Chapter 2: Web Archives and the Problem of Access: Prototyping a Researcher Dashboard for the UK Government Web Archive

Mark Bell, The National Archives, London | Tom Storrar, The National Archives, London | Jane Winters, School of Advanced Study, University of London.

Abstract

There is a burgeoning secondary literature concerning the use of the archived web as a primary source for Humanities research, but it remains centrally concerned with how to work around problems of scale, complexity and access. The manifold barriers encountered include the inability to download and take away data; the opacity of web harvesting processes; the (unknown) scale of content duplication; and the unsuitability of keyword searching as a primary means of exploration. An important record of the recent past remains tantalizingly out of reach for the majority of historians, political scientists, literary scholars and others.

This chapter will explore how a combination of Humanities methodological and research concerns, the expert knowledge of archivists, and machine learning solutions can work together to transform access to the open UK Government Web Archive (UKGWA). We will outline the theory behind and the steps towards building a prototype researcher dashboard for the UKGWA allowing multiple routes into and views of the archives of government online over more than two decades.

1. Introduction

This chapter begins by describing the history and current status of the UK Government Web Archive (UKGWA),[1] which is provided by The National Archives of the UK (TNA), before addressing some of the specific and more general challenges associated with archiving the web. It moves on to discuss the different ways in which researchers in the Humanities and Social Sciences might want to make use of the UKGWA, and outlines some of the factors that currently hinder, or prevent entirely,

[1] http://www.nationalarchives.gov.uk/webarchive/ [last accessed: April 2, 2021].

optimal access. Next, it introduces the concept of co-design, involving archivists, researchers and technologists, as a means for developing useful and sustainable tools and modes of access for web archives in general, and the UKGWA in particular, before describing what a co-designed researcher dashboard might include. This section of the chapter also explores the kinds of research that would be enabled by the provision of a suite of tools sitting on top of the web archive, as well as the utility of such a service for the host institution. It concludes by reflecting on the value of, and mechanisms for, collaboration to enhance access to web archives.

1.1 What is the UK Government Web Archive (UKGWA)?

The UKGWA was established in 2003 by The National Archives, the official archive and publisher for the UK government, and for England and Wales. The rapid rise in the use of the web as a platform for disseminating information began in the mid to late 1990s and it had become clear that TNA, as an institution concerned with gathering the evidential record of government, and how the state interacts with the citizen, would need to collect public websites.

The initial archive was formed of a small number of key sites, and some content from the mid-1990s to mid-2000s was added, ingested from the Internet Archive.[2] The scope of the collection was then widened dramatically in 2008 to agencies and "arm's length bodies," and this collecting remit has remained in place ever since. The archive has evolved to archive resources published on other platforms, most notably social media, including Twitter, YouTube and Flickr.[3] This expansion led the latest of TNA's collection policy documents to refer to the collection's scope as the "UK Central Government Web Estate,"[4] broadening it from "traditional" websites.

While the collection remains limited to the UK central government, its departments, agencies, arm's length bodies, the National Health Service (NHS) and public inquiries, the UKGWA, as a web archive,[5] is a complex and varied collection, which is constantly accruing material and adapting to the capturing challenges of the present, while simultaneously accommodating the technology of the past. As

[2] The Internet Archive "is a non-profit library of millions of free books, movies, software, music, websites, and more." At the time of writing, its Wayback Machine offers access to more than 486 billion archived web pages from around the world. https://archive.org/ [last accessed: April 2, 2021].

[3] https://webarchive.nationalarchives.gov.uk/social/search/ [last accessed: April 2, 2021].

[4] https://www.nationalarchives.gov.uk/documents/information-management/osp27.pdf [last accessed: April 2, 2021].

[5] https://netpreserve.org/web-archiving/ [last accessed: April 2, 2021].

of autumn 2020, the UKGWA contains approximately 6 billion resources[6] across the 24 years of archives it hosts.

