BIG DATA & SOCIETY

Commentary

Archival strategies for contemporary collecting in a world of big data: Challenges and opportunities with curating the UK web archive

Big Data & Society
January–June: 1–6
© The Author(s) 2021
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/2053951721990409
journals.sagepub.com/home/bds
SAGE

Nicola Jayne Bingham[1] and Helena Byrne[2]

Abstract

In this contribution, we will discuss the opportunities and challenges arising from memory institutions' need to redefine their archival strategies for contemporary collecting in a world of big data. We will reflect on this topic by critically examining the case study of the UK Web Archive, which is made up of the six UK Legal Deposit Libraries: the British Library, National Library of Scotland, National Library of Wales, Bodleian Libraries Oxford, Cambridge University Library and Trinity College Dublin. The UK Web Archive aims to archive, preserve and give access to the UK web space. This is achieved through an annual domain crawl, first undertaken in 2013, in addition to more frequent crawls of key websites and specially curated collections which date back as far as 2005. These collections reflect important aspects of British culture and events that shape society. This commentary will explore a number of questions including: what heritage is captured and what heritage is instead neglected by the UK Web archive? What heritage is created in the form of new data and what are its properties? What are the ethical issues that memory institutions face when developing these web archiving practices? What transformations are required to overcome such challenges and what institutional futures can we envisage?

Keywords

Web archiving, big data, ethics, legal deposit, researcher access, heritage

Introduction

In this commentary, the authors address questions of heritage preserved and heritage lost in the UK Web Archive (UKWA) utilising definitions of heritage proposed by Bonacchi and Krzyzanska who offer a characterisation of digital heritage as interactions enabled 'by the Internet and the outcomes of such processes' (the footprints – including data – that are produced) and consider 'heritage production to be any activity, occurring online or offline, as part of which human or non-human engage with the past more or less incidentally' (2019). Our reflections will focus on the case study of UKWA and will exclude a full exploration of the archive's position in the wider landscape of archival practice. This is out of the scope in the context of this commentary, and a comprehensive account of diverse global approaches to web archiving has already been provided by Webster (2019).

The foundations of UKWA were established in 2002 following a report commissioned by the Wellcome Trust and the Joint Information Systems Committee (JISC) on the feasibility of creating a UK web archiving service (Wellcome Library, 2020). This led to the formation of the UK Web Archive Consortium (UKWAC) in 2003, which consisted of the British Library, National Library of Scotland, National Library of Wales, The National Archives, JISC and the Wellcome Trust, against the background of the Legal Deposit Libraries Act 2003, which provided the legislative framework for the six UK Legal Deposit Libraries (LDLs) to archive web content with permission from the copyright holders (UK Web Archive

[1] British Library, Wetherby, UK
[2] British Library, London, UK

Corresponding author:
Nicola Jayne Bingham, British Library, Boston Spa Building 26, Wetherby LS23 7BQ, UK.
Email: nicola.bingham@bl.uk

FAQ 2020). With the full enactment of Non-Print Legal Deposit Regulations (NPLD) in 2013, the LDLs came together, as UKWA (Meilke, 2013), to collect content published on the UK web reflecting important aspects of British culture and events that shape society (Pennock, 2011).

What heritage is captured and created?

The terms of NPLD have significant bearing on the heritage acquired by UKWA, for example, the Libraries are only permitted to archive an 'on line work published in the United Kingdom'. (The Legal Deposit Libraries (Non-Print Works) Regulations, 2013). In practice, this is applied by restricting the archival web crawler (a software application or bot which runs over the internet to fetch, analyse and file information from web servers) to websites on UK top level domain names (.uk, .scot., .cymru etc.) and/or websites hosted on servers physically located in the UK. This results in significant gaps in the heritage acquired as websites on 'non-UK' top level domain names, such as .com, are not automatically identified. This is particularly challenging with the acquisition of social media which is usually hosted on web domains outside of the UK (facebook.com, twitter.com). Web archivists may acquire social media accounts of named individuals/organisations yet must exclude the replies and comments, as the provenance is difficult to ascertain especially when working at scale. For the same reason, UKWA rarely archives Twitter hashtags, which potentially limits the heritage preserved to those individuals and organisations that represent the more official voice.

