International Journal of Digital Humanities (2021) 2:129–144
https://doi.org/10.1007/s42803-021-00037-0

ORIGINAL PAPER

The values of web archives

Valérie Schafer[1] · **Jane Winters**[2]

Received: 13 September 2020 / Accepted: 18 April 2021 / Published online: 10 June 2021

Abstract

This article considers how the development, promotion and adoption of a set of core values for web archives, linked to principles of "good governance", will help them to tackle the challenges of sustainability, accountability and inclusiveness that are central to their long-term societal and cultural worth. It outlines the work that has already been done to address these questions, as web archiving begins to move out of its establishment phase, and then discusses seven key principles of good governance that might be adapted by and embedded within web archives: participation, consensus, accountability, transparency, effectiveness and efficiency, inclusivity and legality. The article concludes with a call to action for researchers and archivists to co-create the core values for web archives that will be required if they are to remain a vital part of our cultural heritage infrastructure.

Keywords Web archives · Good governance · Sustainability · Inclusiveness · FAIR data · Openness

In gathering, preserving, curating, publishing and/or analysing an intangible and massive born-digital heritage, key stakeholders, whether they are libraries, private companies, scholars or others, face challenges which are particularly critical when they concern the sustainability, accountability and inclusiveness of, and engagement with, web archives. Indeed, will web archives be valuable in the long-term without establishing and embedding well-articulated core values? How can "good governance" in web archiving and web practices be ensured? Calls for FAIR[1] data (Wilkinson et al., 2016), for openness in the publishing and academic worlds (Mons, 2018), for "good governance" in knowledge infrastructures, for sustainability in digital studies (Barats et al., 2020) and for inclusiveness in living archives (Rhodes,

[1] FAIR stands for Findable, Accessible, Interoperable, and Reusable.

✉ Valérie Schafer
valerie.schafer@uni.lu

[1] C²DH, University of Luxembourg, Esch-sur-Alzette, Luxembourg

[2] School of Advanced Study, University of London, London, UK

2013; Rollason-Cass & Reed, 2015) are just a few of the issues that challenge their practices and purposes, as web archives exit the early age of preservation and begin to grow to maturity (e.g. Teszelszky, 2019). Our article aims to analyse these challenges, as well as steps towards ethical and public engagement with web archives.

The first section will delineate the evolution of web archives since 1996, to show how far the field has come and the road ahead. It will specifically focus on values, ethics, and public and researcher engagement, as much has already been written on the history of web archiving itself.

The second section will demonstrate why strong core values are now needed to help web archives mature further, to encourage greater and more informed use of the archived web, and to develop the role of web archives in both a "post-truth world" and a changing academic ecosystem. The general public and civil society must also be taken into account, as they ask for greater inclusiveness (or indeed to be explicitly excluded) or wish to engage with this born-digital heritage through digital public history. Whose stories are held within web archives, who owns those stories, and how can trust be engendered effectively?

We will conclude by briefly reflecting on the paths towards multi-stakeholder engagement, and by envisioning changes arising from developments in digital history, technology-assisted curation and artificial intelligence, as well as transnational metadata repositories and other future challenges. These questions are key to the long-term value and sustainability of web archives.

1 From ad hoc solutions and pioneering work to some kind of maturity …

"I like the idea that there's a God somewhere looking after us. But sometimes I think it's a wrathful God. Who are the right gods? Janus sounds nice, but the idea of Sisyphus having to roll a rock continuously up a hill strikes me as more appropriate. When tending files on spinning magnetic storage systems and trying to preserve them for centuries, I think of Damocles' sword hanging over us." (Kahle, 2007: 23).

Undoubtedly, if the Greek pantheon watches over the Internet Archive (IA), its founder, Brewster Kahle, is not wrong to refer to Sisyphus and Damocles. The former hints at the titanic task to be accomplished in terms of capturing and preserving digitized and born-digital heritage, a responsibility so huge and complex that, over the years, the IA has given up archiving "the entire World Wide Web", even though it remains a pioneering and gigantic collection, dating back to 1996 (468 billion web pages had been saved as of 5 September 2020). As for Damocles, his shadow is present in the actions taken against the IA, whether the recent copyright lawsuit brought by a group of publishers[2] or cases of censorship (in Russia, China or Jordan).[3] But two thoughts spring to mind when faced with this pantheon: Janus

[2] For further details: https://blog.archive.org/2020/07/29/internet-archive-responds-to-publishers-lawsuit/. All URLs cited in this article were accessed and checked on 5 September 2020.

[3] https://blog.archive.org/2017/04/11/who-blocked-the-archive-in-jordan/

undoubtedly "sounds nice" but he invites us to think about duality and gatekeeping; and what about the Greek goddesses in this exclusively male pantheon?

When the IA was founded, the comparison with the Library of Alexandria and the motto *Universal access to knowledge* were prominent, but as the live web grew and the uses of web archives also developed, the project has evolved. We will not revisit here the history of web archiving, which has already been well covered (Masanes, 2006; Musiani et al., 2019; Webster, 2017). More important for our purpose is to underline the importance of tracing the frameworks, values and imaginaries expressed throughout the ages of web archiving. In this respect, the "time travel" offered by the IA (Ankerson, 2015) provides interesting examples, from the choice of the Wayback Machine's name itself to the button on its home page which changes from "Take Me Back" in 2001 to "Browse History" in the course of a major redesign of the website in 2014. The IA can become, thanks to its own web archives, in a remarkable *mise en abyme*, an object of study to help us interrogate the paradigm shifts at work, through analysis of expressed intentions, the evolution of the FAQs or the tools both proposed and developed.