Beyond providing a record for posterity and future research, the UKGWA plays a key role in "Web Continuity."[7] This initiative seeks to reduce the number of broken links on government websites by providing public access to highly complete web archive snapshots, while also redirecting users to the web archive when a resource is no longer available on the original website. Furthermore, the archive has been used as a trusted home for websites closed in the process of consolidating government information onto websites such as gov.uk and previously Directgov and Business Link. It is the combination of these initiatives, and the fact that the UKGWA is open to anyone with an internet connection, that means the archive has many thousands of daily users, according to web server log analysis and Google Analytics.

The archive is updated continually through a number of collection processes, including scheduled captures of websites and exceptional, high priority captures, often in response to events of national significance. The latter collection method is often employed to make time-critical captures, for example the government's response to the COVID-19 pandemic, or the UK's exit from the European Union. This will be described in more detail later, as these factors have an influence on the use of the archive.

The UKGWA, as a well-used and trusted service, meets its core mission of capturing the published government record and providing access to it. As the collection has grown and matured, and the service is often and increasingly the only reliable source of this information, researcher interest in exploiting the collection has increased and there is every indication that that interest will continue to accelerate. While TNA has supported several research events and projects, these are normally large undertakings that require the production of tailored datasets (for example, Computational Archival Science[8] and Alan Turing Institute Data Study Group[9] events held in 2019).

A key element in supporting research is an understanding not only of what the collection contains, but how the collection came to contain it. The original content

[6] A resource is anything with a Uniform Resource Locator (URL) and includes everything from HTML pages to images to the JavaScript files necessary to reproduce websites via replay software.

[7] https://webarchive.nationalarchives.gov.uk/ukgwa/20130102170449/http://nationalarchives.gov.uk/information-management/policies/web-continuity.htm [last accessed: August 31, 2021].

[8] https://blog.nationalarchives.gov.uk/network-analysis-of-the-uk-government-web-archive/ [last accessed: April 2, 2021].

[9] https://www.turing.ac.uk/events/data-study-group-december-2019 [last accessed: April 2, 2021].

creators were government departments but the act of archiving the information relies on a series of complex interactions between human actors (members of various teams), the technologies used to create and capture the resources (web technologies and capture techniques) and the points in time at which they are captured.

1.2 Structure and Collection Process

In contrast to many other types of archives and collections, web archives do not normally enjoy the same degree of intellectual control in describing their contents. This is necessary because the services often need to prioritize capture at scale above description, in order to minimize the risk of loss.

However, there are sources of contextualizing information relating to provenance or how a resource may relate to others within or outside the collection. The real challenge is to capture, convert and convey this knowledge in a way that can be easily consumed by researchers.

The web itself inherently contains a rich amount of contextual information. These characteristics include URLs, linkages between resources and a wealth of other structures, from unstructured text to highly structured forms, such as XML. The vast majority of this is preserved in the web archive.

Aside from typical structural data that the UKGWA inherits from the resources it captures, the archiving team also has a selection of web archive specific tools available, such as CDX,[10] which presents some of this data in machine-readable formats. The UKGWA has other valuable sources of context too. First, the UKGWA uses a database to manage the archiving process and stores archivist decisions and explanations. These might include, for example, the reason a website was archived on a particular day, which has enormous potential for providing a rich commentary, and may help users to make sense of the shape of the archive.

Second, XML files generated by this database, which act as messenger files between it and the crawler, contain specific technical information for each web crawl, such as "include" and "exclude" rules which often change between crawls. Therefore, associating each XML file with its respective "snapshot" crawl may be desirable. Third, TNA's catalogue service, Discovery,[11] contains a wealth of knowledge relating to the government bodies responsible for each website (i.e. domain) in the collection. Also of significance is that Discovery holds millions of other records at The National Archives, reducing any barrier between them and the web archive. This again conveys essential context that not only explains why a website was se-

[10] https://archive.org/web/researcher/cdx_file_format.php [last accessed: April 2, 2021].

[11] http://discovery.nationalarchives.gov.uk/, see e.g. http://discovery.nationalarchives.gov.uk/details/r/C16668 for a specific series-level description [last accessed: April 2, 2021].

lected but what happened to it, and where it fits into the wider patchwork of the collection.