In addition to the legislative environment, myriad technical factors influence UKWA's ability to collect and provide access to online heritage. Not all content on the web can be archived, as it is ephemeral with web pages lasting on average between 75 and 100 days (Brown, 2006: 3). The web is a constantly shifting, fluid space requiring bespoke technologies to archive it, the funding for which is often not available to public funded or smaller web archiving institutions.

There are two main acquisition strategies for web archives; whole top-level domain archiving and selective archiving, many [archives] do both and some focus on one of these strategies (Pennock, 2013b: 10). UKWA employs both a whole domain collection strategy together with various selective approaches. Along with the majority of web archives undertaking large-scale web archiving, UKWA employs the Heritrix web crawler developed by the Internet Archive. Heritrix follows a list of starting seeds, or URLs and requests content from the hosting server. It attempts to capture websites comprehensively and to reproduce the 'look and feel' of the website but is not capable, in all cases, of archiving JavaScripts, dynamic webpages, web content retrieved by querying a database or streaming video/audio (Brügger, 2018: 88).

Websites can be more or less archivable depending on the platform or underlying technology used to produce and serve the website. Archiving proprietary software is a major challenge. An example is Facebook which deliberately blocks web crawlers and has proved impossible to archive (Grimshaw, 2020). Wix (a software company, providing cloud-based web development services) is another platform which is technically challenging to archive due to a combination of factors including the use of content delivery networks (CDNs), geographically distributed servers which are difficult for the crawler to reach and are not often in scope. Objects such as comments appended to online news articles or blogs are particularly elusive to web archiving technologies, as they are often loaded into the browser dynamically, e.g., with Ajax functionality, which is beyond the reach of most automated web crawlers. Combined and at scale, these technical limitations mean that the heritage created by the web archiving process disproportionately represents the more official voice, or those individuals and organisations that have the funding and technical infrastructure to produce their own websites. Free platforms such as Facebook, Twitter, Wix etc. are utilised more often by private individuals and small community groups meaning that these are more likely to be excluded from the archive.

To address some of the technical challenges associated with archiving dynamic websites, UKWA has made exploratory steps with the Webrecorder app and, the hosted browser-based version, Conifer, which work by recording user-driven browsing sessions in real time, enabling archivists to achieve very high fidelity captures of web content, including social media (UK Web Archive, 2020a). In theory, Webrecorder offers potential for collecting and creating much more heritage; in practice however, 'recording' websites is a manual, extremely time-consuming process and can only be used very selectively due to resource constraints (Bingham et al., 2020).

In addition to whole domain web crawling, which captures the 'big picture', UKWA employs selective approaches to fulfil collecting purposes, including curated collections or 'topics and themes' which fall into three broad categories: thematic, event-based and rapid response (UK Web Archive, 2020b). Collections are driven by 'curators' comprising library, archive and reference specialists, academic researchers or other partners external to the LDLs. Curators or 'collectors have substantial epistemological weight, because they decide what material should be gathered and in what form' (Bonacchi and Krzyzanska, 2019). Some of UKWA's academic partners show a preference towards selecting content 'offering the most scholarly and verifiable information' related to their chosen subject while other curators focus on selecting content from ordinary people that are involved in grassroots campaigns or online hobbies (Huc-Hepher, 2015: 5). Staff at the LDLs often follow a similar strategy to subject experts at the Bibliothèque nationale de France (BnF) as described by Milligan (2019). Key considerations include that the material is 'public' – so that it can be truly considered a 'publication' under the auspices of legal deposit (Milligan, 2019: 164–165).

