media/rep/

Repositorium für die Medienwissenschaft

Josh Cowls

Cultures of the UK web

2017

https://doi.org/10.25969/mediarep/12526

Veröffentlichungsversion / published version
Sammelbandbeitrag / collection article

Empfohlene Zitierung / Suggested Citation:
Cowls, Josh: Cultures of the UK web. In: Niels Brügger, Ralph Schroeder (Hg.): *The Web as History. Using Web Archives to Understand the Past and the Present.* London: UCL Press 2017, S. 220–237. DOI: https://doi.org/10.25969/mediarep/12526.

Nutzungsbedingungen:
Dieser Text wird unter einer Creative Commons - Namensnennung 4.0 Lizenz zur Verfügung gestellt. Nähere Auskünfte zu dieser Lizenz finden Sie hier:
https://creativecommons.org/licenses/by/4.0

Terms of use:
This document is made available under a creative commons - Attribution 4.0 License. For more information see:
https://creativecommons.org/licenses/by/4.0

DFG Deutsche Forschungsgemeinschaft

Philipps Universität Marburg

11
Cultures of the UK web

Josh Cowls

Introduction

This chapter reports findings and insights from 'Big UK Domain Data for the Arts and Humanities' (BUDDAH), a project led by the British Library, the Institute for Historical Research at the University of London, and the Oxford Internet Institute at the University of Oxford. This project ran from January 2014 to March 2015. The primary aim of the project was to facilitate the use of a 65 terabyte dataset containing crawls of the .uk domain from 1996 to 2013. The crawls were conducted by the Internet Archive, which captures and archives web pages on a massive scale (Kahle, 1997). This dataset (hereafter 'the web archive') was acquired for the use of the British Library by Jisc, a charitable organization which facilitates the use of digital technologies in UK education and research. As becomes clear in the researchers' reflections in this chapter, the dataset shared many of the limitations and challenges of other web archives. Although enormous, the archive does not hold every site or page in the .uk domain, raising questions around the representativeness of the archive, and those resources that were captured are time-stamped in relation to the date they were captured, rather than originally created.

As part of the project, ten arts and humanities researchers were invited to use this web archive dataset to conduct cutting-edge research. The researchers, who received a bursary for their participation in the project, were all studying at a doctoral level or higher at the time of the project, and were thus experts in their respective areas of scholarship.

This chapter describes how the researchers utilized the dataset, and explores the findings which emerged from their research. The initiative was distinctive for several reasons. It represented a rare opportunity for researchers with little or no existing expertise in the use of web archives to utilize this exciting but challenging source of data. Moreover, the process was structured so that development of the archive could proceed iteratively, in response to the researchers' feedback; regular meetings between the researchers and developers facilitated this process. This collaborative, iterative process was especially important for the development of 'Shine', the interface used to conduct full-text searches of the archive. Due to the scale and diversity of the data contained in the archive, the search interface became an essential tool for navigation of the archive. Yet, as will be seen, the search interface also served as a platform for *conceptual* navigation of web archive research itself.

The chapter begins with a description of the ten projects, briefly outlining the research foci, the approaches taken and the findings which emerged. These experiences are then synthesized in the discussion section, with a series of wider reflections on the challenges of conducting research using web archives, and the implications for arts and humanities scholarship that result from the use of this valuable but as yet underexplored resource.

Project summaries

Online reactions to institutional crises: BBC Online and the aftermath of Jimmy Savile

For her case study, Rowan Aust of Royal Holloway, University of London focused on the aftermath of a major scandal at the heart of the British Broadcasting Corporation (BBC). Following his death in 2011, a string of sexual abuse allegations emerged against the iconic broadcaster Jimmy Savile, who was famous for decades fronting BBC radio and television programmes and for his prolific charity work. Aust described the revelations about Savile as 'ruptur[ing] the stability' which undergirds cultural memory, and her research focused on understanding how the BBC as a prominent institution reacted to the allegations, analysing how content relating to Savile changed over time on the BBC's website (Aust, 2015: 3).

Aust began the research by conducting an iterative series of searches relating to Savile and the BBC within the archive as a whole and the BBC website in particular. Through comparison with the live site, this yielded a series of instances in which changes had been made as a reaction to the scandal. Aust found only one instance – an interview with Savile on the long-running BBC radio series 'Desert Island Discs' – in which explicit reference had been made to the removal of online material relating to Savile. Elsewhere on the website, attempts to erase or modify content had been applied inconsistently and haphazardly. In some cases, links had been removed or broken; in others, content had been modified or greyed out.