Human actions involve decisions on when and how to capture a resource or a website but also why that effort was made. Data on this is kept as part of the archive but most of it is not public, being historically considered purely "administrative" in nature. However, being that web archives are created through actions and decisions, both human and machine, these are rooted in the time and the context in which they are made. It is often necessary for a web archivist to modify rules to include or exclude elements to successfully capture a resource or set of resources. This may be to avoid crawler traps[12] or, more often, to expand the scope of a capture so that it captures a sub-domain, or some externally-hosted content, pertinent to the website. These decisions and rules are easy to implement but can have a significant bearing on the completeness of the archive, the boundary around it, and ultimately on its users' abilities to comprehend it.

As the model trusted to comprehensively capture the published government record, the web archive needs to be of sufficiently high quality. To achieve this, quality assurance is performed by members of the web archiving team. Using a mixed approach of manual and automated methods, tools and experience, web archivists verify the capture of content and its rendering in replay tools. Web archivists do, however, need to prioritize certain aspects of quality assurance, for example capture over some elements of replay. While this is no surprise, as web archiving is a "lossy" process, most current tools and approaches keep it to a minimum in the UKGWA. Decisions relating to this are recorded in some way, be it via checklists, logs, database systems, or archivists' notes. However, these are often only understood by trained web archivists and therefore would not necessarily facilitate greater understanding of the collection.

The issue of scale makes it necessary for high-volume capturing, a process which is not always compatible with producing and disseminating detailed information about the collection process. A good example of this is an average crawl of the gov.uk[13] website, which is archived monthly, and contains over 1.8 million resources. This also largely explains why the UKGWA is only catalogued at TNA at website level. However, a dashboard could still exploit tools intrinsic to the web archive: data from CDX, and potentially from logs generated by the crawler, could

[12] "A crawler trap is a set of web pages that create an infinite number of URLs (documents) for [a] crawler to find, meaning that such a crawl could infinitely keep running and finding 'new' URLs." One example of a crawler trap is an online calendar with an almost infinite date range. https://support.archive-it.org/hc/en-us/articles/208332943-Identify-and-avoid-crawler-traps- [last accessed: April 2, 2021].

[13] https://webarchive.nationalarchives.gov.uk/ukgwa/20200901093455/https://www.gov.uk/ [last accessed: August 31, 2021].

be presented to show when, how and why a resource was captured on a particular date. Machine learning and AI techniques are likely to be extremely useful in addressing these challenges in the future. However, we already have many promising routes to exposing some of this valuable context, from static dataset files (for example, csv files) to services that support querying to produce machine-readable and "at-scale" data (for example application programming interfaces, or APIs). It is important that these approaches are documented openly, allowing for collaboration between web archiving institutions and a common understanding among researchers of their potential use. Such an approach may lead eventually to forming widely-adopted conventions, or even standards, that will support researchers moving between collections without having to navigate the nuances within each separate collection.

Beyond publishing explanatory metadata, capturing and conveying useful data relating to decision making is challenging in a number of ways. It is therefore likely to be desirable for the web archive tools to do the "heavy lifting" with the dashboard, providing useful functionality to ease digestion of that information. Nevertheless, it is worth briefly discussing some of the challenges that need to be overcome.

In common with all web archives, usability is a difficult challenge and our research supports the notion that new users often need to spend some time with the web archive before becoming confident in using it. The provision of the data and metadata described will be driven by collaboration between the web archiving team and researchers as they use the collection.

The UKGWA is not only well used but also serves a broad user base. A combination of online and "in person" user testing projects have shown users range from members of the public seeking historical reports or data to journalists wishing to see the evolution of policy on a particular topic; from civil servants researching previous policies to solicitors accessing historical guidance.