To help mitigate potential biases in the institutional or 'top down' approach and to ensure that the heritage created is as diverse as possible, UKWA aims to widen participation in the curation process as widely as possible. Access is made available to the Annotation Curation Tool (ACT), bespoke software which enables any logged-in user to curate web content (Bingham, 2015). However, access is usually restricted to LDL staff or trusted partners to help ensure compliance with the regulations.

Webrecorder/Conifer has great potential to democratise the web archiving process as websites archived by individuals external to the LDLs can be added to UKWA, creating possibilities for more diversity within the archive. In addition, the Save a UK Website facility on UKWA website allows any member of the public to nominate websites for preservation in the archive.

What are the ethical issues?

Despite the scarcity of literature or documented practice to support web archivists in ethical collection management (Dougherty, 2003: 1; Graham, 2017: 103; Lomborg, 2019: 2), the community has started to engage in productive dialogue which could lead to a conceptual framework in the future. The National Forum on Ethics and Archiving the Web (EAW) held at the New Museum in New York City in 2018 (Rhizome) focussed on 'Opportunities and challenges in ethics and web archiving', for example. Several guideline documents aiming to increase transparency in the web archiving process have been produced by individual organisations such as the Documenting the Now project which has developed a tool to integrate and document the choices made by archivists, librarians and researchers in the archiving process (Dolan-Mescal, 2017). Social Feed Manager developed by George Washington University Libraries is accompanied by one of the few sets of guidelines that exist for ethics and privacy in social media research (George Washington University Libraries, 2018).

Web archiving is a complex, obscure and largely dark process. The curatorial and collection activities the researcher would ordinarily engage with to understand the affordances of the archive are much less accessible than for traditional collections. This is even more apparent with large, domain scale collections such as UKWA when compared to web archives that are narrower in size and scope. Questions about inclusion, exclusion, acquisition, description, access and authenticity are more challenging to unpick due to myriad layers of technicity in the archiving process, the legislative environment and the sheer volume of data. Whilst this information is not deliberately obscured, it is contained in disparate systems and is difficult to interpret without unfettered access to archive staff, crawl logs and an understanding of the organisational and legal environment in which the collection has been created.

UKWA strives to build transparencies into its processes, for example by making collection scoping documents available to the public in the British Library Research Repository. Furthermore, efforts are made to communicate to audiences through various channels (the website, blogs, conferences and papers). In addition, for the small number of researchers who have worked with UKWA in a big data capacity (Basile and Mcgillivray, 2018; Tranos, 2018), tailored orientations to the archive have been facilitated by the UKWA team.

UKWA is a national collection and as such aims to be as comprehensive, inclusive and representative as possible, within the technical and legislative frameworks. Key collection principles include that material is collected without pre-selection or filtering; information is not endorsed or censored; nothing is deleted or redacted unless it is found to be illegal. Automated, whole domain archiving mitigates to some extent, curatorial bias in acquisition; however, for various reasons, automated crawling is far from representative or inclusive and in fact amplifies inequalities on the web, which is not a neutral space in the first place (Noble, 2018).

A major responsibility archives uphold is to collect and provide access to materials that represent diverse populations (de Klerk, 2018). Harrison et al. discuss (in the context of analogue collections) how 'diversity' has emerged historically as a normative conservation target; yet, there are a range of ways in which diversity is created, understood and preserved by cultural heritage institutions (2020: 466).

Web archiving represents a unique opportunity to represent communities that may be marginalised for various reasons. As Lomborg suggests, 'there is a democratising potential but an increased ethical responsibility' in web archiving (2019: 6). When building community-centred collections, UKWA attempts to engage with those communities that are the subject of focus. Recent examples include the LGBTQ+ Collection, which was led by the CILIP LGBTQ+ Network and the British Library

LGBTQ+ Staff Network and later opened up to the wider LGBTQ community to submit selections UK Web Archive (2020b). Engagement with the LGBTQ community was handled sensitively, for example in allowing individuals to submit websites anonymously and casting a curatorial eye over any content that was submitted.

