d.-e); Posthumus and van Luin 2017b; BnF 2017c; Maurer and Els 2017b; UK Web Archive (n.d.-c); British Library (n.d.-a); Webber 2017; National Archives (n.d.-b); Moesgaard and Larsen 2017b; Arquivo.pt (n.d.-b); National Library of Ireland 2017a, b

embargo period of one year after the collection of the website to avoid that the archived content competes with the online website (Gomes 2017a) Table 4.

Finally, web archives raise the question of how to proceed in relation to illegal content. Since most of web archiving procedures are automatic, it is inevitable that sometimes so-called 'illegal' content is collected. This was also noted by the National Library of France (BnF) where Sara Aubry stated:

> We will not collect them, we will not take active steps to collect [illegal content] in the context of selective crawls [thematic or events-based collections, for example]. In the broad crawls [covering the capture of a representative sample of the French web], however, we will not refrain from collecting them. (Tanésie et al. 2017, translated from French)

Table 4 Overview of allowed use of the web archives

Country	Institution	Functionalities
The Netherlands	National Library	Copy only for themselves.
The Netherlands	National Archive	Not specified.
France	National Library	Short quotations and screenshots only for teaching and research. Forbidden to download archived files and other technical restrictions may prevent copying of texts or screenshots.
Luxembourg	National Library	No functionalities as no access is currently provided.
UK	British Library	Printing of material in the legal deposit UK web archive is allowed, but very limited.
UK	National Archives	Most Crown copyright material within the Web Archive can be used without formal permission under the terms of the Open Government Licence. Where the copyright of material is owned by a third-party, it is the responsibility of the user to obtain the necessary permission for re-use.
Denmark	Royal Danish Library	Possible to make a copy of the website for personal use, display the archive or websites from the archive for teaching (non-public classes or courses). Use in public, scientific and television presentations and for scientific publications is also possible but with certain restrictions.
Portugal	Foundation for Science and Technology (FCT) (Arquivo.pt)	Access is intended to support work of an educational, scientific or research nature. Use for commercial purposes is strictly forbidden.
Ireland	National Library	Available for the purposes of research and private study only. For publication the permission is needed from the National Library. When copyright exists and is not held by the National Library, the copyright holder's permission is also needed.

The information included in this table can be found in: KB Nederland (n.d.-e); BnF (2017c); Maurer and Els 2017b; UK Web Archive (n.d.-c); National Archives (n.d.-a); Netarkivet.dk 2016b; Arquivo.pt (n.d.-d); National Library of Ireland (n.d.-b); National Library of Ireland (n.d.-c)

If web archives contain illegal content, heritage institutions usually ensure that these archived web pages are not made accessible to the public. Nevertheless, such contents might be of interest for digital scholars and researchers in certain disciplines to understand and analyse the history and the culture of the country. To paraphrase Valérie Schafer; having, for example, access to past Neo-Nazi websites, which, in fact, contain hate speech, is of utmost importance, both for the study of digital cultures and of history in general (Tanésie and Aubry 2017).

4.2 What makes a web archive easy to use?

Once users have obtained access to a web archive, archived websites are often not easily discoverable via the available search and browse methods (see Table 5). This inhibits use (Dooley 2016). Two main challenges were revealed to ensure discoverability in the context of a web archive; the lack of descriptive metadata guidelines and the lack of a clear understanding of user needs and behaviour (Dooley et al. 2017). It is necessary to address these two challenges in order to guarantee the discoverability of web archives.

The lack of descriptive metadata guidelines related to web archiving is also problematic for initiatives where the aim is to link different web archives, as is the case for the National Coalition for Digital Preservation (NCDD) in The Netherlands. The NCDD is working on promoting cooperation and creating an inventory of which material is present in which web archive (NCDD n.d.). Related to this initiative, Teszelszky (2017a) said: 'If we want to have a national web collection, we need to use the same software. We need to have common standards and that's something that

Table 5 Overview of search options in the web archives

Country	Institution	Search options			
		URL	Full-text	Topical browsing	Alphabetic browsing
The Netherlands	National Library	Yes	No	No	No
The Netherlands	National Archive	No	No	No	No
France	National Library	Yes	Yes	Yes	No
Luxembourg	National Library	Not open for the public yet.	Not open for the public yet.	Not open for the public yet.	Not open for the public yet.
UK	British Library	Yes	Yes	Yes	No
UK	National Archives	Yes	Yes	No	Yes
Denmark	Royal Danish Library	Yes	Yes	No	No
Portugal	Foundation for Science and Technology	Yes	Yes	No	No
Ireland	National Library	Yes	Yes	No	Yes

The information included in this table can be found in: Teszelszky 2017b; Posthumus and van Luin 2017b; BnF (2017c); Maurer and Els 2017b; UK Web Archive (n.d.-b); The National Archives n.d.; Gomes 2017b; National Library of Ireland (n.d.-a)

will be worked on'. Increasing standardisation of metadata management would, therefore, be advantageous for the users.