1.3 What Do Researchers Want to Do with the UKGWA?

The majority of national web archives impose access restrictions, ranging from complete closure to the public (for example, in Sweden) to off-site access for bona fide researchers located in the host country (for example, in Denmark).[14] The most common form of restriction tends to arise from the existence of Legal Deposit Legislation, which allows the harvesting of national web domains at scale but often limits access to browsing only on the premises of the archiving institution. The challenges that this poses for researchers and other users have been well

[14] International Internet Preservation Consortium, Legal Deposit, n.d. https://netpreserve.org/web-archiving/legal-deposit/ [last accessed: April 2, 2021].

documented.[15] The UKGWA, however, like the Croatian, Portuguese and Icelandic web archives, permits unrestricted access online to its data, from anywhere in the world. Researchers can readily consult this essential primary source for the history of the late twentieth and early twenty-first century, either through its own search interface or via TNA's Discovery catalogue. As noted above, a full-text search is available, and there is also a browse option for those who are more familiar with the structure of the UK government and its departments and ministries. Given the restrictions that exist for other web archives, the value of this open access, and the broad permission to copy and reproduce that arises from Crown Copyright, should not be underestimated.

Access by means of a public search interface meets many of the needs of users of the UKGWA. It is easy to search for a particular government report if you roughly remember its name (even if you cannot, you might eventually find it); you can carry out a case study of the Department for Culture, Media and Sport by locating it on the site browse list (which handily takes account of the fact that it was renamed the Department of Digital, Culture, Media and Sport in 2017). But the limitations of search for an archive of this size soon become apparent. The UKGWA interface gives "Budget 2010" as an example search. At the time of writing, without placing the phrase in double quotation marks, this generates 100,525,737 results. Even with the quotation marks in place, so that the full phrase is being searched for, 105,052 results are generated. These results, which are not presented in any particular order or ranking, can be browsed and read 25 results at a time, which is not a task that is reasonable for anyone to undertake.[16] Overwhelmed by volume, and with no means to extract and refine the data offline, the researcher effectively hits a dead end.

Quite apart from the challenge posed by scale, and the limitations of in-browser searching, the modes of access currently available to the user fail to provide a sense of the scope of the archive. It is one thing to know that more than 5,000 websites have been archived between 2003 and 2020,[17] but what kind of information does this include, how often has data for particular sites been collected, and how has the UK government web estate changed over time? This is true for many digital

[15] See, for example, Ian Milligan, Web Archive Legal Deposit: A Double-edged Sword, 14.07.2015, URL: https://ianmilli.wordpress.com/2015/07/14/web-archive-legal-deposit-a-double-edged-sword/ [last accessed: April 2, 2021]; Jane Winters, Giving with One Click, Taking with the Other: Electronic Legal Deposit, Web Archives and Researcher Access, in: Melissa Terras/Paul Gooding (eds.), *Electronic Legal Deposit: Shaping the Library Collections of the Future*, London 2020, 159-178.

[16] At the time of writing, a Google search for "Budget 2010," without the enclosing quotation marks, returns 1.8 billion results (516,000 with the quotation marks), but the top few are highly relevant.

[17] The National Archives of the UK, How to Use the Web Archive, n.d. http://www.nationalarchives.gov.uk/webarchive/information/ [last accessed: April 2, 2021].

archives, but in the case of the UKGWA and other web archives the position is further complicated by the vagaries and particularities of the crawling process through which data is harvested. There are multiple levels of complexity here, not least the fact that parts of the UKGWA (pre-2003) are derived from the Internet Archive, which has its own crawling protocols and criteria for collecting websites on the gov.uk domain.[18] This archival and algorithmic context is essential for researchers approaching the UKGWA with anything more than an interest in a single government report or news announcement.

What, then, would help researchers to make the most effective use of the UKGWA? Building on the foundation of open data, we can envisage a researcher dashboard catering for a wide range of use cases and accommodating different levels of technical expertise. A non-exhaustive list of user requirements might include: access to metadata and statistics; the ability to export different kinds of data from the archive (metadata, images, page content stripped of menus, headers and footers); tools for analyzing trends in the data, for example linguistic and cultural change; the option to analyze online networks of government and the flow of information between departments; and visualization tools assisting both navigation and analysis. The context for this data would also be presented to the user, allowing them to explore the ebbs and flows of archiving the gov.uk domain, which is affected by changes in technology and web design as much as by political crises and the transfer of power between administrations.