A purer design, stronger visual aspects, enhanced tools and increased user participation are at the heart of a project that nonetheless retains remarkable stability, particularly in the FAQs. Indeed, despite its reorganization, which reflects changes in the audiences that the site primarily addresses – in 2013 it was still website owners who were the first to find answers to their questions,[4] but in 2015 search tips and therefore the users of the Wayback Machine were prioritized[5] – elements that might have been expected to have evolved in line with the digitization of society and related policies, in particular those on legal issues and on digital heritage, remain very stable. This is true of the copyright policy, for example, which shows little change between 2004 and 2015.

Search tips became more prominent as the IA and the Wayback Machine began to enhance their toolbox in the second half of 2010. Enhancements notably included search of the full text of home pages; timestamps, allowing users to identify the temporal patchwork within an archived web page; and a changes function, revealing the degree of modification of a website. This wide range of tools, which rely both on data and metadata, allow new kinds of research, drawing on both close and distant reading techniques. Evolving documentary models and paradigm shifts in web archiving have a profound impact on collections, but also on research. Thus scholars must get used to seeing their research composed, on the one hand, of archives that have responded to different documentary paradigms, and on the other hand, if their research is conducted over a long period of time, to seeing the tools modify their research capacity.

However, it would be unfair to focus here only on the IA. In the 2000s, state initiatives to collect "national" domains, whose complexity of definition is equalled

[4] Archive.org FAQ, archived on 30/05/2013 http://web.archive.org/web/20130530012907/http://archive.org/about/faqs.php#20

[5] Archive.org FAQ, archived on 03/06/2015 http://web.archive.org/web/20150603185729/https://archive.org/about/faqs.php

only by the diversity of legislation or the scope of collection, have increased (Gomes et al., 2011). To this should obviously be added the recent initiatives aimed at archiving social-digital networks, the most prominent of which is the hosting of the Twitter archives by the Library of Congress, with the well-documented difficulties of making them accessible (Bruns, 2018). However, each institution also follows its own chronology, according to particular considerations, specific constraints, etc. The French National library (BnF) thus distinguishes several stages in its archiving: 1999–2004, or the time of experimentation; 2004–2007, or the implementation of an "integrated model", legally stabilized by the extension of French legal deposit legislation to include the Internet; and 2007–2012, with the completion of a whole archiving cycle. To these three periods, we may add a more recent one as the BnF has increasingly turned to the idea of a service enabling researchers to create and share corpora (Moiraghi, 2018). The story of web archiving at the UK's British Library (BL) is similarly marked by different phases of development. As at the BnF, the first phase, from 2001 to 2003 was marked by experimentation and feasibility assessment; from 2003 to 2008, collaboration was key, with the BL forming part of the UK Web Archiving Consortium, alongside the then Joint Information Systems Committee (JISC), The National Archives of the UK, the National Library of Wales, the National Library of Scotland and the Wellcome Library; from 2008 to 2013, the BL continued selectively to archive the UK web, on a permissions basis; and from April 2013 to the present, annual domain crawls have been undertaken on the basis of legal deposit legislation, with data gradually included in an open full-text search from late 2017 (Bingham, 2015; Webber, 2017).

These temporalities are based both on internal evolution and the sharing of international experience within the International Internet Preservation Consortium (IIPC), but also on socio-technical changes and in particular the emergence of new formats or platforms. But there are limiting factors that can slow down these evolutions. For example, when dealing with social media, to the diverse legal status of social networks, and in particular private content on Facebook, may be added constraints linked to APIs. Twitter is probably the most captured social network because of its public API. There are also constraints linked to collection techniques, with Heritrix, a crawler widely used in the world of web archives, struggling to capture certain content. Sometimes Instagram or TikTok are included in special collections, such as the one on COVID-19 coordinated by the IIPC,[6] but they are rarely archived on a regular basis:

> "More or less successfully, we tried to capture content from Facebook, TikTok, Twitter, YouTube, Instagram, Reddit, Imgur, Soundcloud, and Pinterest. Twitter is the platform we are able to crawl with Heritrix with rather good results. We collect Facebook profiles with an account at Archive-It, as they have a better set of tools for capturing Facebook. With frequent Quality Assur-

[6] From mid-February 2020 the IIPC launched a collaborative collection, available on Archive-It. National institutions have provided this international initiative with the benefit of their detailed knowledge of their web space. https://archive-it.org/collections/13529

> ance and follow-ups, we also get rather good results from Instagram, TikTok and Reddit. […]
> As Heritrix has problems with dynamic web content and streaming, we also used Webrecorder.io […]. However, captures with Webrecorder.io are only drops in the ocean. The use of Webrecorder.io is manual" (Schostag, 2020).