Aust followed up this comparative analysis of primary sources by writing to the BBC's Controller of Editorial Policy. Through back-and-forth correspondence, she learned that the BBC had procedural guidelines for the removal of online content, but that these had only been implemented in 2014, well after the first allegations. Aust's research suggests that, contrary to the 'presumption that material published online will become part of a permanent accessible archive' described in these guidelines, at least some sensitive content has been quietly removed (2015: 8). Yet while attempting to control the narrative around the Savile case, the BBC appears to have been hamstrung by the size and scale of its online presence, which has perhaps proved too diverse and diffuse for a blanket policy of removal or modification to be effectively implemented. Aust's research is thus significant on its own terms – shining light on a serious and significant case – but it also holds lessons for the maintenance and modification of large institutional archives more generally, and the implications of this for cultural memory.

The web archive and Beat literature

For her project, Rona Cran of University College London sought to discover academic and public receptions to Beat literature in the UK as reflected in the web archive, and to establish whether web archives were therefore a useful research tool for literary studies more broadly. Overall, Cran found that the archive 'has great validity and enormous potential as a research tool for literary researchers', describing a 'liberating sense, when working within the archive, of exploring both the past and the future simultaneously – of entering uncharted territory whilst also rediscovering forgotten artefacts' (Cran, 2015: 3).

Yet while it clearly holds much promise for study in this area, Cran encountered a number of challenges with using the archive for conducting research. One was the 'geographic' limitation of the archive dataset: much relevant material was likely to have been published on domains other than .uk. Moreover, the data that was available in the

UK archive was ‘far more fragmented and disparate’ than in more consolidated and comprehensive literary collections elsewhere on the web (2015: 2). What is needed, then, is a process of ‘foraging and sensemaking’: the ‘territory’ represented by the web archive ‘needs to be mapped’ by scholars and the interested public (2015: 3).

Cran found, however, that the archive in its current messy, unmapped and arbitrary state in fact meshes perfectly with a sensitive understanding of Beat literature. Cran draws parallels between the writing styles of Beat writers – such as William Burroughs, whose novels ‘read […] like the uncontrolled spewings of an ailing machine’ and the haphazard nature of the dataset (2015: 5). Further, Beat writers ‘treasured notions of fragmentation, ellipsis and inherent unknowability’, which are ‘positive aspects of the web archive in its current form’ (2015: 6). From this perspective, Cran’s research shows that not only can we learn much about the Beats from the web archive, so too can we learn plenty about the web archive from the Beats.

Revealing British Euroscepticism in the web archive

In his project, Richard Deswarte of the University of East Anglia sought to discover whether and how British Euroscepticism has been recorded in the archive. Deswarte started by creating a list of keywords relevant to Britain’s place in the EU, including ‘referendum’, ‘Eurosceptic’ and ‘UKIP’ (the acronym for the Eurosceptic United Kingdom Independence Party), which were then searched for within the archive (Deswarte, 2015: 3). Unfortunately, the volume of results returned for these queries was enormous, which precluded closer analysis of all or even a meaningful sample of the available resources. Even when more filtered searching was conducted – for example by limiting the date range and removing the most prolific subdomains such as news sites – large numbers of results were returned.

Nonetheless, Deswarte was able to find a number of individual items of relevance to his research focus, including the full text of an old speech by UKIP’s leader Nigel Farage and a series of documentary films. However, despite the academic value of these discoveries, as Deswarte notes, ‘their discovery was serendipitous rather than based on a sound methodological approach to analysing the increasing mountains of materials’ (2015: 4). Various challenges, including the abundance of data available, and issues over the quality and consistency of web page capture, precluded large-scale quantitative analysis to draw out more general patterns regarding Euroscepticism, or to relate them to offline trends. In one instance, searching for 'UKIP' returned hundreds of results contained on the sports pages of a regional newspaper. It took considerable effort on the part of the researcher to establish that this was the result of a rolling news banner. This example demonstrates how seemingly minor elements on a web page can create major issues which are extremely time-consuming to weed out. Deswarte's project serves to show that, while web archives may host valuable material, locating this material, and relating it to broader societal trends across time, currently represents a major challenge for historians.