We would argue that the prototyping of a suite of tools of this kind requires collaboration between archivists, researchers and technologists. The value of collaborative working and co-design for web archives has already been demonstrated by the "Big UK Domain Data for the Arts and Humanities" (BUDDAH) project, whose co-created SHINE interface influenced the development of online access provision for the UK Web Archive at the British Library.[19] The difficulties of contextualizing, accessing and analyzing the archived web are too complex to be addressed by individuals or single institutions, and there is now an opportunity to bring together the three main stakeholder groups to design tools and services that will be robust, flexible, customizable and sustainable. This requires long-term collaboration and engagement: researchers' needs will change over time, as will the challenges of archiving an ever-moving digital target.

[18] The National Archives of the UK, Information on Web Archiving, n.d. https://webarchive.nationalarchives.gov.uk/ukgwa/20170608213215/https://www.nationalarchives.gov.uk/webarchive/information.htm [last accessed: April 2, 2021].

[19] For more information about the BUDDAH project (funded by the Arts and Humanities Research Council, grant reference AH/L009854/1), see Josh Cowls, Cultures of the UK Web, in: Niels Brügger/Ralph Schroeder (eds.), *The Web as History: Using Web Archives to Understand the Past and Present*, London 2017, 220-237.

The suggestions for a dashboard for the UKGWA that follow are the result of a number of meetings and conversations between the authors of this chapter, who bring the perspectives of, respectively, a research software engineer, a web archivist and a Digital Humanities researcher. A different configuration of contributors would no doubt result in other proposals; and not everything that is outlined below would be possible for all web archives. The list does, however, serve as a starting point, highlighting key research themes but adopting a realistic approach to what can be achieved within archiving institutions which have limited resources at their disposal, and multiple competing priorities.

2. Towards a Prototype Dashboard

Our prototype dashboard will allow the researcher to define the scope of their analysis along three dimensions: breadth, depth, and temporality. Breadth is the selection of websites to be included, depth defines the parts of each website to be included, and temporality defines the period of analysis. Brügger proposes five strata to delimit the web as an object of study: web element (for example, a piece of text or an image on a page), web page, website, web sphere (web activity related to an event, concept or theme) and the web itself.[20] The first two map to the depth dimension, selecting the types of content (elements) and setting selection criteria for the pages. For example, the depth could be defined as all hyperlinks appearing on homepages, the text content of accessibility pages, or images extracted from a random sample. Depth could also relate to an entire site, which is the third of the strata and also the minimum value for the breadth dimension. The web sphere, first proposed by Steven Schneider and Kirsten Foot, is defined as activity related to an event, concept or theme, and aligns with the breadth dimension.[21] We would expand the configuration of the breadth dimension to enable the definition of any subset of sites, including random sampling, but the web sphere idea of filtering according to a theme rather than an explicit list of sites is an important feature.

The temporal dimension is arguably the one that sets exploration of the web archive apart from that of the live web. The archive is constructed from multiple snapshots of web pages, the distance between snapshots differs by domain and can be influenced by events, and a snapshot is taken irrespective of whether the page has changed since it was last captured. The depth of crawl can change over time

[20] Niels Brügger, Website History and the Website as an Object of Study, in: *New Media & Society* 11 (1-2/2019), 129, doi:10.1177/1461444808099574.

[21] Steven M. Schneider/Kirsten A. Foot, Web Sphere Analysis: An Approach to Studying Online Action, in: Christine Hine (ed.), *Virtual Methods: Issues in Social Science Research on the Internet*, Oxford 2005, 157-170.