To come back to the motivations of national institutions with responsibility for archiving the web, although these are often libraries, the comparison with the Library of Alexandria is rarely chosen. It is indeed the image of the conservation of publications that is highlighted in the context of legal deposit. However, some web archives are not under the control of libraries, such as arquivo.pt, which is linked to the Portuguese national research and education network. The Archive Team, composed of volunteers, and described by Jason Scott as having "started out of anger and a feeling of powerlessness, this feeling that we were letting companies decide for us what was going to survive and what was going to die",[7] has made a specialty of preserving endangered vernacular heritage, like Geocities. In the Doc Now project, the aim is to document social movements, in collaboration with academics and activists in particular, and the motivations of these living archives go beyond the work of knowledge infrastructures.

Paradigm shifts are sometimes disconcerting but also enriching for the different actors of web archiving, insofar as the "new" participants can usefully capitalize on the experiences of the pioneers. This was well understood in Belgium, for example, by the teams behind the PROMISE (Geeraert, 2020) and BESOCIAL projects, which carried out an important and wide-ranging analysis of the experiences of its neighbours. We can also highlight a form of maturity in archiving, with some processes already well integrated, such as emergency collections responding to sudden and unprecedented events of future historical interest. Moreover, developments in the field to date have been marked by close collaboration between archivists and researchers, which allows for the early identification of researcher needs – whether through questionnaires, pioneering studies like BUDDAH,[8] joint research projects like ASAP[9] or datathons – and in some instances the co-creation of tools and interfaces.

However, some controversies have also arisen in recent months, which testify to the fact that the agreement between the multiple stakeholders is fragile, as web archives become a political as well as a scientific and societal issue. As mentioned above, there were lawsuits and debates between the IA and editors, but also between the IA and Doc Now during the George Floyd protests, as exemplified by some tweets when Jason Scott and others invited people to participate in the collecting effort: "Don't listen to this person. You will be putting protesters lives at risk. Police were being violent against protesters all night & they'll also come looking for them later via video & photographs. The Internet Archive doesn't care about

[7] https://en.wikipedia.org/wiki/Archive_Team
[8] Big UK Domain Data for the Arts and Humanities, https://buddah.projects.history.ac.uk/
[9] Archives Sauvegarde Attentats Paris. https://asap.hypotheses.org

this #georgefloyd", wrote the official account of Doc Now on 31 May 2020. There were also unfavourable reactions to the column published in a French newspaper by researchers claiming to be creating an "ordinary memory of the extraordinary" (Piguet & Montebello, 2020) related to COVID-19. An archivist at the BnF reacted immediately: "It is a pity that the authors are obviously not aware of the periodic legal deposit or the @DLwebBnF from @laBnF". Within the ranks of archivists themselves, there are also calls for archiving policies to be opened up more widely, especially to Black people,[10] joined by civil society. These few examples testify to a demand for inclusive archiving and an increasingly diverse range of stakeholders. Certainly "demonstrating the value of Internet (and web) histories" (Winters, 2017) is on the right track. We should, however, sound a note of caution. More inclusive archiving will meet the needs of many stakeholders, but how do we account for those individuals and groups who do not want to be included and who assert ownership of their stories by rejecting the archiving efforts of even the most well-intentioned third parties? "How can individuals and communities using social media consent to archiving, or at least be meaningfully informed of it?" (Taylor, 2015). On 3 December 2018, Tumblr announced that a range of "adult content" would be removed from the platform on 17 December, leaving Archive Team volunteers just two weeks to try to capture around 700,000 blogs (Captain, 2018). One of the first responses to a call for volunteer archivists on Twitter, from an independent user, claimed that "Some of this content is intensely personal and intimate and people should retain control over it". This kind of response to web archiving initiatives will not be unique.

2 Valuing web archives

The recent controversies invite us to place web archiving within a broader societal dynamic with regard to heritage and archives and to revisit the issue raised a few years ago: do web archives have politics (Musiani & Schafer, 2017)? To take up the Langdon Winner (1980) formula, web archives certainly do have politics. These policies are both internal and external to institutions. We will not go into detail here about what happened on November 2013, when the UK's Conservative party deleted more than a decade's worth of speeches from its website and temporarily blocked access to the IA's Wayback Machine (Winters, 2017), or the call for "Backing up the history of the Internet in Canada to save it from Trump" (Conger, 2016). One could also mention the political role of Archive-It, which contains collections dedicated to Wikileaks, or to the Jasmine Revolution and the Ukrainian conflict. Social events in Ferguson, Missouri served as a reference point for the DocNow project, launched in 2016. We have also discussed above the role of internal, national and institutional policies at work. Alongside questions of politics, there are issues of governance and management. In this respect, the central question of "good governance" seems to be a relevant framework for including web archives and thinking about them. A

[10] https://twitter.com/blkgrlarchivist/status/1269415733106290688

United Nations paper on this question identifies eight major characteristics of good governance, which provide a useful framework for thinking about web (and other born-digital) archives: "Good governance … is participatory, consensus oriented, accountable, transparent, responsive, effective and efficient, equitable and inclusive and follows the rule of law".[11] Let us consider each of these characteristics in turn, and their application to and relevance for web archives.