Searching for home in the historic web: An ethnosemiotic study of London-French habitus as displayed in blogs

Saskia Huc-Hepher, of the University of Westminster, used a number of web archives to conduct an ethnosemiotic case study – an approach combining ethnographic research and semiotic analysis – of the French community in London. Huc-Hepher sought to 'think small when handling big data [to] inject new, deeper meanings', by focusing on a small set of primary sources written in French, using 'the storytelling of individual lives serving as a guiding light' to illuminate the enormous web archive (Huc-Hepher, 2015).

Huc-Hepher searched the archive to develop a corpus of relevant sites: blogs written by French émigrés living in London. There were both advantages and limitations to this approach. The use of French-language search terms proved an effective filter for sites in the .uk archive. However, visual components of the blogs in the corpus – including fonts, photos and videos – were often deficient or broken, threatening to 'ultimately jeapardis[e] the very validity of the multimodal semiotic approach'.

Despite these shortcomings, Huc-Hepher was able to locate a number of blogs authored by French people in London across a range of archives. Due to the restrictions of the web archive – for example, the fact that it only contains .uk sites – only one blog was found here; other examples were identified in different web archives. Through the multimodal analysis, Huc-Hepher was able to assess a whole raft of components on each blog including colour palettes, the content and layout of banner images, typography and text. Crucially, through analysis of these blogs over time – using captures of the same blog at different times in different archives – Huc-Hepher was able to detect subtle but meaningful changes in the emotional position of the blogger in relation to London. In many cases she observed a gradual integration of bloggers into their new environments, finding a 'half-way habitus' or 'hybridisation of habituation'. It is notable that Huc-Hepher was able to conduct a rich, illuminating analysis with only a small number of resources. This points to the contribution to research that even a single page or object can play, and thus reminds us of the importance of archiving as much as possible for the benefit of future research.

Capture, commemoration and the citizen-historian: Digital shoebox archives relating to POWs in WW2

In her research project, Dr Alison Kay of Northumbria University focused on using the web archive to locate and explore 'digital shoebox archives' – micro-collections and narratives of lived experience – relating to Prisoners of War (POWs) in the Second World War (Kay, 2015). At the core of Kay's approach was the iterative development of search strings which would return results relating primarily to Second World War POWs, especially in regard to personal narratives and commemoration. This involved various filtering techniques, including the exclusion of various domains, such as media organizations (like bbc.co.uk) and commercial sites (like amazon.co.uk) and proximity searches of key phrases, to limit irrelevant results. This strategy proved effective: without these filters, the number of results returned for one of Kay's basic search strings was an impenetrable 53,638; with them, it was a much more manageable 206. Overall, for 11 distinct search terms, this figure was 2,894. On one hand, this represented a sizeable decrease from the 24,727 pre-filtered results; yet on the other, it still remained too large for a researcher to single-handedly tackle in the course of the project.

Despite the volume of results and the limited time available to assess them, Kay's project offered an illuminating overview of the sort of valuable historical material captured by the web archive. By identifying a number of online projects gathering memories of war, and filtering to only include results from these domains, Kay was able to investigate what proportion of memories from the live web had made it into the web archive. The findings here varied by project. For the Wartime Memories Project, nearly 10,000 results were returned. This might represent up to two thirds of the 15,000 or so wartime stories and testimonies, though as Kay notes, duplicate captures may have increased this total artificially. For the BBC's People's War project, however, only 346 results were returned – a tiny proportion of the 47,000 stories the project claims to have on the live web. Kay's research therefore suggests that, however imperfect and incomplete (or in the case of duplicates, 'over-complete'), the collection of 'shoebox' memories contained in the archive might well prove a valuable source of materials for historians and the public at large.

Digital barriers and the accessible web: Disabled people, information and the internet

In his project Gareth Millward, of the London School of Hygiene and Tropical Medicine, sought to investigate how information was presented on the internet in a format accessible to disabled users. In particular, it focused on the accessibility of information made available about, and by, disabled organizations themselves. This initial investigation began with a series of searches of the entire dataset, seeking to discover how well represented disability organizations were over time – this analysis found that overall, the Royal National Institute of the Blind stood above its peers in terms of references in the dataset (Millward, 2015). More generally, public-facing charities seem to enjoy better coverage compared with more focused lobbying organizations. Millward also investigated the extent to which disability organizations' websites adhered to the World Wide Web Consortium's Web Accessibility Initiative accessibility guidelines using code validation tools.