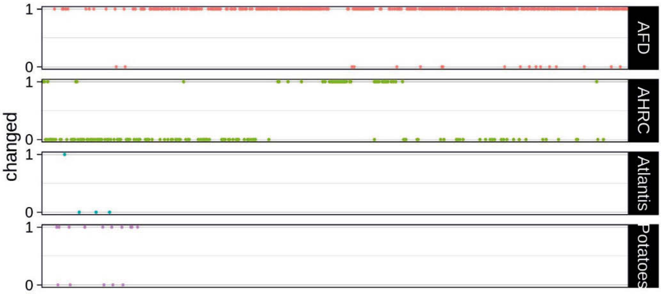
so a page archived in one crawl may have been missed in the previous one, and may not be crawled again. The researcher should be able to define the temporal aspect of their analysis by a single point of time (the closest snapshot to 1/1/2017), a range (all snapshots in the year 2015) or combinations thereof (closest snapshots to 1 January every year, all snapshots in the first quarter of each year).

An obvious starting point for a dashboard is summarized aggregations, in tabular or visual form. Using existing data held in the UKGWA in CDX format, it is relatively straightforward to produce summary statistics, or bar and line charts summarizing page/resource captures, optionally by domain, over time. These summaries give the researcher a sense of scale but they are open to misinterpretation without an understanding of the capture process. For example, setting the dimensions to homepages of all websites and all snapshots from 2003 to 2016 produces counts showing a rise from just 80 captures in 2003 to 20,429 by 2012. This value almost halves in 2014 (10,511 captures) and halves again by 2016 (5,300 captures). This can be somewhat explained by understanding frequency of capture during the period, rising from an average of only 1.4 captures per page in 2003 to a peak of 9.57 in 2012, and falling to 3.43 by 2016. The decrease in volume since 2012 was also driven by a trend of centralization of government on the web towards the gov.uk domain, with 2135 unique home pages captured in 2012, falling to 1383 in 2016. During this period the volume of resource captures has increased exponentially.

Returning to the search problem raised earlier, the dashboard could summarize search results rather than presenting a list. Analysis of the first 10,000 results of the "Budget 2010" search found 7,731 unique URLs. The most common one was www.gov.uk/government/publications/budget-2010, which appears 27 times and in this case would be a good result, although it is important not to conflate the number of snapshots with relevance. It first appears on page 4 of the search results when the search is set to return 100 results per page. Remarkably, 5,038 of the results were for URLs in the www.cotswold.gov.uk domain, the website of an English district council. The dashboard could allow the researcher to filter out domains they deem irrelevant, and if this were coupled with returning only one result per URL (with the ability to view all snapshots) the task of sifting through search results could be greatly reduced. A more sophisticated approach would be to use page contents to group them by subject matter and enable a more semantic search.

Rather than analyze overall volumes the researcher may wish to visualize change. CDX files contain checksums, file sizes, and mime types for every captured resource in the archive. Using CDX data we can visualize the capture frequency of a page and derive an indicator showing whether the checksum has changed between snapshots. Fig 2.1 shows four such visualizations, for www.armedforcesday.org.uk (AFD), www.ahrc.ac.uk (AHRC), www.projectatlantis.net (Atlantis) and potatoesforschools.org.uk (Potatoes).[22]

Fig 2.1: Changes over time derived from CDX files. Crown Copyright, licensed under the Open Government Licence.



The x-axis ranges from 20080604224039 (4 June 2008 at 22:40:39) to 20201006224341 (6 October 2020 at 22:43:41), and the y-axis for each graph is a binary value where 1 indicates that the checksum differed from the previous snapshot. Each page shows a different pattern of activity. Both AFD and AHRC were frequently (but not uniformly) captured throughout the period while the capture of Atlantis and Potatoes ceased in 2010 and 2011 respectively. While AFD appears to be under almost constant change, the AHRC page appears almost static for the first half of the period, is more active in the third quarter, and then returns to being static. The Potatoes site seems to experience intermittent change between less frequent captures, while Atlantis's activity pattern suggests a site which was first captured at the end of its active lifetime. While this is useful to understand archiving activity for a page and gives an idea of how dynamic a page may be, it is misleading in its current form and does not give an idea of what has changed.