The second most frequently mentioned challenge is the need for a better understanding of user needs and behaviour to ensure discoverability for archived websites (Costa and Silva 2010; Dougherty et al. 2010). Many web archiving institutions do not have accurate statistics on the number of visitors of their web archive. Often the number of visitors to the web archive are merged with the number of visitors of the whole website (as is the case at the National Archive of the Netherlands) or in other cases the internal use of the staff was included (as is the case at the National Library of the Netherlands). Furthermore, numbers like these do not indicate who these visitors are; why they are visiting; what they expect to find; what they take away with them and whether they experienced any degree of satisfaction. As Maria Ryan (2017) of the National Library of Ireland stated: 'It's difficult to get good analytics on web archive users, due to the fact the selective web archive can be accessed remotely'. In the case of Arquivo.pt., efforts are made to target the right people to stimulate them to make use of the web archive. In this regard, they have a well-defined communication strategy in place to encourage researchers and academia to use their collections. For example, they organise contests offering prize money to researchers working with their collections (Gomes 2017b). That user engagement is also considered an important matter at the British Library is underscored by the fact that they have a 'Web Archiving Engagement Manager' for the web archive (British Library 2017b). This contrasts with other web archiving initiatives that find it hard to attract users:

> Not many people are using our web archive. I think we have 100 visitors a year [...] We only see this year that these kind of researchers come to our web archive because some websites are not in the Internet Archive. (Teszelszky 2017a)

In general, most interfaces of web archives afford a form of URL search (either searching for an exact URL or a specific part of a URL), combined with full-text searches. The URL approach has been dominant for years (Ben-David and Huurdeman 2014) but, recently, full-text search is also supported by most of the web archives. Research by Costa and Silva (2010) shows that users prefer full-text search to URL search. However, some web archives have also permitted other types of searches for some time now. In such web archives, the user can also explore topical collections or undertake alphabetical browsing (see Table 5).

5 Overview of tools used in web archiving

This section briefly describes web archives from a technical viewpoint. In particular, it discusses software tools involved in the process of gathering web content and analysing this content that might be relevant for digital scholars. Not all available tools are described, however, nor are the long-term preservation systems or the back-ends of archives.

Web archiving starts with harvesting or crawling websites, which means trying to get a copy of websites. Since web content is diverse—static pages, dynamic pages, multimedia, social media, etc.— different harvesting tools focus on different types of

content. Typically they produce output that can be stored or archived, for instance, as a directory structure on disk, mimicking the original website or as Web Archive (WARC) files (ISO 2017).

HTTrack (Roche 2018) copies the website(s) to disk so the user can simply open it in a browser. It uses a single thread so one instance is only suited for limited crawls. Webrecorder (Webrecorder n.d.) uses a browser to harvest content of websites, hereby addressing typical issues of other harvesting tools: dynamic content, flash, multimedia, etc. It 'records' web pages as the user browses them, so it is suited for very selective, high quality crawling. Although it requires some technical skills to install, an online demo is available. The content is saved in the WARC format.

Wget (Free Software Foundation 2017) and the similar tool Wpull (Foo 2016) are versatile command line tools that have built-in web crawling functionality, comparable to HTTrack. They can write to a directory structure or to WARC files. Wpull is better suited for large crawls because it stores detected URLs to disk as opposed to WGet which stores them in often limited computer memory, and it offers deduplication (i.e. crawls a page only once). Both tools are rather easy to install and to run; the art is to compose the right commands to instruct them. Grab-site (Grab-site GitHub 2018) provides a graphical interface for Wpull.

Social media require specialised tools to capture their content because of their very dynamic nature. Capturing content is typically done programmatically using Application Programming Interfaces or APIs, offered by the social media providers. F(b)arc (Fbarc GitHub 2018) is a command line tool that can be used to archive data using the Facebook Graph interface. Twarc (Twarc GitHub 2018) is a command line tool and library that makes using the Twitter APIs easy. It can be used to archive data, detect trends, search friends, etc. Social Feed Manager (Social Feed Manager 2018) can harvest data from Twitter, Tumblr, Flickr, and Sina Weibo.

Web archiving organisations tend to use more advanced tools, which often require technical skills to install and use. Heritrix (Webarchive.jira.com, July 2016) is a general purpose web crawler designed with web archiving in mind. It can be configured for broad crawls or targeted crawls, on one machine or in clusters, it can be extended with custom code, etc. It is suited for large scale crawling activities, but less so for dynamic pages or social media. It produces WARC files. The NetarchiveSuite (Rosenthal 2017) is built with Heritrix at its core, but provides extra functionality in the area of deployment, long term preservation and access. Brozzler (Brozzler GitHub 2018) uses the engine of the Chrome browser to harvest pages, which offers the same advantages Webrecorder offers, but it requires no user interaction during crawling. It can be set up on a cluster.

Besides tools to get the data, there are also tools for doing something with the data. Tools to view the archived websites include Webrecorder Player (Webrecorder Player for Desktop GitHub 2018), OpenWayback (IIPC 2018), pywb (Pywb GitHub 2018) and WAIL (Web Archiving Integration Layer) (Kelly 2017). Webrecorder Player is relatively easy to install and use and can open content from ARC, WARC and HAR (http Archive) files. OpenWayback reads and indexes WARC files and lets users browse or search the archived content in a web browser. Pywb offers OpenWayback functionality, but it also enables web pages to be recorded while the user surfs the web. It is the software used in Webrecorder and Webrecorder Player. Note that OpenWayback and pywb require technical skills to set up. WAIL is an easy-to-use

tool with a graphical user interface that combines Heritrix for capturing websites and OpenWayback for viewing the captured content.

Several tools and libraries exist to *enable* processing archived data, but don't *do* a lot of actual processing. Tools that can read and write data, or validate and extract metadata from WARC files include JWAT (Clarke 2016), node-warc (Node-warc GitHub 2018), WARCAT (WARCAT GitHub 2017) (Web ARChive (WARC) Archiving Tool), warcio (Warcio GitHub 2017) and warctools (Warctools GitHub 2016). These tools often require programming skills to write software that processes the data itself.

Some tools go a step further and provide a framework for analysing web archives. The Archives Unleashed Toolkit (AUT), part of the Archives Unleashed Project (Archives Unleashed Project 2018), provides a flexible data model for storing and managing raw content as well as metadata and extracted knowledge. Although basic programming or scripting skills are required, a lot of built-in functions (including, extracting links, popular images, and named entity extraction) help the writing of powerful code. A version running in the cloud, providing a user interface, is currently being developed. A tool similar to AUT is ArchiveSpark (ArchiveSpark GitHub 2018). This tool focuses somewhat more on entity recognition and linking than AUT. Another difference is that ArchiveSpark extensively uses CDX files, which are indexes generated from WARC files to speed up some processing. Both tools are built using the Apache Spark analytics engine, enabling a plethora of (big) data processing and analysis tools on top of their own functionality.