Yet in addition to these findings, Millward's report pointed to a series of challenges in conducting analysis of this sort. First, the names of disability organizations had a significant bearing on whether their reach could be accurately analysed: organizations with common names such as Scope and Mind returned a large proportion of irrelevant results, even with additional terms such as 'disability' included in the search string. Secondly, the code validation did not allow a like-for-like comparison between websites: as web pages became longer, the number of accessibility errors in the code would also increase. A final, more general challenge that Millward pointed to was the enormous size of the dataset and the amount of potentially relevant material. This necessitated a series of blanket decisions taken to try to reduce the size of the corpus to an extent that more sensitive analysis would be impossible. As such, while interesting discoveries could and did emerge from this approach, ultimately 'there was very little academic validity to the corpus, and it was difficult to defend the results as representative or in any way objective' (2015: 8).

Nonetheless, Millward sketched a number of future areas to extend this analysis. Instead of conducting large-scale searches of the whole archive, link analysis could be a more accurate way of assessing the relative reach and influence of different organizations. Qualitative analysis of individual websites as well as oral histories with the figures responsible for organizations' online strategy could augment quantitative findings. Moreover, the importance of this line of research is not in question: in the case of the RNIB, for example, we can see 'a continuity from braille through to web access as a core part of the charity's remit' (2010: 8). Engaging with web archives as a primary source for exploring this phenomena is therefore inescapable.

A history of UK companies on the web

Marta Musso, of the University of Cambridge, sought to track the diffusion of internet use among UK companies, with a focus on the period between 1996 and 1999, when websites of companies were generally in their infancy. Musso utilized a range of sources of data and approaches, comparing captures of company websites in the archive with contemporary versions on the live web, as well as conducting a questionnaire and examining records of website registrations and contemporary newspaper articles (Musso, 2015).

Musso began by attempting to sketch the gradual uptake of website registration by major companies in the early stages of the web. Combining information from various archival and other sources, she found that, despite a UK-centric sample of companies, many used the generic .com domain as opposed to UK alternatives. Moreover, over the course of the period she observed 'an overall tendency [among companies] to switch to a .com domain and to simplify the address altogether'. She also found that older and larger companies tended to register their company websites earlier than others, usually before 1996. Through responses to the questionnaire, she learned that BP registered their domain address bp.com as early as 1989, although this served merely as a placeholder for many years.

Following up these findings, Musso conducted a series of searches of the archive for words relating to business and commerce, which yielded a series of company directories from the period. These directories, however, only existed briefly; from around 2000 onwards, the sophistication of search engines had rendered them far less useful. Examining some of these sites in more detail, Musso observed that many of them were 'mimicking physical commercial spaces', borrowing elements from an offline commercial setting – for example by using a ‘front door’ and user-friendly menus – to make the online browsing experience more familiar to users. Taking these findings together, Musso suggests that the early experiences of companies on the web reflects more general patterns in internet use, resulting in a ‘shift from a private means of communication […] to a public space, a virtual reality in which everyone [can] have access to every space’.

The online development of the Ministry of Defence

In his project, Harry Raffal of the University of Hull explored the web archive to trace the development of the websites for Britain’s Armed Forces – the Army, the Navy and the Air Force as well as the umbrella Ministry of Defence (MoD). Raffal began by creating a corpus of five iterations of each of the websites across the period 1996–2013 (Raffal, 2015). This corpus was thematically analysed by coding various elements on the page, chiefly text, videos, images and navigational elements. Raffal also conducted link analysis of the MoD site, in relation to a number of subsidiary recruitment and educational organization websites.

These analyses yielded various findings about the purposes behind the Armed Forces’ web presence. Through thematic analysis Raffal found that, although the content and design of the MoD website has changed over time – in line with developments in web standards and trends more widely – the initial intentions for the Ministry’s online presence – promoting a ‘corporate image’ and serving ‘business and presentational needs’ – remain largely in place. In the case of the Armed Forces, recruitment has remained a chief concern: especially in the case of the Army, which integrated television commercials with interactive website content, and shifted its recruitment terminology from the word ‘career’ to the more informal ‘join’. Raffal contrasts this with the continued use of ‘career’ among the Navy and Air Force, for whom longer-term, more technical appointments remain the norm.