2.1 Participatory

Participation in web archiving can take many forms, from the participatory design of tools and interfaces evidenced in the BUDDAH project to the co-creation of collections. The former remains relatively rare, but there are numerous examples of successful participatory collection practice. This is most often evident in special collections, focused on particular events, and takes two main forms: an open invitation to suggest websites for inclusion in a collection; and the solicitation of individual memories or testimony. Occasionally, an idea for a special collection might originate from outside the collecting institutions themselves, for example the collection on the French in London in the UK Web Archive which arose from the research of Saskia Huc-Hepher,[12] but the topics of the collections are generally decided in advance by the host libraries and archives. There are some striking examples of fully participatory archiving, including the pioneering Charlie Archive, based at Harvard Library.[13] Participatory approaches have also been common in the numerous COVID-19 collecting initiatives which have characterised 2020. The National Library of Luxembourg launched a call for participation which allowed it to integrate new websites which would otherwise have been overlooked ("For example, the Muslim community and shoura.lu. I hadn't thought of looking for religious communities. The Muslim community posted online information and recommendations for its members about services in mosques, religious holidays, etc. Based on this suggestion, we then looked more closely at other religious communities" (Els & Schafer, 2020)). Under the rubric "Days with Corona", Danish citizens were asked to help the Royal Library document the COVID-19 lockdown (Schostag, 2020) by submitting photos and stories from their own lives, as well as by nominating websites and social media accounts for inclusion in the web archive (the two kinds of participation we have already identified).

If special collections frequently have a participatory element, this is less often the case with regular "national" collections, which require automated crawling on a huge scale. Nevertheless, options for individuals to nominate websites for inclusion in an archive are becoming more widespread. The IA has for some time had an option to "Save page now", linking the feature to the need for reliable digital

[11] What is good governance? https://www.unescap.org/sites/default/files/good-governance.pdf
[12] https://www.webarchive.org.uk/en/ukwa/collection/309
[13] https://library.harvard.edu/collections/charlie-archive

citation.[14] The possibility of nominating websites for inclusion has also long been an option for users of the UK Web Archive, who are now encouraged by a top-level menu item to "Save a UK website". There is explicit acknowledgement that "there are vast numbers of websites that we [archivists] miss simply because we don't know about them".[15] Participation enriches the archive, but it remains relatively ad hoc and, of course, only reaches those who are already digitally engaged. Wider engagement will lead to different forms of participation and help to shape inclusive collections.

2.2 Consensus oriented

If participation is already firmly on the agenda for web archives, consensus is a rather more challenging concept. As noted in the UN paper, "There are several actors and as many viewpoints in a given society", which need to be mediated. Consensus can only be achieved through "an understanding of the historical, cultural and social contexts of a given society or community". Archives may aim to represent the societies within which they are positioned – "The concept of representation is firmly embedded in archival processes, whether articulated or not" (Charlton, 2017: 2) – and to reflect the varied perspectives of individuals and organisations, but this is more often achieved through generous collecting and heterogeneous collections than through any attempt to reach consensus. For example, the special collections relating to political elections aim to collect comprehensively, without discriminating between different parties or political positions. Perhaps, then, where web archives are concerned, it is more useful to think in terms of trust rather than consensus. If web archives are trusted, and their role in society understood and valued, then the inclusion of material which might be considered offensive by some – or at least not representative of them as individuals – is understood as important for society at large and for history. This requires careful contextualisation, which can bring challenges for web archives, particularly at scale. The live web is complex and multi-layered, and this complexity only increases as it is archived and republished.

Context is one of the first casualties of keyword searching, which is the primary form of access to most web archive collections: "it is already evident that we need to move away from a search-oriented approach towards one that reflects classic archival methods, with an emphasis on hierarchy and context" (Winters & Prescott, 2019: 393). Researchers and archivists are working together to address these questions, and to ensure that the required context is both captured and made available in an accessible and readily legible form. The macro-level context provided by an organisation like the IIPC is a useful starting point, but attention needs to be paid to the many different levels of the archived web identified by Brügger (2009): the web as a whole, the web sphere, the web page (and even individual web elements). This is not an easy task, but it is an important one.

[14] The mention of trust is instructive, and a characteristic of good governance that we would add to the UN framework. We will return to this question below.

[15] https://www.webarchive.org.uk/en/ukwa/info/nominate

2.3 Accountable

What does accountability mean for web archives? To whom are they accountable, and for what aspects of their work? What are the mechanisms and standards required to ensure accountability? Zumofen (2015: 1–2) notes that, while accountability is increasingly prominent in contemporary debates, "its sense remains elusive"; "Sometimes defined as a mechanism, accountability could also be a virtue, a social relation, a function or sometimes it is just used as the synonym of transparency." The number of stakeholders in web archives complicates accountability further. At a fundamental level, most web archiving institutions are accountable to their funders, according to more or less metrics-based criteria. Web archives themselves will most likely be only one activity within a library or archive, and will have to account for their performance, successes and failures alongside other departments and initiatives. In many instances, however, it is not clear what successful performance actually looks like for web archives. One key metric, which is often publicly documented, is size (x terabytes of data added over the course of a year), but this is in many ways something which is beyond the influence of the archive. The amount of data to be harvested is dependent on the growth of the live web, of a country code Top Level Domain, during an arbitrary period of time or even in some institutions of the budget allowed for crawling. A more useful, although still problematic criterion, is levels of usage. The IA, for example, publishes detailed server and archive statistics, which reveal that 191,727 new users were added in August 2020, and even graphs the number of page views per second.[16] Usage information, however, has little relevance for accountability for web archives that are able to offer only limited user access because of restrictions arising from legal deposit legislation.