A last aspect worth mentioning is how to access publicly available archived data from organisations. As described before, most organisations make this data accessible by means of a web page. However, there is a standardized way of getting web resources near a given timestamp, with a specific URL: Memento (Van de Sompel et al. 2013). It is not necessary to know which organisation holds the data, as long as it runs a Memento aware web service. Organisations supporting Memento are Arquivo.pt., National Library of Ireland, UK Government Web Archive, UK Web Archive, Internet Archive and many more (Kremer 2016). OpenWayback and pywb for instance are tools that provide Memento functionality. A number of Memento clients exist, as standalone libraries or browser plugins, which can be used to access data in this way. A demo is available online.[18]

6 Discussion and conclusion

Our explorative analysis of European web archives for use by digital scholars underlines four important considerations:

(1) Digital scholars need to investigate why, by whom and on whose behalf web archiving is being done. This is important because it '(…) serves to orient users as to some of the questions they should be asking of their sources, and of the institutions that provide them' (Webster 2017: 176). With regard to why the content in question has been archived, it has been shown that the selection is based on a variety of strategies and criteria. Sometimes the collection scope is

[18] See http://timetravel.mementoweb.org/

defined by law as is the case in countries with legal deposit legislation; in other cases, the scope is defined by the heritage institution itself. In the case of national libraries and Arquivo.pt., it has been shown that two main approaches exist: broad crawls and selective crawls although some institutions combine both. Broad crawls cover top-level domain crawls and relevant content outside of the national domain(s) and selective crawls mostly focus on specific events, themes or emergencies. When it comes to social media in the studied web archives, it has been demonstrated that approaches differ widely: some institutions do not (yet) include any social media content, while others cover several platforms. Twitter, Facebook and YouTube are the social network platforms that are most often included by web archives.

Who does the web archiving is another important factor from the user perspective. Sometimes specific collection specialists are responsible, whereas in other cases, selection is a responsibility that is shared collectively by a large number of people. A number of institutions also work together with external experts for the selection of web content and most of the studied initiatives offer the public the possibility to submit suggestions to be included in the web archive.

The context in which web archiving has taken place is, therefore, very important for researchers as it has a significant impact on its use as a source in scholarly studies (Webster 2017).

(2) Access conditions differ widely between web archives and the vast majority of archived online contents is protected in order to respect the legal provisions relating to copyright. Once access to the archive is gained, most web archive interfaces only afford simple tools (e.g. URL or full-text searches). Researchers also need to take into account the integrity and authenticity of the information captured (which is strongly linked with quality assurance and metadata management procedures of the webarchive). Nielsen remarks that the 'ongoing efforts that are being made to enhance access to the archives' (2016: 22) also reveal new challenges. For example, full-text keyword search provides different possibilities for finding material, but can potentially create challenges for digital scholars such as data overload and the task of filtering out the relevant results by themselves. The latter is difficult as most digital scholars are so accustomed to seeking information through querying search engines, such as Google, where the results are ordered by relevance, which means they expect to find information in web archives in the same way (Costa and Silva 2011).

Ultimately, digital scholars need data-level access to web archives to undertake analysis using digital tools and methods. A pioneer in this area is Ian Milligan, who made, in the context of the Web Archives for Longitudinal Knowledge project (WALK n.d.), a number of datasets available, including information about those datasets and how to cite them. It is anticipated that data-level access to web archives will increase in the future (Lin et al. 2017). Digital scholars will thus need to become aware of the characteristics of web archive search results and of the fact that they can be sometimes problematic. For example, web archive search results will often be very numerous, neither ordered by relevance nor importance and full of irrelevant material and false returns (Deswarte 2015). The tools and the interfaces offered by web archives are very much in an early stage of development and web archivists are only

beginning to tackle the strengths and weaknesses of both their data and interfaces.

Another challenge identified is the need for a better understanding of user needs and behaviour because of the lack of available resources of web archiving institutions. However, a variety of studies have investigated the practices of web archiving and researchers using web archives. Studies done by BUDDAH underlined, amongst other issues, the lack of guidance for humanities researchers: 'A shared conceptual framework of the web archives research process is essential to systematize practices, advance the field, and to welcome new entrants to this area. [...] Such a framework would be structurally useful to describe any research that investigates social questions based on web archives' (Maemura et al. 2016: 3251–3252). Web archiving institutions could play a role in providing this guidance for researchers.

(3) With regard to legal frameworks for web archiving, it is important for digital scholars to understand the general legal frameworks governing web archiving. Increasingly, many European countries have extended their legal deposit legislation to include web archiving. While this means that national libraries have a formal mandate to archive the web content of their nation, there are still challenges to providing access to this content, for example, for research use. Sometimes this access can, for example, only be provided on-site within a national library. For countries where there is no legal deposit legislation in place, a number of, often pragmatic solutions—such as approaching website owners to ask permission to archive their website content—are in place to enable cultural heritage organisations to archive websites. It should also be important to have in mind that the General Data Protection Regulation (GDPR) gives the Member States the possibility to put in place a softened regime when personal data are processed in specific contexts such as archiving in the public interest, historical or scientific research and statistical purposes

(4) Using web archives as a basis for research requires, perhaps even more than other digital research materials, a relatively high level of technical knowledge. Not only is it important to understand the context in which the websites were archived (e.g. how they were selected, when and with tools were they archived), but there are also technical challenges to accessing this content (e.g. full-text search is not always readily available), and understanding the file formats (e.g. WARC) that have been used for web-archiving. However, thanks to the increasing community that is building around web-archiving (e.g. IIPC) and research using web archives (e.g. RESAW), the expertise, tools and knowledge are also growing.