Raffal’s link analysis also yielded interesting results. It highlighted an unexpectedly prominent role in the network of armed forces websites for the MoD’s Supporting Britain’s Reservists and Employers Agency (SaBRE). A high concentration of inbound links from local authorities and reserve associations suggested that SaBRE was ‘achieving at least part of its remit as an organisation that aims to build support for members of the Armed Forces’. Overall, Raffal’s research benefited from the creation of a systematic, self-contained corpus and the utilization of mixed methods to uncover meaningful patterns among the UK Armed Forces' online presence.

Looking for public archaeology in the web archives

In her project Lorna Richardson, now of Umeå University, sought to explore representations of public archaeology in the web archive (Richardson, 2015). In common with other projects, the mass of resources available led Richardson to narrow and fine-tune her search criteria, from the well-known archaeological site Stonehenge to three slightly more obscure sites, which respectively returned substantially fewer results. Richardson employed the archive's n-gram function to display the results over time, and found that, for the 'Ness of Brodgar' site, an archaeological find in Orkney made in the late 1990s, 'the N-gram visualises beautifully the release of information as the archaeological site progressed in its discoveries'. Richardson is able to show here that information released from the archaeological site gradually makes its way onto archaeological websites.

Richardson notes that overall, there are hundreds of thousands of pages in the archive containing material potentially relevant to public archaeology. Richardson's approach to working through this mass of data evolved into a 'manageable scoping exercise for a handful of key archaeological sites and terms'. Indeed, this approach can be extended to web archive research more generally: as Richardson suggests, 'using an archaeological approach to explore, reconstruct and reimagine the technologies of past iterations of the World Wide Web' could improve academic understanding of how the archive can be used effectively.

Do online networks exist for the poetry community?

Helen Taylor of Royal Holloway, University of London conducted research into the presence of poetry networks in the web archive (Taylor, 2015). Poetry networks long predate the internet, allowing content to be spread informally between different locales; Taylor sought to discover whether online poetry networks 'exist in and of themselves, or whether the online presence of a group is merely a kind of "placeholder" or "directory"' (2015: 2). Taylor restricted her analysis to two poetry websites which represent both ends of this spectrum: The Poetry

Forum, a place for people to share and comment on their own poetry, and the Oxford University Poetry Society's website.

In her analysis of The Poetry Forum, Taylor observed a number of features which point towards it being a genuine community. The sign-up process, while optional, encourages contributors to have a profile – although these need not faithfully reflect their offline persona – promoting the sharing of original work and comment. Taylor surmises that the site 'is an example of how poetry networks do exist online […] this kind of interaction and exchange would not have been possible before the internet' (2015: 3). This contrasts sharply, Taylor finds, with the case of the Oxford University Poetry Society, wherein the site's 'online presence is only there in order to get you to engage offline' (2015: 4). Having taken a temporally representative series of captures, Taylor found that members' poems only seldom appear on the site, and even when they do there is no facility for comment or discussion. Taylor further speculates that there was a deliberate decision not to bring the discussion online, given the haphazard use of different URLs over time, and the frequent appearance of 'we have moved' messages.

By highlighting these two examples located in the archive, Taylor's research demonstrates the diversity of approaches taken to the creation and maintenance of poetry networks online. She concludes that, although the web has been transformative in facilitating connections between geographically diffuse participants, in circumstances where these virtual connections are not required, another website can play an entirely different role.

Discussion

Giving the ten researchers direct access to the data through the search interface, and closely involving them in the development process, yielded a range of insights regarding the utility of web archives for humanities research. Each of the researchers dedicated significant portions of their project reports to reflect on the potentials and pitfalls of conducting research using web archive datasets. In this discussion section, these reflections are synthesized, yielding three topics which are drawn out in greater depth: how the researchers conceptualized web archives; the strategies for research that were taken; and the importance of search tools for navigating the massive archive.