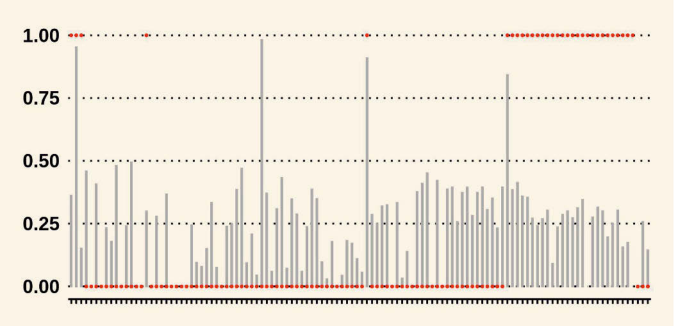
The visualization is misleading because the AHRC website was active throughout the period, and the long run of unchanged checksums is in fact due to the page being redirected for the majority of snapshots. The checksum is therefore of a redirection message and not main site content. This suggests pre-processing is required before creating a visualization of this kind, to include redirections in the

[22] Two of the URLs are self-explanatory - Armed Forces Day and Potatoes for Schools. The AHRC is the UK's Arts and Humanities Research Council; the Atlantis Initiative was "a public-sector initiative to understand the underlying issues and agree the standards to collectively provide interoperable base geographic and environmental datasets to better support water management in flooding and water quality for the twenty-first century."

analysis. This is not straightforward, however, since www.ahrc.ac.uk redirects variously to www.ahrc.ac.uk/Pages/default.aspx, www.ahrc.ac.uk/Pages/Home.aspx, and ahrc.ukri.org, and while it would not be unreasonable to disambiguate the first two with the home page, it is not obvious that the third should be treated in the same way, as it is now in the ukri.org domain. Similarly, www.eatwell.gov.uk redirects to http://www.nhs.uk/Livewell/healthy-eating/Pages/Healthyeating.aspx from April 2011. Using checksums to identify change is a blunt tool since the checksum of a file will change if only a single character is amended.

Fig 2.2 shows changes in hyperlinks on the www.ahrc.ac.uk home page over time, with the checksum changes from Fig 2.1 overlaid.

Fig 2.2: Link structure changes over time for www.ahrc.ac.uk. Crown Copyright, licensed under the Open Government Licence.



The x-axis of the graph represents snapshots, as before, while the y-axis measures the Jaccard distance of a page's links versus those on the previous snapshot. The Jaccard similarity score is a ratio of the number of items in common between two sets against the number of distinct items in the sets. The Jaccard distance is one minus the similarity, and so if the hyperlinks are identical between two pages the score is zero, while complete difference results in a score of one. The graph suggests that changes occur frequently and that there have been four occasions (the bars above 0.75) involving a major restructuring of the web page. Hyperlink based analysis can be performed using WAT files which contain metadata extracted from WARC files.[23] To improve these graphs further, they could be annotated with the aforementioned administrative data to provide context for the peaks and troughs of capture activity. The dashboard will need to allow researchers to combine different methods in this way in order to interrogate the archived data effectively.

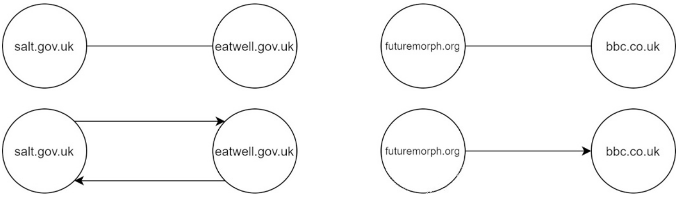
[23] https://support.archive-it.org/hc/en-us/articles/360039686611-Web-Archive-Transformation-WAT-files [last accessed: April 2, 2021].

A second type of analysis enabled by WAT files is network analysis. A network graph can be built by representing each web page with a node (visually represented as a circle), and creating an edge (visually a line) between two nodes (or pages) if they are connected by a hyperlink. The graph can be directed or undirected. In a directed graph edges can be thought of as arrows going from the source page of a hyperlink to the page it is linking to, so that a pair of nodes can have up to two edges in opposing directions between them. In the undirected case a link between two nodes would indicate that at least one of the pages has a hyperlink to the other.