Other mechanisms that might be considered include user groups, the provision of enhanced documentation (both on- and off-line), improved public access, openness about processes and quality assurance, and so on. Many web archives already do some or all of this, but web archiving is a costly activity, and with cost come ever greater requirements for accountability.

A far more intractable question is how users can hold accountable those commercial digital platforms and services which are archives by default, like Facebook, Instagram, YouTube and so many others. The lack of accountability here is abundantly clear. We have already mentioned the two weeks' notice of content removal given by Tumblr to its users in 2017, but there are numerous other examples of digital data loss and/or destruction. One of the most well documented is the closure of GeoCities by Yahoo! in 2009 (and the subsequent rescue efforts of the IA and Archive Team) (Milligan, 2017), but there are others, such as the "sunsetting" of FriendsReunited in February 2016 (Pankhurst, 2016) or the huge loss of data from MySpace that occurred following "a server migration project" (Chokshi, 2019). The terms and conditions of use for such platforms ensure that there is no real accountability for commercial services, which are, but should not be, treated as de facto web archives.

[16] User statistics https://archive.org/about/stats.php; Pageview stats https://analytics0.archive.org/stats/pageviews.php

2.4 Transparent

Transparency is closely connected with trust, accountability and openness, and this is an area where huge progress has been made in recent years. Several archiving institutions, for example, regularly publish lists of crawled URLs as open data, so that users can explore what has and has not been included in their web archives. The UK Web Archive publishes its code openly, alongside a range of tools for working with the archived web, via GitHub.[17] But this is not the whole story. Web archives are surprisingly poor at consistently documenting their own histories and practices. It is not an easy task to discover key dates or to piece together the history of technical and organisational change that has helped to determine the current state of a web archive. It may be possible to establish the size of a web archive at the time one is using it (many provide a prominent rolling tally of the number of web pages indexed), but what was the size of the archive in 2010? Sometimes even the current situation can only be inferred. At the time of writing, the most up-to-date public information for the UK Web Archive is that "As of 2017 we have collected approximately 500 TB of data and [are] increasing this by over roughly 60–70 TB a year."[18] Some of the gaps can be filled by looking through the UKWA's very informative blog, but the picture is by no means comprehensive.[19] It is much easier to establish this kind of information for the commercial platforms and services that we have already identified as generally lacking in accountability, hinting at the difference between these two concepts. A notable exception was the UK Government Web Archive, based at The National Archives of the UK. For several years a detailed account of the UKGWA's history was published on its website, but some time before April 2018 it was removed from the live web and is only accessible via the archive itself (Winters, 2019: 85 n. 10). The community of web archivists and researchers is currently sufficiently small that some of the background knowledge essential for the interpretation of web archives can be gleaned simply by speaking to the archivists involved in the creation of these vital resources, but this will not be the case forever.

Moreover, researchers are not the only community of users for whom transparent data collection processes are important. How do web archives meet the challenge of ensuring that "information is freely available and directly accessible to those who will be affected by [it]"?[20] At present, it is simply not possible to know in any detail what kinds of information about which people are captured in web archives. There is clear conflict here too with the requirements of legal deposit legislation in some countries, which may mandate that the data collected be kept closed, or access limited to particular groups.

[17] UK Web Archive repositories https://github.com/ukwa

[18] UK Web Archive FAQs https://www.webarchive.org.uk/en/ukwa/info/faq

[19] https://britishlibrary.typepad.co.uk/webarchive/?_ga=2.228305567.388124317.1598974250-519666553.1598256809

[20] https://www.unescap.org/sites/default/files/good-governance.pdf

2.5 Responsive

In one sense, responsiveness is the default position of most web archives. They have to be ready to archive digital responses to sudden and often catastrophic events, such as natural disasters or the COVID-19 pandemic. These events are not bounded by our normal working routines, but unfold at inconvenient hours, across different time zones and in multiple languages. We have already mentioned the ASAP project, which has documented just what this real-time collection involves (see, for example, the interview with T. Drugeon at the French Institut national de l'audiovisuel (INA), which collected more than 30 million tweets related to the series of terrorist attacks that affected France in 2015).[21]

Web archives also have to be responsive to changing technology and digital platform updates, which may render a previously effective crawling process unreliable, or even prevent a website from being archived at all. Platforms change continually: some of those changes are obvious – the doubling of characters in a tweet or the addition of warning labels to indicate misleading content – but others are both much less apparent and just as potentially problematic for archiving processes.

This level of responsiveness is impressive, but even more remarkable given the small size of the teams involved and the relatively low of level of funding provided to web archives of all kinds. The web archives based in national memory institutions are usually staffed by a handful of people, who are competing internally for limited resources. Increased investment in web archives, as custodians of the world's digital cultural heritage, is essential if greater responsiveness is to be delivered.