Conceptualizing web archives

Although prior to this project the ten researchers had had little or no experience with web archives, they came across many characteristics of this field that have already been highlighted by scholars elsewhere (Brügger, 2012; Schneider and Foot, 2004). For example, many researchers noted both similarities and differences between the web archive dataset and older, traditional archives. Richard Deswarte suggested that 'in many ways, the term "archive" is a misnomer', since what the researcher really faces is not a web page in its original form, but a reconstruction of a pre-existing web page – and often an incomplete one' (Deswarte, 2015: 6). In a similar vein, Alison Kay noted that 'as a historian I have to remind myself that the online web is gone. We [only] have representations'. Web archives might therefore have less in common with historical archives – which are typically text-based in nature – than with archaeological artefacts, a metaphor proposed by Lorna Richardson. A technical disjuncture between how a web page appeared on the live web and how it is rendered in a web archive thus exists alongside a temporal disjuncture between when an archaeological artefact was originally used and when it was found. Rona Cran also observed conceptual similarities between the web archive and the 'uncontrolled spewings of an ailing machine' characteristic to Beat literature (Cran, 2015: 5).

Yet this is not to overstate the conceptual departure from traditional archives represented by web archives. As Alison Kay noted, 'mass printing was worrisome in its volume in years past, in the way that the archived web is challenging today'. Indeed, just as with web archives, most historical archives are subject to some degree of arbitrariness regarding what is and is not preserved. In the case of web archiving, both technical and curatorial factors can affect what is kept and what is discarded. The web archive used here essentially represents the .uk portion of the far larger Internet Archive, which hoovers up vast tracts of the live web for archiving on the basis of links between sites. Yet even at this huge scale, there is a role for curation, as with the Archive's policy to respect the robots.txt protocol when crawling. Whether a page does or does not appear in the Internet Archive is therefore the result of both technical contingencies and curatorial considerations. In a certain respect, the fact that technical as well as more subjective factors affect what appears in a web archive 'could be regarded [as] a refreshing objectivising tool […] a means of making the final collection less a reflection of [the archivist] and more about the material itself', as Saskia Huc-Hepher suggested.

The researchers thus found that web archives, at least in their current state, represent a curious position in relation to previous sources of data. They are both similar to and distinctive from traditional historical archives, whilst also holding conceptual affinities to archaeological and literary traditions. The following section explores how these different perspectives on what the web archive represents fuels alternative approaches to utilizing it for research.

Strategies taken

The breadth of research interests pursued by the ten researchers on the project was reflected in the large degree of diversity in terms of the methodological approaches taken, despite the fact that all the researchers had access to the same data and the same tools. These diverse approaches can nonetheless be roughly clustered into two contrasting strategies, which can be labelled the 'part of the whole' approach and the 'whole of a part' approach. Should they tackle the entire archive dataset in all its enormity and complexity, using the search engine to isolate specific items across the archived web relating to their research (the 'part of the whole' approach)? Or should they restrict their research focus to a pre-defined, substantively meaningful subset of resources (the 'whole of a part' approach)? These strategies are not mutually exclusive, even in a single research project: many of the researchers ultimately used both in their own projects. But the distinction drawn here helps to illuminate the strengths and limitations of each approach for conducting valuable research.

An example of the 'part of the whole' approach is Richard Deswarte's investigation into Euroscepticism. One clear advantage of Deswarte's strategy of searching the archive as a whole is that results which emerge are representative of the archive itself, allowing longitudinal analysis of the prevalence of a given phenomenon (in this case Euroscepticism). Of course this is not to suggest that the archive itself is necessarily representative of society: numerous aforementioned issues, such as the inconsistent capture of pages, cast doubt on the true representativeness or reflectiveness of society in the archive. Furthermore, other practical issues limit the utility of this strategy. As Deswarte found, the number of results returned when searching the archive as a whole can be huge, particularly when researchers are seeking evidence of general phenomena (Deswarte, 2015). This strategy also gives rise to the 'so what?' problem described by Gareth Millward, who argued that 'only through disaggregating these results can we gain any real meaning that might be of use' to researchers (Millward, 2015: 5). With the net cast so wide, trawling through the vast catch can simply be too time-consuming, a challenge also noted by Alison Kay and Lorna Richardson.