Given the importance of legislative, technical and policy-related elements linked to the creation of a web archive as a research object, it is paramount to provide adequate information and documentation about this context to the users of the web archive in order to open up the black box of web archiving. The Portuguese web archive can be considered a good example in this context as videos are created that shed light onto the inner workings of the web archive, thereby furthering transparency (Arquivo.pt 2018). The features and history of a web archive are pertinent to all its users. It is particularly relevant in order to evaluate the web archive as a data source. As Laursen states: 'In short, the story of an archive is relevant for the trustworthiness of the archive' (2017: 223). We have shown that many challenges are associated with web archiving. However, some of the greatest challenges, seen from the user's perspective, come down to two factors. Firstly, that it is impossible to save everything, and that the

choices made are significant for the research object. As Masanès states: 'Web archiving is often a matter of choices, as perfect and complete archiving is unreachable' (Masanès 2005: 77). Secondly, in most cases the object researchers are attempting to preserve when creating a web archive will be distorted by the actual archiving process (Nielsen 2016). It could be argued that it is unlikely, if not impossible, that we can preserve all of the attributes and functionality of digital materials. However, little is known about the levels of loss that are acceptable for digital scholars (Harvey 2005).

7 Limitations and future research

Although this research produced useful insights on how European web archiving initiatives select and open up archived web content, the research design had some limitations. Most importantly, this study was limited to a sample of only nine web archives, eight of which are managed by heritage institutions. This is not meant to be a representative sample for the web archiving landscape, as it only includes European web archives. In addition all these web archives are members of the International Internet Preservation Consortium (IIPC), except for the National Archive of the Netherlands.

Despite these limitations, this article can function as a point of departure for more extensive and qualitative research. With regard to selection, research into the retrieval of examples of the earliest web pages of a national web domain would be very interesting, as would studies about how to ensure the representative inclusion of web material about and from minority groups in web archives. Furthermore, the different models of collaboration with partners external to national heritage institutions for selection, such as digital scholars and members of the general public, could also be an interesting research subject. From an access perspective, it could be worthwhile to explore how secure remote access to web archives could be provided for researchers, in compliance with the related legal provisions.

Furthermore, research related to data-level access to web archives would be another valuable research area, backed by a solid evidence-base from user studies. From a legal point of view, future research can center around two legal developments that will impact web archiving: on the one hand, the impact of the GDPR on the legislation of the various EU member states; on the other hand, the reform of copyright exceptions and limitations at the European level. From a technical point of view, it has been noted that 'the archive separates itself increasingly from the live web the archive tries to preserve' (Laursen and Møldrup-Dalum 2017: 216) and that further research into the development of solutions and tools for the various technical challenges web archives are confronted with is, therefore, essential.

Acknowledgements The research outlined in this article was conducted in the context of the PROMISE-project. This project received funding from the Belgian Science Policy Office (BELSPO) in December 2016, through their Belgian Research Action through Interdisciplinary Networks (BRAIN) research programme, for a 24-month period. The project was initiated by the Royal Library of Belgium and the State Archives of Belgium and the project consortium also includes the universities of Ghent and Namur and the Information and Documentation School of the Brussels-Brabant Institute of Higher Education (HE^{2}B IESSID). We would like to thank the interviewees and their colleagues for taking the time to answer our many questions.

List of institutions and representatives consulted

- National Library of The Netherlands: Kees Teszelszky (Researcher web archiving, Digital Preservation Department)
- National Archive of The Netherlands: Antal Posthumus (Adviser recordkeeping, Directie Infrastructuur & Advies) and Jeroen van Luin (Acquisition and Maintenance of Digital Archives)
- National Library of France (BnF): Pascal Tanésie (Assistant to the head of the department of digital legal deposit), Sara Aubry (Web Archiving Project Manager, IT department) and Bert Wendland (IT Department)
- National Library of Luxembourg: Yves Maurer (Webarchiving Technical Manager) and Ben Els (Digital Curator)
- The Royal Danish Library: Jakob Moesgaard (Specialkonsulent, Department of Digital Legal Deposit and Preservation) and Tue Hejlskov Larsen (IT analyst)
- The UK National Archives: Tom Storrar (Head of Web Archiving) and Claire Newing (Web Archivist)
- The British Library: Jason Webber (Web Archiving Engagement and Liaison Manager)
- Arquivo.pt.: Daniel Gomes (Head of Arquivo.pt., the Portuguese web-archive, Advanced Services Department)
- National Library of Ireland (NLI): Maria Ryan (Web Archivist)

References

Archives Unleashed Project. (2018). *The Archives Unleashed Project*. Retrieved from http://archivesunleashed.org/. Last accessed on 20/04/2018.

ArchiveSpark GitHub. (2018). Helgeho/ArchiveSpark: *An Apache Spark framework for easy data processing, extraction as well as derivation for Web archives and archival collections, developed by the Internet Archive and L3S Research Center*. Retrieved from https://github.com/helgeho/ArchiveSpark. Last accessed on 20/04/2018.

Arquivo.pt. (2018). *Arquivo.pt (Portuguese web-archive): official playlist.* Retrieved from https://www.youtube.com/playlist?list=PLKfzD5UuSdETtSCX_TM02nSP7JDmGFGIE. Last accessed on 12/02/2018.

Arquivo.pt. (n.d.-a). *Arquivo.pt*. Retrieved from http://www.arquivo.pt. Last accessed on 22/01/2018.