2.6 Effective and efficient

We have already established that web archives are extraordinarily efficient given the limited resources at their disposal. Under the framework that we are making use of here, however, "The concept of efficiency in the context of good governance also covers the sustainable use of natural resources and the protection of the environment".[22] This is not a question that faces web archives alone, but it is crucial to the future of the whole field of digital preservation: "Digital preservation relies on technological infrastructure … that has considerable negative environmental impacts, which in turn threaten the very organizations tasked with preserving digital content" (Pendergrass et al., 2019). Collective action, and a willingness to rethink decades-old practice, will be required (Pendergrass et al. argue for "shifting cultural heritage professionals' paradigm of appraisal, permanence, and availability of digital content"). Radical steps will need to be taken at the societal level, but web archives and digital preservation specialists can make an important contribution.

[21] https://asap.hypotheses.org/173

[22] https://www.unescap.org/sites/default/files/good-governance.pdf

2.7 Equitable and inclusive

Questions of equality, diversity and inclusion rose to global prominence during 2020, thanks to the combination of inequalities brought into sharp focus by the COVID-19 pandemic and the increased prominence of the Black Lives Matter movement. Web archives include information for and about diverse groups in society and hold out the promise of preserving the voices of individuals who in previous centuries would only have featured in an archive if they interacted with the church, central government or the law. But this does not mean that web archives are equitable and inclusive by default. As we have already described, inequalities of access are built in to the legal frameworks governing web archiving in many countries. Schostag (2020) notes, for example, that "In accordance with the Danish personal data protection law, the public has no access to the archived web material. Only researchers affiliated with Danish research institutions can apply for access in connection with specific research projects". Where web archives are open to everyone but can only be consulted in the physical premises of a national library, there are people for whom the costs of travel will be prohibitive. As libraries closed during the COVID-19 pandemic, access was closed off to everyone.

With regards to inclusiveness, there have been some notable efforts to diversify special collections, so that web archives become visibly more inclusive. In the UK Web Archive, there are special collections concerned with 'Black and Asian Britain', 'Caribbean communities in the UK', 'Gender equality', 'LGBTA + lives online', 'Muslims, trust and cultural dialogue', 'Russia in the UK', 'Latin America UK', etc. Most of these special collections have been curated by library staff, but community collecting practices are beginning to be incorporated. The 'Latin America UK' collection, for example, has been "produced by Latin American communities in the UK or by UK organisations with direct links to these communities and to the region".[23] These are small steps, but they indicate an important direction of travel for web archives. And it is a direction which brings not only opportunities, but also ethical challenges. As Lomborg (2019) notes, web archives and researchers have "an obligation to reflect upon whose stories we are telling, to what extent we are equipped to tell their story, and what kinds of vulnerability and harm we might encounter and nurture when doing it".

2.8 Follows the rule of law

Ethical practices are essential, but they are no substitute for enlightened legal frameworks and protection. In most countries, web archiving is heavily regulated, subject to various combinations of legal deposit and copyright law. In some instances, the relevant law(s) explicitly address the archiving of born-digital content, as in France, Luxembourg and the UK following the extension of earlier legal deposit legislation to include new digital formats. This is both enabling (it allows collection and preservation) and constraining (it can restrict access) (Winters, 2021). In other

[23] https://www.webarchive.org.uk/en/ukwa/collection/2384

countries, it is copyright and intellectual property law that prevails, ensuring that such web archiving remains selective and permissions based. Following the rule of law is a vital protection for web archives, and for society, but the failure of national and international law to keep pace with technological change can negatively affect some of the other criteria for good governance that we have outlined above. Faced with restrictions to access, rather than using national web archives, researchers, and even more the public at large, may be tempted to rely on the IA as the sole source for retrieving web archives, while missing more complete material preserved within other archiving institutions. Indeed, the IA places no restrictions on online access, thanks to a copyright policy based on removing content or disabling access to content infringing copyright or other intellectual property rights "in appropriate circumstances and at its discretion", in response to well-evidenced take-down requests.[24] There is certainly a balance to be struck, which retains safeguards for citizens (the IA recognizes: "We collect Web pages that are publicly accessible. These may include pages with personal information")[25] but also increases access and gives web archives greater room to experiment and innovate.

3 Conclusion: Towards multi-stakeholder and multi-scale engagement

Web archives are part of a complex and rapidly evolving social, political, technical and legal ecosystem. As previously underlined, they require increased access (Winters, 2021), but the transparency, accountability, reproducibility, sustainability and legibility of these complex knowledge infrastructures and born-digital heritage are key for their current as well as future use. The progress made in just over two decades is remarkable and many initiatives have already been highlighted in terms of collaboration, awareness-raising, coordination and responsiveness. However, web archives are still underused compared to their huge potential. How can they engage the academic world as well as stakeholders like IT companies, journalists and the wider public in a time of widespread misinformation?

The issue here is not only that of preservation, even if it is also urgent to consider the increasing digital divide between areas with "national" web archives and parts of the world where preservation is based on IA efforts, with all the gaps that may result from archiving processes which are not necessarily based on a detailed knowledge of the local web space (although the case of the IA's archiving of the North Korean Web (Ben-David & Amram, 2018) shows an ultimately very effective curation network). The question is not only one of being able to access the resources but also of being able to understand them, which leads on to wider concerns about digital literacy to allow both technical and cognitive understanding of these archives (Bachimont, 2017): they require comprehension of both their shaping and their contextualization (Schafer & Thierry, 2018).

[24] https://help.archive.org/hc/en-us/articles/360004716091-Wayback-Machine-General-Information
[25] Idem.