Another example is UKWA's partnership with the Boredom Project, a collaboration between Anglia Ruskin University, Chelmsford Museum's Creatives network and the British Science Association. Their 'Young Creatives' project enables young people in Essex to share experiences of the Covid-19 pandemic through poetry, photography, paintings, short stories, doodles, etc. (Woolman, 2020). The group have helped to co-curate the Coronavirus collection by submitting websites for archiving, which adds an authentic and cognizant dimension to the collection.

Ethical considerations around description and cataloguing are not new to web archives; different actors will have different values and identities that the archivist must consider when ascribing meaning. Two collections in UKWA illustrate this point: in 2005 subject specialists debated whether the London Terrorist Attack 7 July 2005 collection should be described as such given that 'terrorist' is a loaded term, before concluding that this phrasing did accurately describe the events of July 2005. The second example is presented by Huc-Hepher, who examines considerations involved in assigning labels of 'community' in the London French Special Collection UK Web Archive (2020b) Topics and themes.

The web has facilitated an unprecedented upheaval in publishing leading to a democratisation of the creative and publishing process. Although UKWA does not collect private data, data that is personal is included, if it is published, due to researcher demand. However, this raises several ethical difficulties. The question of what is public and what is private on the web is complex, raising ethical concerns about future access and potential harm (Lomborg, 2019: 2). The private individual posting to a blog, Twitter or commenting on a news article may not understand this act as 'publishing' and most likely will not have an expectation that this content will be preserved for the public record; yet, this is precisely the mission of organisations such as UKWA (Smith and Cooke, 2018: 2).

Graham raises further important questions in the discussion of ethics in web archiving: 'what about collecting material that is morally questionable but nevertheless important to document (hate speech/bigotry/far right groups)? does the value of collecting this material outweigh the rights' of those users to privacy – or does collecting this material give a platform to these messages and amplify them' (2017: 107)? The size and scale of web archive collections exacerbate the problem as the contents of the collection are largely unknown yet at the same time indexing (particularly free text indexing) potentially makes named individuals easier to discover. This was an important consideration in archiving recent anti-racism protests in the UK, for example, as activists could be involuntarily exposed or left vulnerable to persecution or endangerment through their online presence.

Conclusion: Institutional futures

This commentary has illustrated the opportunities and challenges involved in managing a colossal, noisy and unstructured data set. It has looked at heritage created and heritage lost by the archival process and has discussed some of the ethical issues involved in web archiving.

Whilst the UKWA aims to be as comprehensive as possible, it is evident that the range of web content included is constrained by the UKWA's legislative framework and the technical processes used to carry out archiving. Only websites published on a narrow definition of the UK web domain can be archived at scale and, even then, websites cannot always be archived completely due to issues with archiving dynamic web technologies, leading to gaps in the heritage created. We have also seen that interactions on the web that might be seen to represent the 'ordinary people' in a community, country, society or organization are not as well represented in the archive as the more official voice. Despite the huge size of the UKWA, it can probably be best described as representative at best. Having said this, UKWA is actively pursuing co-curation, by reaching out to disparate agencies and individuals and by utilising new software to help increase the type and variety of heritage created and preserved in the archive.

Despite technical and legislative barriers to access, UKWA affords an exciting opportunity to explore the history of the mid-1990s onwards. Although there is a growing literature in this area (Basile and Mcgillivray, 2018; Tranos, 2018; Ben-David and Amram, 2018), web archives remain an area of study which is under exploited. UKWA is building on work already begun in engaging with the researcher community through its involvement with networks such as RESAW (a Research infrastructure for the Study of Archived Web materials) and WARCnet (Web ARChive studies network researching web domains and events) (RESAW, 2012; WARCnet, 2020). Work such as the recent project, 'asking questions with web archives – introductory notebooks for historians'[1] aims to enable researchers to explore and analyse web archives without needing advanced coding skills, which often prove a barrier to working with web archives. The opening up of UKWA to interrogation by researchers and a participatory curation process are perhaps helping to shape a more democratic, open and inclusive institutional future.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

ORCID iDs

Nicola Jayne Bingham https://orcid.org/0000-0002-5510-9869
Helena Byrne https://orcid.org/0000-0002-0966-4685

Note

[1] The project is supported by the International Internet Preservation Consortium (IIPC) and led by Andrew Jackson, UKWA Technical Lead and Tim Sherratt (Sherratt, 2020).