Arquivo.pt. (n.d.-b). *Knowledge*. Retrieved from https://www.fccn.pt/en/knowledge/arquivo-pt/. Last accessed on 20/10/2017.

Arquivo.pt. (n.d.-c). *Crawling and archiving Web content*. Retrieved from http://sobre.arquivo.pt/en/help/crawling-and-archiving-web-content/#qe-faq-2416. Last accessed on 20/10/2017.

Arquivo.pt. (n.d.-d). *Terms and conditions*. Retrieved from http://sobre.arquivo.pt/en/about/terms-and-conditions/. Last accessed on 31/01/2017.

Arquivo.pt. (n.d.-e). *What is Arquivo.pt - the Portuguese Web Archive?* Retrieved from http://sobre.arquivo.pt/en/help/what-is-arquivo-pt/. Last accessed on 20/10/2017.

Ben-David, A., & Huurdeman, H. (2014). Web archive search as research: Methodological and theoretical implications. *Alexandria, 25*(1–2), 93–111.

Beunen, A. & Schiphof, T. (2006). *Legal aspects of web archiving from a Dutch perspective* (report commissioned by the National Library in The Hague).

BnF. (2014). *Historique de l'archivage web.* Retrieved from http://www.bnf.fr/fr/professionnels/archivage_web_bnf/a.depot_legal_internet_histoire.html#SHDC__Attribute_BlocArticle1BnF. Last accessed on 22/01/2018.

BnF. (2016). *BnF Collecte de web (BCWeb).* Retrieved from https://collecteweb.bnf.fr/login.html. Last accessed on 04/02/2018.

BnF. (2017a, February). *Collectes ciblées de l'internet français*. Retrieved from http://www.bnf.fr/fr/collections_et_services/anx_pres/a.collectes_ciblees_arch_internet.html. Last accessed on 16/12/2017.

BnF. (2017b). *Internet archives.* Retrieved from http://www.bnf.fr/en/collections_and_services/book_press_media/a.internet_archives.html. Last accessed on 21/09/2017.

BnF. (2017c). *Guide des archives de l'Internet* [Brochure]. Retrieved from http://www.bnf.fr/documents/guide_archives_internet.pdf. Last accessed on 20/09/2017.

BnL. (n.d.). *Appel à participation - Bibliothèque nationale de Luxembourg.* Retrieved from: http://crawl.bnl.lu/2017/06/appel-a-participation-bibliotheque-nationale-de-luxembourg-web-archive/. Last accessed on 26/01/2018.

British Library. (2017a, April 18). *The challenges of web archiving social media* [web log message]. Retrieved from http://blogs.bl.uk/webarchive/2017/04/the-challenges-of-web-archiving-social-media.html. Last accessed on 30/10/2017.

British Library. (2017b, May 17). *Web Archiving Engagement Manager.* Retrieved from https://www.bl.uk/people/experts/jason-webber. Last accessed on 04/02/2018.

British Library. (n.d.-a). *UK web archive*. Retrieved from https://www.bl.uk/collection-guides/uk-web-archive. Last accessed on 31/10/2017.

British Library. (n.d.-b). *Explore the British Library. Non-print legal deposit: FAQs.* Retrieved from http://www.bl.uk/catalogues/search/non-print_legal_deposit.html. Last accessed on 31/10/2017.

Brozzler GitHub. (2018). *internetarchive/brozzler: brozzler - distributed browser-based web crawler*. Retrieved from https://github.com/internetarchive/brozzler. Last accessed on 20/04/2018.

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 present* (pp. 62–80). London: UCL Press.

BUDDAH, Big UK Domain Data for the Arts and Humanities. (2014) *Bursaries.* Retrieved from https://buddah.projects.history.ac.uk/news/bursaries/ . Last accessed on 04/02/2018.

Chakraborty, A., & Nanni, F. (2017). The changing digital faces of science museums: A diachronic analysis of museum websites. In N. Brügger (Ed.), *Web 25. Histories from the first 25 years of the world wide web* (pp. 157–174). New York: Peter Lang.

Clarke, N. (2016). *JWAT.* Retrieved from https://sbforge.org/display/JWAT/JWAT. Last accessed on 20/04/2018.

Costa, M. & Silva, M. (2010). Understanding the information needs of web archive users. In *Proceedings of the 10th International Web Archiving Workshop* (pp. 9-16).

Costa, M. & Silva, M. (2011). Characterizing search behavior in web archives. In *Proceedings of the 1st International Temporal Web Analytics Workshop*.

Deswarte, R. (2015). *Revealing British euroscepticism in the UK web domain and archive case study.* Retrieved from http://sas-space.sas.ac.uk/6103/#undefined. Last accessed on 25/01/2018.

Dooley, J. (2016 October). *Metadata to meet user needs*. Presented at the OCLC Member Forum. Los Angeles.

Dooley, J. M., Farrell, K. S., Kim, T. & Venlet, J. (2017). Developing web archiving metadata best practices to meet user needs. *Journal of Webstern Archives, 8*(2), Art. 5, 15 pp.

Dougherty, M., Meyer, E. T., Madsen, C., van den Heuvel, C., Thomas, A., & Wyatt, S. (2010). *Researcher engagement with web Archives: State of the art*. London: JISC.

Fbarc GitHub. (2018). *justinlittman/fbarc: A commandline tool and Python library for archiving data from Facebook using the Graph API.* Retrieved from https://github.com/justinlittman/fbarc. Last accessed on 20/04/2018.

Foo, C. (2016). *Welcome to Wpull's documentation! - Wpull 2.0.1 documentation*. Retrieved from https://wpull.readthedocs.io/en/master/#. Last accessed on 20/04/2018.

Free Software Foundation. (2017) *Wget - GNU Project - Free Software Foundation*. Retrieved from https://www.gnu.org/software/wget/. Last accessed on 20/04/2018.