How can we create news forms of literacy, teaching and engagement around web archives? As explained by Milligan et al. (2019: 168), based on their experience of datathons: "This is an educational model that differs from the traditional classroom mode of lab and lecture". As Brügger (2018) has shown, born-digital heritage is more akin to a reborn-digital heritage, which invites us also to rethink the framework of digital hermeneutics applied to digital heritage and the user experience (Fickers, 2020). While web archives are becoming records of multiple human activities, their legal and political uses, as well as the ethical and philosophical issues related to the right to oblivion and the right to be remembered, invite us to think about questions of proof, authenticity and traceability in an interdisciplinary manner.

Web archives will also have to respond to new forms of engagement, publication and public history. How can we take into account and value the creation of tools, data publications, and other activities that are sometimes considered as trivial and ancillary but remain time-consuming and highly valuable for academic research and, more broadly, for the entire community interested in web archives? How can we better recognize the collective in research initiatives and multi-stakeholder participation in public history, which is sometimes still too asymmetrical?

The concept of "sharing zones" in academia also applies in the web archiving world, notably with regards to gateways and interoperability: transnational metadata and derivative data repositories are probably the new frontier that has to be reached in order to ensure that web archives meet researchers' needs. This would help users to move from one archive to another, whether remotely or physically, and allow for more comprehensive documentation of collections as well as interfaces, tools, etc. A common portal could help researchers to enter more easily the heterogeneous world of web archiving. Artificial Intelligence will undoubtedly help to move the field forward, but it must not be used uncritically. Machine learning approaches that improve traceability, for example, may also be co-opted by those who are interested in predicting individual and collective behaviour, with the result that web archives may become an arena for future social struggles. Reinforcing "good governance" in web archives may help to address these controversies.

All of this is a goal towards which web archives can work in the medium term rather than something which can be delivered overnight. Some important steps to be taken along the path could be the development of committees composed of multi-stakeholders to assess "good governance", the sharing of curricula dedicated to web archives, increased open and regular dialogue gathering representatives of all stakeholders, and the development of engagement by design, which means lowering the barriers to access, as Milligan (2020) has suggested.

Web archives are at a crossroads, with contradictory instructions about which path to take, but they cannot be the sole "property" of experts. In 2017 the question related to web archives was "how can value be demonstrated to the wider general public?" (Winters, 2017); in 2021 it is time to provide an answer to a second one: "how can *values* be demonstrated to the wider general public?".

Data availability Not applicable

Code availability Not applicable

Declarations

Conflicts of interest Not applicable

References

Ankerson, M. (2015). Take me back! Web history as chronotourism of the digital archive. "Times and Temporalities of the Web" International symposium, Paris.

Bachimont, B. (2017). *Patrimoine et numérique. Technique et politique de la mémoire*. Ina Editions.

Barats, C., Schafer, V., & Fickers, A. (2020). Fading Away… The challenge of sustainability in digital studies. *Digital Humanities Quaterly,* 14 (3). http://www.digitalhumanities.org/dhq/vol/14/3/000484/000484.html.

Ben-David, A., & Amram, A. (2018). The internet archive and the socio-technical construction of historical facts. *Internet Histories, 2*(1–2), 179–201.

Bingham, N. (2015). Ten years of archiving the Web. *UK Web Archive Blog*. https://blogs.bl.uk/webarchive/2015/06/ten-years-of-archiving-the-web.html.

Brügger, N. (2018). *The archived web. Doing history in the digital age*. MIT Press.

Brügger, N. (2009). Website history and the website as an object of study. *New Media and Society, 11*(1–2), 115–132.

Bruns, A. (2018). The library of congress: A failure of historic proportions. *Medium.com.* https://urlz.fr/dKwg.

Captain, S. (2018). The frantic, unprecedented race to save 700,000 NSFW Tumblrs for posterity. *Fast Company* https://urlz.fr/dKwi.

Charlton, T. (2017). The treachery of archives: Representation, power, and the urgency for self-reflexivity in archival arrangement and description. *The iJournal, 3*(1), 1–8.

Chokshi, N. (2019). Myspace, once the king of social networks, lost years of data from its heyday. *New York Times*. https://www.nytimes.com/2019/03/19/business/myspace-user-data.html.

Conger, K. (2016). Backing up the history of the Internet in Canada to save it from Trump. *Techcrunch.com*. https://urlz.fr/dKkI.

Els, B., & Schafer, V. (2020). Exploring special web archive collections related to COVID-19: The case of the BnL. *WARCnet Paper*. Aarhus: University of Aarhus. https://cc.au.dk/fileadmin/user_upload/WARCnet/Schafer_et_al_COVID-19_BnL.pdf.

Fickers, A. (2020). Update für die Hermeneutik. Geschichtswissenschaft auf dem Weg zur digitalen Forensik? *Zeithistorische Forschungen/studies in Contemporary History, 17*, 157–168.

Geeraert, F. (2020). The PROMISE of a Belgian web archive. *IIPC Blog*. https://netpreserveblog.wordpress.com/2020/02/06/the-promise-of-a-belgian-web-archive/.

Gomes, D. Miranda, J., & Costa, M. (2011). A Survey on Web Archiving Initiatives. *TPDL 2011*, 408–420. Berlin/Heidelberg/New York: Springer.