References

Basile P and Mcgillivray B (2018) Exploiting the web for semantic change detection. In: Soldatova L, Vanschoren J, Papadopoulos G, Ceci M (eds) *International Conference on Discovery Science*. Berlin: Springer, pp.194–208.

Ben-David A and Amram A (2018) Computational methods for web history. In: Brügger N and Milligan I (eds) *The SAGE Handbook of Web History*, pp.153–168. London: Sage.

Bingham N (2015) W3ACT User Guide. Report, British Library, UK. Available at: https://github.com/ukwa/w3act/wiki/W3ACT-User-Guide (accessed 26 May 2020).

Bingham N, Byrne H, Lelkes-Rarugal C, et al. (2020) Using webrecorder to archive UK Party Leaders after the UK General Election 2019. *UK Web Archive Blog*. Available at: https://blogs.bl.uk/webarchive/2020/05/using-webrecorder-to-archive-uk-political-party-leaders-social-media-after-the-uk-general-election-2.html (accessed 29 May 2020).

Bonacchi C and Krzyzanska M (2019) Digital heritage research retheorised: Ontologies and epistemologies in a world of big data. *International Journal of Heritage Studies*. DOI: 10.1080/13527258.2019.1578989. https://www.researchgate.net/publication/330831166_Digital_heritage_research_re-theorised_Ontologies_and_epistemologies_in_a_world_of_big_data

Brown A (2006) *Archiving Websites: A Practical Guide for Information Management Professionals*. London: Facet.

British Library Repository. Available at: https://bl.iro.bl.uk/search?f%5Bcreator_search_sim%5D%5B%5D=UK%20Web%20 Archive (accessed 26 November 2020).

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

de Klerk T (2018) Ethics in archives: decisions in digital archiving. *NC State University Libraries Blog*. Available at: www.lib.ncsu.edu/news/special-collections/ethics-in-archives%3A-decisions-indigital-archiving (accessed 31 January 2020).

Dolan-Mescal A (2017) Opportunities for making appraisal transparent when documenting the now. *Documenting DocNow* (blog), 14 June 2017. Available at: https://news.docnow.io/opportunities-for-making-appraisal-transparent-whendocumenting-the-now-10b807606d39 (accessed 20 May 2020).

Dougherty M (2003) Ethics in/of web archiving. Available at: www.academia.edu/901106/Ethics_in_of_Web_Archiving (accessed 20 February 2020).

George Washington University Libraries and Social Feed Manager (2018) Social media research ethical and privacy guidelines. Available at: https://gwu-libraries.github.io/sfmui/resources/social_media_research_ethical_and_privacy_guid elines.pdf (accessed 20 February 2020).

Graham P (2017) Guest editorial: Reflections on the ethics of web archiving. *Journal of Archival Organization* 14(3–4): 103–110.

Grimshaw J (2020) Web Archiving the UK General Election 2019. *UK Web Archive Blog*. Available at: https://blogs.bl.uk/webarchive/2020/05/web-archiving-the-uk-general-election-2019.html (accessed 26 November 2020).

Harrison R, et al. (2020) *Heritage Futures: Comparative Approaches to Natural and Cultural Heritage Practices*. London: UCL Press.