Gomes, D. (2017a, November 30). *Web preservation demands access*. Retrieved from http://www.dpconline.org/blog/idpd/web-preservation-demands-access. Last accessed 14/12/2017.

Gomes, D. (2017b, November 24) *Personal interview via Zoom with Daniel Gomes /Interviewers: Sally Chambers, Friedel Geeraert, Gerald Haesendonck, Alejandra Michel and Eveline Vlassenroot.* [M4A file].

Grab-site GitHub. (2018). *ludios/grab-site: The archivist's web crawler: WARC output, dashboard for all crawls, dynamic ignore patterns*. Retrieved from https://github.com/ludios/grab-site. Last accessed on 20/04/2018.

Graff, E. & Sepetjan, S. (2011). Le dépôt légal en France. *Les cahiers de la propriété intellectuelle*, 2011/1, 179–180.

Harvey, D. R. (2005). *Preserving digital materials*. München: KG Saur.

Helmond, A., Nieborg, D., & van der Vlist, F. N. (2017). The political economy of social data: A historical analysis of platform–industry partnerships. In *Proceedings of the 8th International Conference on Social Media & Society* (SMSociety 17) New York: ACM Press. https://doi.org/10.1145/3097286.3097324.

Hockx-Yu, H. (2014). *Archiving social media in the context of non-print legal deposit.* Paper presented at IFLA, Lyon.

IIPC. (2017). *Why archive the web?* Retrieved from http://netpreserve.org/web-archiving/. Last accessed on 22/01/2018.

IIPC. (2018). *OpenWayback*. Retrieved from http://netpreserve.org/web-archiving/openwayback/. Last accessed on 09/02/2018.

ISO. (2017). *Information and documentation - WARC file format (ISO 28500:2017).*

KB Nederland (n.d.-a) *Selectie bij webarchivering*. Retrieved from https://www.kb.nl/organisatie/onderzoek-expertise/e-depot-duurzame-opslag/webarchivering/selectie-bij-webarchivering. Last accessed on 19/12/2017.

KB Nederland (n.d.-b). *Legal issues*. Retrieved from https://www.kb.nl/en/organisation/research-expertise/long-term-usability-of-digital-resources/web-archiving/legal-issues. Last accessed on 22/09/17.

KB Nederland (n.d.-c). *Web archiving*. Retrieved from https://www.kb.nl/en/organisation/research-expertise/long-term-usability-of-digital-resources/web-archiving. Last accessed on 22/09/17.

KB Nederland (n.d.-d) *KB-webarchief: veelgestelde vragen*. Retrieved from https://www.kb.nl/organisatie/onderzoek-expertise/e-depot-duurzame-opslag/webarchivering/kb-webarchief-veelgestelde-vragen. Last accessed 08/12/2017.

KB Nederland (n.d.-e) *Gebruiksvoorwaarden webarchief Koninklijke Bibliotheek.* Retrieved from https://www.kb.nl/bronnen-zoekwijzers/databanken-mede-gemaakt-door-de-kb/webarchief-kb/gebruiksvoorwaarden-webarchief-koninklijke-bibliotheek. Last accessed on 08/12/2017.

Kelly, M. (2017). *Web Archiving Integration Layer (WAIL)*. Retrieved from https://machawk1.github.io/wail/. Last accessed on 20/04/2018.

Koerbin, P. (2017). Revisiting the world wide web as artefact: Case studies in archiving small data for the National Library of Australia's PANDORA archive. In N. Brügger (Ed.), *Web 25. Histories from the first 25 years of the world wide web* (pp. 191–206). New York: Peter Lang.

Kremer, I. (2016). *About the Time Travel Service*. Retrieved from http://timetravel.mementoweb.org/about/. Last accessed on 20/04/2018.

Kunze, S. & Power, B. (n.d.). *The 1916 Easter Rising Web Archive* Project, p. 2. Retrieved from https://archivedweb.blogs.sas.ac.uk/files/2017/06/RESAW2017-PowerKunze-The_1916_Easter_Rising_web_archive_Project.pdfp_.pdf. Last accessed on 2/11/2017.

Laursen, D., & Møldrup-Dalum, P. (2017). Looking back, looking forward: 10 years of web development to collect, preserve and access the Danish web. In N. Brügger (Ed.), *Web 25. Histories from the first 25 years of the world wide web* (pp. 207–228). New York: Peter Lang.

Lin, J., Milligan, I., Wiebe, J., & Zhou, A. (2017). Warcbase: Scalable analytics infrastructure for exploring web archives. *Journal on Computing and Cultural Heritage, 10*(4), 1–30. https://doi.org/10.1145/3097570.

Maemura, E., Becker, C., & Milligan, I. (2016). *Understanding computational web archives research methods using research objects*. In James Joshi, George Karypis, Ling Liu, et al., *2016 IEEE International Conference on Big Data* (Big Data)(pp. 3250–3259).

Masanès, J. (2005). Web archiving methods and approaches: A comparative study. *Library Trends, 54*(1), 72–90.

Maurer, Y. & Els, B. (2017a, November 24). *Personal interview via GoToMeeting with Yves Maurer and Ben Els/Interviewers: Emmanuel Di Pretoro, Friedel Geeraert, Gerald Haesendonck, Eveline Vlassenroot.* [M4A file].

Maurer, Y. & Els, B. (2017b, November 24). *Written answers given by the Bibliothèque nationale de Luxembourg via Google Docs before the personal interview with Yves Maurer and Ben Els/Interviewers: Emmanuel Di Pretoro, Friedel Geeraert, Gerald Haesendonck, Eveline Vlassenroot.*

Moesgaard, J. & Larsen, T. H. (2017a, November 30). P*ersonal interview via GoToMeeting with Jakob Moesgaard & Tue Hejlskov Larsen/Interviewers: Emmanuel Di Pretoro,* Friedel Geeraert, Gerald Haesendonck, Sally Chambers and Alejandra Michel.