Seeking to collect and analyse a part of the whole archive is therefore replete with challenges. Other researchers, in contrast, adopted what might be called the 'whole of a part' strategy. This approach meant focusing squarely on a pre-defined set of resources – usually one or a small number of websites – and analysing them in their entirety. Research projects primarily using this approach include Harry Raffal's investigation of the Ministry of Defence websites, Helen Taylor's analysis of two distinct poetry networks, and Rowan Aust's investigation of the Jimmy Savile scandal as reflected on the BBC website. Again, there are obvious advantages to this approach. Researchers can use the filter-by-domain feature of the search interface at the outset, resulting in a far smaller set of results, which is likely to allow more sensitive, 'line-by-line' analysis of all the results returned. Yet though this approach may yield a tighter and more internally coherent group of resources on which qualitative analysis can be performed, care must be taken to highlight that the items selected for analysis are not representative of the archive as a whole.

Both the 'part of the whole' and 'whole of a part' strategies therefore have strengths and limitations for producing valuable research using the archive; these are summarized in Table 11.1. Overall, it seems sensible to make the decision of which strategy to adopt based on the nature of the research question. Where the research question centres on a broad social or historical phenomenon it may make most sense to pursue the 'part of the whole' strategy, all the while bearing in mind the significant challenge of scale that this approach often entails. In contrast, where the research is focused on a specific entity or event, particularly where this is associated a priori to a particular subdomain or website, the 'whole of a part' approach may work best.

This section has explored two contrasting ways of conducting valid research using large-scale web archive datasets. Of course, the two strategies presented here are not mutually exclusive, since many researchers utilized both strategies at different points. Nor are they strictly dichotomous: a 'web sphere' (Schneider and Foot, 2004) – the third largest of the five 'strata' of web analysis listed by Brügger (2012) – could be the unit of analysis in either approach, depending on the size and nature of the 'sphere' in question. Moreover, both strategies involve the use of searching the archive at some stage in the research process. In the following section, therefore, the purpose and process of searching is explored in more detail.

Table 11.1 Comparing strategies for web archive research

Approach taken	**Summary of process**	**Key advantage**	**Key limitations**
'Part of the whole'	Searching the entire archive for a broad historical/ social phenomenon (filtering occurs mostly a posteriori)	Ability to treat the archive as a whole and make definitive statements about the archive	– Archive not necessarily representative of society – Structured data important for quantitative analysis, yet web archive data is not (usefully) structured
'Whole of a part'	Searching a particular subdomain or website for a specific entity or event (filtering occurs mostly a priori)	Able to adopt a sensitive, grounded understanding of web pages and elements	– Difficult to make definitive statements about how resources analysed relate to or represent the archive as a whole – Meaningful findings are discovered serendipitously not systematically

The use of search tools

Whichever strategy the researchers employed, the enormous size of the database meant that researchers required a search interface through which to access, assemble and analyse the resources relevant to their research question. Shine, the search interface which the researchers used, was developed over the course of the project, largely informed by the researchers' experiences. The development and use of the interface inevitably opened up another raft of conceptual questions and challenges. Research does not take place in a vacuum – something that humanities scholars appreciate more than most – and many of the researchers noted that their previous experience with using search engines to navigate the live web affected their assumptions about searching web archives. Richard Deswarte coined the phrase 'Google mindset' to describe the set of expectations that researchers had about how the search interface would or should work (Deswarte, 2015: 9).

The core difference – straightforward in principle but disorientating in practice – relates to the ordering of search results. The algorithms developed by Google and other search engines are as lucrative as they are elusive, ranking billions of results by perceived relevance in a split second. In the case of the Shine search interface, the size of the index is enormous – with hundreds of thousands of results common for basic queries – but the ordering of results is far less sophisticated. Instead of 'relevance' as an option, researchers can order results by, for example, an item's title or the date it was crawled. Thus as Richard Deswarte pointed out, 'all of the results' – not just the first few – 'will potentially be of interest' (Deswarte, 2015: 7).

Viewed from a different perspective, however, some researchers found this limitation liberating and even empowering. Rona Cran took a 'deliberately unsystematic approach to the archive, by treating it as something akin to a vast bundle of unsorted papers rather than, say, Google'. In doing so, she 'was able to confront it with my own perspectives'. For Cran, this process 'heightened [the] intellectual integrity' of her study, since she was 'using processes of reasoning and selection which were unique' to her as a humanities researcher (Cran, 2015: 5). Thus through the limitation of the search interface her research expertise had fresh importance.