Huc-Hepher S (2015) Big web data, small focus: An ethno-semiotic approach to culturally themed selective web archiving. *Big Data & Society*. DOI:10.1177/205395171559582. 3. https://journals.sagepub.com/doi/full/10.1177/2053951715595823

Lomborg S (2019) Ethical considerations for web archives and web history research. In: Brügger N and Milligan I (eds) *The SAGE Handbook of Web History*. London: Sage, pp. 99–111

Meilke J (2013) British Library adds billions of webpages and tweets to archive. *The Guardian*, 5 April 2013. Available at: www.theguardian.c; om/technology/2013/apr/05/british-library-archive-webpagestweets (accessed 28 May 2020).

Milligan I (2019) *History in the Age of Abundance: How the Web is Transforming Historical Research*. Canada: McGill-Queen's University Press.

Noble S (2018) Ethics of archiving the web: Algorithms of oppression. Invited Keynote at the New Museum, New

York. Available at: https://eaw.rhizome.org/ (accessed 28 May 2020).

Pennock M (2011) UK web archive #FastFacts: A starter for 10. *UK Web Archive Blog*. Available at: https://blogs.bl.uk/webarchive/2011/11/fastfacts.html (accessed 28 May 2020).

Pennock M (2013a) Public Consultation on non-print legal deposit. *UK Web Archive Blog*. Available at: https://blogs.bl.uk/webarchive/2012/03/consultation-nonprintlegaldeposit.html (accessed 28 May 2020).

Pennock M (2013b) Web-archiving – DPC Technology Watch Report 13 – 01 March 2013, York, UK: Digital Preservation Coalition. Available at: www.dpconline.org/docs/technology-watch-reports/865-dpctw13-01-pdf/file (accessed 25 May 2020).

RESAW (2012) Available at: http://resaw.eu/ (accessed 28 May 2020).

Rhizome (2018) National Forum on ethics & archiving the web. New York, US, 22–24 March, 2018. Available at: https://eaw.rhizome.org/ (accessed 8 May 2020).

Sherratt T (2020) GLAM Workbench Web Archives blogpost. Available at: https://glamworkbench.github.io/web-archives/ (accessed 31 May 2020).

Smith C and Cooke I (2018) Emerging formats: Complex digital media and its impact on the UK legal deposit libraries. *Alexandria: The Journal of National and International Library and Information Issues* 27(3): 175–187.

Tranos E (2018) Web archives: A tool for geographical research? *UK Web Archive blog*. Available at: https://blogs.bl.uk/webarchive/2018/09/web-archives-a-tool-for-geographicalresearch.html (accessed 19 February 2020).

UK Government (2013) The Legal Deposit Libraries (Non-Print Works) Regulations 2013. Available at: www.legislation.gov.uk/ukdsi/2013/9780111533703 (accessed 1 November 2020).

UK Web Archive F.A.Q. Available at: www.webarchive.org.uk/wayback/en/archive/20091130210036/http:/www.webarchive.org.uk/ukwa/info/faq (accessed 29 September 2020).

UK Web Archive (2020a) UK Web Archive Annual Report 2019. Report, British Library, UK,April. Available at: https://bl.iro.bl.uk/work/ns/1f290ae6-1ba8-4642-bb74-5894631a17f2 (accessed 25 November 2020).

UK Web Archive (2020b) Topics and themes. Available at: www.webarchive.org.uk/en/ukwa/collection/202 (accessed 27 May 2020).

WARCnet (2020) Available at: https://cc.au.dk/en/warcnet/ (accessed 28 May 2020).

Webster P (2019) Existing web archives. In: Brügger N and Milligan I (eds) *the SAGE Handbook of Web History*. London: SAGE, pp.30–41.

Wellcome Library. Web archiving: A feasibility study for JISC and the Wellcome Trust. Available at: www.webarchive.org.uk/wayback/archive/20080922220153/http://library.wellcome.ac.uk/node228.html (accessed 28 May 2020).

Woolman A (2020) Introducing: The Boredom Project. British Science Association Blog. Available at: www.britishscienceassociation.org/blog/introducing-the-boredom-project (accessed 26 November 2020).