Moesgaard, J. & Larsen, T. H. (2017b, November 30). *Written answers given by the Danish Royal Library via Google Docs before the personal interview with Jakob Moesgaard & Tue Hejlskov Larsen/Interviewers: Emmanuel Di Pretoro, Friedel Geeraert, Gerald Haesendonck, Sally Chambers and Alejandra Michel*.

National Archives. (n.d.-a). How to use the web archive. Retrieved from http://www.nationalarchives.gov.uk/webarchive/information/. Last accessed on 19/10/2017.

National Archives. (n.d.-b). *UK Government web archive*. Retrieved from http://www.nationalarchives.gov.uk/webarchive/. Last accessed on 31/10/2017.

National Library of Ireland. (2017a). *NLI Review 2016*. Retrieved from https://www.nli.ie/GetAttachment.aspx?id=011e629f-1a5a-4cde-91d7-8a62ccf84bef. Last accessed 9/10/2017.

National Library of Ireland. (2017b). *Web Archive FAQ & Resources*. Retrieved from https://www.nli.ie/en/web-archive-faq.aspx. Last accessed on 9/10/2017.

National Library of Ireland. (n.d.-a) *NLI Web Archive: A record of the online life in Ireland*. Retrieved from http://collection.europarchive.org/nli. Last accessed on 1/02/2018.

National Library of Ireland. (n.d.-b). *Rights and Reproductions.* Retrieved from https://www.nli.ie/en/rights-reproductions.aspx. Last accessed on 31/01/2018.

National Library of Ireland. (n.d.-c). *Web Archive*. Retrieved from https://www.nli.ie/en/web_archive.aspx. Last accessed on 31/01/2018.

National Library of Ireland. (n.d.-d). *Web archive collections*. Retrieved from http://www.nli.ie/en/udlist/web-archive-collections.aspx. Last accessed on 20/10/2017.

NCDD. (n.d.). *Expertgroep webarchivering*. Retrieved from http://www.ncdd.nl/kennis-en-advies/expertgroepen/expertgroep-webarchivering/. Last accessed on 08/12/2017.

Netarkivet.dk. (2016a). *Selektive høstninger*. Retrieved from http://netarkivet.dk/om-netarkivet/selektive-hostninger_2016/. Last accessed on 31/10/2017.

Netarkivet.dk. (2016b). *Adgang til Netarkivet*. Retrieved from http://netarkivet.dk/adgang/. Last accessed on 31/01/2018.

Netarkivet.dk. (2017). *Brugermanual til Netarkivet*. Retrieved from: http://netarkivet.dk/wp-content/uploads/2015/03/Netarkivet_Strategi_Langtidsbevaring_1.0_150115.pdf . Last accessed on 1/02/2018.

Nielsen, J. (2016). *Using web archives in research - an introduction*. Retrieved from http://www.netlab.dk/wp-content/uploads/2016/10/Nielsen_Using_Web_Archives_in_Research.pdf. Last accessed on 18/01/2018.

Node-warc GitHub. (2018). *N0taN3rd/node-warc: Parse And Create Web ARChive (WARC) files with node.js*. Retrieved from https://github.com/N0taN3rd/node-warc. Last accessed on 20/04/2018.

Ogden, J., Halford, S. & Carr, L. (2017). Observing web archives. The case for an ethnographic study of web archiving. *WebSci. June* (25-28). https://doi.org/10.1145/3091478.3091506.

Posthumus A. and van Luin, J. (2017a, December 6). Personal interview via UC4all with Antal Posthumus *and Jeroen van Luin/Interviewers: Eveline Vlassenroot and Friedel Geeraert.*

Posthumus A. and van Luin, J. (2017b, December 6). *Written answers given via Google Docs by the National Archive before the personal interview with Antal Posthumus and Jeroen van* Luin/Interviewers*:* Eveline Vlassenroot and Friedel Geeraert.

Pywb GitHub. (2018). *webrecorder/pywb: Core Python Web Archiving Toolkit for replay and recording of web archives https://pypi.python.org/pypi/pywb*. Retrieved from https://github.com/webrecorder/pywb. Last accessed on 20/04/2018.

RESAW (Research Infrastructure for the Study of Archived Web Materials). (2012). *About RESAW.* Retrieved from http://resaw.eu/about/. Last accessed on 04/02/2018.

Reyes Ayala, B. (2013). *Web archiving bibliography 2013*. Texas: UNT Digital Library.

Roche, X. (2018). *HTTrack Website Copier*. Retrieved from http://www.httrack.com/. Last accessed on 20/04/2018.

Rosenthal, C. (2017, July). *NetarchiveSuite*. Retrieved from https://sbforge.org/display/NAS/NetarchiveSuite. Last accessed on 20/04/2018.

Ryan, M. (2017, November 16). *Personal interview via GoToMeeting with Maria Ryan/Interviewers: Gerald Haesendonck, Alejandra Michel and Eveline Vlassenroot.* [M4A file].

Schneider, S. M., & Foot, K. A. (2005). Web sphere analysis: An approach to studying online action. In C. Hine (Ed.), *Virtual Methods - Issues in Social Research on the Internet*. Oxford: Berg Publishers, 157–171.

Schneider, S., & Foot, K. (2008). Archiving of internet content. In W. Donsbach (Ed.), *The international encyclopedia of communication*. Oxford: Blackwell. https://doi.org/10.1002/9781405186407.wbieca051.

Schroeder, R., & Brügger, N. (2017). Introduction: The web as history. In N. Brügger & R. Schroeder (Eds.), *The web as history. Using web archives to understand the past and present* (pp. 1–19). London: UCL Press.