This sense of greater control over the research process was bolstered by later developments of the interface. As noted, this development process took place iteratively, directly in response to the feedback of the researchers. Most significantly, the ability to create a personal corpus of results – extracting individual results from search results into a persistent collection – was the new feature most requested by researchers across the project, and was well received when introduced. When combined with advanced search tools, which allowed researchers to search only particular domains or across narrower time periods, for example, it gave researchers a more powerful set of tools with which to tackle the data available.

However, even with the addition of these tools, saying anything definitive about the contents of the archive in general remained extremely difficult, particularly for researchers whose research questions were broadly conceived. One approach would be to create a small sample of the data, which could then be sensitively analysed by the researcher. Yet Richard Deswarte noted that while 'structured data mean[s] it is possible to make clear and academically justified decisions on the size and relevance of representative samples […] unfortunately and problematically the data in web archives is almost completely unstructured, at least in terms of content' (Deswarte, 2015: 3). Another problem was explained by Gareth Millward, who suggested that it was in fact the relative scarcity of data that made it possible to answer a historical question in the traditional way. Typically, traditional historians 'identify a question and source base, go back to the archive, and then mine what [they] can until that vein is exhausted. This is [only] possible because we have a relatively small amount of evidence which has survived' (Millward, 2015: 6).

In summary, the researchers' experiences with the search function seem to have been in equal parts frustrating and empowering: frustrating, because of the lack of any substantive ordering through which researchers can get to grips with the voluminous resources available; yet empowering because, in the absence of any such pre-ordained notion of relevance, researchers must make more decisions based on their own domain expertise. These perspectives are not, of course, mutually exclusive: the most frustrating aspects of the experience could be mollified by, for example, more powerful and tailored search functions, whilst still allowing researchers the ability to make informed decisions about what to include in a corpus. Yet in addition to technical improvements, researchers and developers need to continue to engage critically with the utility of full text searching: as Richard Deswarte argued, 'its preeminence as the main approach to accessing web archives cannot remain unquestioned', and for Alison Kay, 'historians need to be contributing to discussions today about the sources of tomorrow' (Deswarte, 2015: 9).

Conclusion

Each of the ten case studies discussed here have moved the web archiving research front forwards, both in the specific areas they covered, and through the necessarily innovative methodological approaches they adopted. This reflects the initial aims of the BUDDAH project, which set out not only to 'highlight the value of web archives as a source for arts and humanities researchers', but also 'to establish a theoretical and methodological framework for the analysis of web archives' (Big UK Domain Data for the Arts and Humanities, n.d.). The previous section suggests that such a framework is indeed in development, albeit at a nascent stage. Moreover, the project demonstrated the importance of incorporating the perspectives of researchers at each stage of development of the dataset.

Yet this chapter has also made clear how much still remains to be done to ensure that the great potential of web archives as a source for arts and humanities research can be realized. The chapter has described the nascent strategies and techniques used by researchers on the project, but clearly many questions remain unresolved. These include, for example, how to handle messy and incomplete data; how web archive research is assimilated into the mainstream of a range of different disciplines; and how the results of search queries of the dataset can be meaningfully presented.

Since the conclusion of the research projects, both the underlying archive dataset and the Shine interface used to access it have been under continuous development, in large part in response to the researchers' experiences described here (Jackson, 2016). The interface is now accessible to everyone interested in conducting their own research about how UK society is reflected on the web between 1996 and 2013. The two main features currently offered are the search tool, which enables faceted browsing of the results, and a trend analysis tool, showing the relative appearance of a given word or phrase in the archive over the 18-year period. These tools, and more guidance, can be found at https://www.webarchive.org.uk/shine.

Crucially, however, those coming into contact with web archives in the future will not enjoy many of the advantages that researchers on this project benefited from, including contact with those developing the dataset, and the ability to share challenges and solutions as a group. As web archives continue to be developed, therefore, it is important that researchers as a user group are kept squarely in mind, even if they are not always in earshot. This chapter has illuminated many of the successes that researchers enjoyed, the challenges they faced, and most significantly, the ways in which they conceptualized and approached web archives as a source for scholarship. It is hoped, therefore, that this chapter – and the research projects that it profiles – can serve as a resource not only for scholars engaged at this emerging research front, but also for those involved at every stage in the development of web archives for research.