Kahle, B. (2007). Universal access to all knowledge. *The American Archivist, 70*, 23–31.

Lomborg, S. (2019). Ethical considerations for web archives and web history research. In N. Brügger & I. Milligan (Eds.), *The SAGE handbook of web history* (pp. 99–111). Sage Publications Ltd.

Masanes, J. (2006). *Web archiving*. Springer.

Milligan, I. (2017). Welcome to the web: The online community of GeoCities during the early years of the World Wide Web. In N. Brügger & R. Schroeder (Eds.), *The web as history: Using web archives to understand the past and the present* (pp. 137–158). UCL Press.

Milligan, I. et al. (2019). Building Community and Tools for Analyzing Web Archives through Datathons. *ACM/IEEE Joint Conference on Digital Libraries* (JCDL). https://yorkspace.library.yorku.ca/xmlui/bitstream/handle/10315/36180/datathon.pdf?sequence=1&isAllowed=y.

Milligan, I. (2020). You shouldn't Need to be a Web Historian to Use Web Archives: Lowering Barriers to Access Through Community and Infrastructure. *WARCnet paper*. Aarhus: University of Aarhus. https://cc.au.dk/fileadmin/user_upload/WARCnet/Milligan_You_shouldn_t_Need_to_be.pdf.

Moiraghi, E. (2018). Le projet Corpus et ses publics potentiels. Une étude prospective sur les besoins et les attentes des futurs usagers. BnF.

Mons, B. (2018). *Data stewardship for open science. Implementing FAIR principles*. CRC Press.
Musiani, F., Paloque-Bergès, C., Schafer, V., & Thierry, B. (2019). *Qu'est-ce qu'une archive du Web?*. OpenEdition Press. https://books.openedition.org/oep/8713.
Musiani, F., & Schafer, V. (2017). *Do web archives have politics?* RESAW conference.
Pankhurst, S. (2016). FriendsReunited – the sunset of an era. *Medium.* https://medium.com/@liife/friendsreunited-the-sunset-of-an-era-3e5b2ea7bb11.
Pendergrass, K. L., Sampson, W., Walsh, T., & Alagna, L. (2019). Toward environmentally sustainable digital preservation. *The American Archivist, 82*(1), 165–206.
Piguet, M., & Montebello, C. (2020). Covid-19: pour une mémoire ordinaire de l'extraordinaire. *Libération*. https://www.liberation.fr/debats/2020/04/25/covid-19-pour-une-memoire-ordinaire-de-l-extraordinaire_1786299.
Rhodes, T. (2013). A Living, Breathing Revolution: How Libraries Can Use "Living Archives" to Support, Engage, and Document Social Movements. Singapour, IFLA WLIC.
Rollason-Cass, S., & Reed, S. (2015). Living movements, living archives: Selecting and archiving web content during times of social unrest. *New Review of Information Networking, 2*(2), 241–247.
Schafer, V., & Thierry, B. (2018). Web history in context. In N. Brügger & I. Milligan (Eds.), *The SAGE handbook of web history* (pp. 59–72). Sage Publications Ltd.
Schostag, S. (2020). The Danish coronavirus web collection – coronavirus on the curators' minds. *International Internet Preservation Consortium Blog*. https://netpreserveblog.wordpress.com/2020/07/29/the-danish-coronavirus-web-collection/.
Taylor, N. (2015). Questions of ethics at Web Archives 2015. Stanford Libraries website. http://library.stanford.edu/blogs/digital-library-blog/2015/12/questions-ethics-web-archives-2015.
Teszelszky, K. (2019). Web archaeology in The Netherlands: The selection and harvest of the Dutch web incunables of provider Euronet (1994–2000). *Internet Histories, 3*(2), 180–194.
Webber, J. (2017). A new (beta) interface for the UK Web Archive. *UK Web Archive Blog*. https://blogs.bl.uk/webarchive/2017/12/a-new-beta-interface-for-the-uk-web-archive.html.
Webster, P. (2017). Users, technologies, organisations: Towards a cultural history of world Web archiving. In N. Brügger (Ed.), *Web 25: Histories from 25 years of the World Wide Web* (pp. 179–190). Peter Lang.
Wilkinson, M., Dumontier, M., Aalbersberg, I., et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data,* 3.
Winner, L. (1980). Do artifacts have politics? *Daedalus, 109*(1), 121–136.
Winters, J. (2021). Giving with one click, taking with the other: E-legal deposit, web archives and researcher access. In P. Gooding & M. Terra (Eds.), *Electronic legal deposit: Shaping the library collections of the future.* Facet Publishing.
Winters, J. (2019). Negotiating the archives of UK web space. In N. Brügger & D. Laursen (Eds.), *The historical web and digital humanities: The case of national web domains* (pp. 75–88). Routledge.
Winters, J., & Prescott, A. (2019). Negotiating the born-digital: A problem of search. *Archives and Manuscripts, 47*(3), 391–403.
Winters, J. (2017). Breaking in to the mainstream: Demonstrating the value of internet (and web) histories. *Internet Histories, 1*(1–2), 173–179.
Zumofen, R. (2015). Redefining accountability in a strategic perspective to enhance performance. *International Research Society for Public Management*, 1–20.