Sierman, B., & Teszelszky, K. (2017). How can we improve our web collection? An evaluation of web archiving at the KB National Library of the Netherlands (2007-2017). *Alexandria, 27*, 94–107. https://doi.org/10.1177/0955749017725930.

Social Feed Manager. (2018). *Social Feed Manager*. Retrieved from https://gwu-libraries.github.io/sfm-ui/. Last accessed on 20/04/2018.

Tanésie, P. & Aubry, S. (2017, December 12). Le dépôt légal du web à la BnF : organisation, procédures et outils. Presentation given at the Bibliothèque nationale de France, Paris.

Tanésie, P., Aubry, S., Wendland, B. (2017, December 12), *Personal interview at the BnF with Pascal Tanésie, Sara Aubry & Bert Wendland/Interviewers: Sally Chambers, Rolande Depoortere, Friedel Geeraert, Alejandra Michel, and Eveline Vlassenroot* [mp3 file].

Teszelszky, K. (2017a, November 8). *Personal interview via GoToMeeting with Kees Teszelszky/Interviewers: Gerald Haesendonck, Alejandra Michel and Eveline Vlassenroot.* [M4A file].

Teszelszky, K. (2017b, November 8). *Written answers given via Google Docs by the KB Nederland before the personal interview with Kees* Teszelszky/Interviewers*:* Gerald Haesendonck, Alejandra Michel and Eveline Vlassenroot.

The National Archives. (n.d.). *UK Government Web Archive*. Retrieved from http://www.nationalarchives.gov.uk/webarchive/. Last accessed on 1/02/2018.

Twarc GitHub. (2018). *DocNow/twarc: A command line tool (and Python library) for archiving Twitter JSON*. Retrieved from https://github.com/DocNow/twarc. Last accessed on 20/04/2018.

UK Web Archive. (n.d.-a). *About*. Retrieved from https://www.webarchive.org.uk/ukwa/info/about. Last accessed on 30/10/2017.

UK Web Archive. (n.d.-b). *Browse*. Retrieved from https://www.webarchive.org.uk/ukwa/browse. Last accessed on 1/02/2018.

UK Web Archive. (n.d.-c). *Frequently asked questions*. Retrieved from https://www.webarchive.org.uk/ukwa/info/faq. Last accessed on 30/10/2017.

UK Web Archive. (n.d.-d). *SHINE*. Retrieved from https://www.webarchive.org.uk/shine. Last accessed on 05/02/2018.

Van de Sompel, H., Nelson, M.L., Sanderson, R. (2013). *RFC 7089: HTTP Framework for Time-Based Access to Resource States—Memento*. Retrieved from http://tools.ietf.org/rfc/rfc7089.txt. Last accessed on 20/04/2018.

WALK (Web Archives for Longitudinal Knowledge). (n.d.). *Datasets*. Retrieved from: http://webarchives.ca/datasets. Last accessed on 04/02/2018.

WARCAT GitHub. (2017). *chfoo/warcat: Tool and library for handling Web ARChive (WARC) files*. Retrieved from https://github.com/chfoo/warcat. Last accessed on 20/04/2018.

Warcio GitHub. (2017). *webrecorder/warcio: Streaming WARC/ARC library for fast web archive IO https://pypi.python.org/pypi/warcio*. Retrieved from https://github.com/webrecorder/warcio. Last accessed on 20/04/2018.

Warctools GitHub. (2016). *internetarchive/warctools: warctools*. Retrieved from https://github.com/internetarchive/warctools. Last accessed on 20/04/2018.

Webber, J. (2017, November 16). *Personal interview via GoToMeeting with Jason Webber/Interviewers: Sally Chambers, Gerald Haesendonck, Alejandra Michel and Eveline Vlassenroot.* [M4A file].

Weber, M. S. (2017). The tumultuous history of news on the web. In N. Brugger & R. Schroeder (Eds.), *The web as history. Using web Archives to understand the past and the present* (pp. 83–100). London: UCL Press.

Webrecorder. (n.d.). Collect & revisit the web. Retrieved from https://webrecorder.io/. Last accessed on 19/02/2019.

Webrecorder Player for Desktop Github. (2018). *webrecorder/webrecorderplayer-electron: Webrecorder Player for Desktop (OSX/Windows/Linux). (Built with Electron + Webrecorder)*. Retrieved from https://github.com/webrecorder/webrecorderplayer-electron. Last accessed on 20/04/2018.

Webster, P. (2017). Users, technologies, organisations: Towards a cultural history of world web archiving. In N. Brügger & N. (Eds.), *Web 25. Histories from 25 years of the world wide web* (pp. 175–190). New York: Peter Lang.

Affiliations

Eveline Vlassenroot[1] · Sally Chambers[2] · Emmanuel Di Pretoro[3] · Friedel Geeraert[4] · Gerald Haesendonck[5] · Alejandra Michel[6] · Peter Mechant[1]

Emmanuel Di Pretoro
edipretoro@he2b.be

Friedel Geeraert
Friedel.Geeraert@kbr.be

Gerald Haesendonck
Gerald.Haesendonck@UGent.be

Alejandra Michel
alejandra.michel@unamur.be

Peter Mechant
Peter.Mechant@UGent.be

1 imec-mict-UGent, Ghent, Belgium

2 Ghent Centre for Digital Humanities, UGent, Ghent, Belgium

3 URF-SID, Haute École Bruxelles-Brabant, Bruxelles, Belgium

4 Royal Library and State Archives of Belgium, Brussel, Belgium

5 Department of Electronics and Information Systems, Ghent University - imec – IDLab, Ghent, Belgium

6 NADI/CRIDS, UNamur, Namur, Belgium