It may seem an obvious point but it is important to be aware that in the directed case that there will not be incoming links from web pages which are not in the archive. This is always the case for sites outside the government web domain but can also apply to websites of which the archive was unaware, or pages which no longer exist and have not been archived.

Fig 2.3 illustrates the difference between a directed graph and undirected for two scenarios. The top row shows two pairs of nodes connected in an undirected graph. In the bottom row, the directed case, the pair on the left are connected in each direction because both sites link to each other, while on the right there is an arrow only in one direction because bbc.co.uk is not archived in UKGWA (it may still have linked back to futuremorph.org[24] but we do not have that evidence).

Fig 2.3: Connection of nodes in undirected and directed graphs. Crown Copyright, licensed under the Open Government Licence.



The previously mentioned Computational Archival Science (CAS) workshop explored network analysis of the UKGWA, comparing network structure at different points in time. The dataset consisted of hyperlinks from pages up to a depth of 2, i.e., homepages and the pages linked to by the homepages. Network graphs were generated at the page level and from an aggregated dataset at the domain level.

This experimentation leads to an important question, and a challenge. What is being visualized and summarized in the dashboard? The charts in Fig 2.1 and 2.2 were based on individual web pages, the network graphs from the CAS workshop

[24] https://webarchive.nationalarchives.gov.uk/ukgwa/20100730145942/http:/www.futuremorph.org/ [last accessed: April 2, 2021].

were summarized at the domain level. This summarization involved identifying child nodes of the home pages, i.e., pages linked to from the home page which were in the same domain, and linking two home pages if they or at least one of their children were linked. The UKGWA holds over 6 billion archived resources and the ability to aggregate is essential to making clear visualizations, or finding meaningful patterns through analytical means. This is the idea behind the *Historian's Macroscope*, which begins by envisioning a researcher zooming in and out of the archive as they follow different paths of inquiry.[25] The macroscope was first proposed by Joel De Rosnay as a theoretical tool for the study of large complex systems, analogous to the microscope or telescope.[26] Staying with the microscope analogy, we can imagine beginning with blots on a slide, one for each website, visible to the naked eye. At full magnification we see the atomic level, billions of web page elements. What do we see at intermediate magnifications, and how many steps are there between zero and maximum?

Starting at the level of the web page, the object of study consists of multiple elements including structural objects (for example, navigation menus), textual content, and images. These can all be extracted from the HTML and treated as text, in the case of images by extracting a label from the HTML or by using machine learning to generate one. The navigation menus can cause a problem as they are repeated throughout a website, creating unwanted noise and duplication of information. A common approach to this problem is boilerplate removal which strips out the menus leaving only text behind.[27] This can be problematic on home pages, for example, in which almost all of the content consists of hyperlinks, and boilerplate removal can return an empty page. Schneider and Foot suggest that studies of web content that overlook the structuring elements of a page or site are limited.[28] Rather than removing boilerplate we suggest that it should be identified and treated as a contextual object.

The website as an object of study presents a challenge, particularly when the breadth of analysis includes multiple sites. How do we represent a website, should it be treated as the sum total of its content, or is the structure important? The homepage is a high-level summary of a site but it is too distant from the content, while working with every page can provide too much detail. To study it as an object, it needs to be organized or aggregated. We have already considered one form

[25] Shawn Graham/Ian Milligan/Scott Weingart, *Exploring Big Historical Data: The Historian's Macroscope*, London 2015, 1-2.

[26] Joël De Rosnay, *The Macroscope: A New World Scientific System*, New York 1979, xiii.

[27] Marco Baroni et al., Cleaneval: a Competition for Cleaning Web Pages, in: *Proceedings of the Sixth International Conference on Language Resources and Evaluation* 2008, 1.

[28] Steven M. Schneider/ Kirsten A. Foot, The Web as an Object of Study, in: *New Media & Society* 6 (1/2004), 114-122, doi:10.1177%2F1461444804039912.

of organization, a network graph, but there are also three sources of hierarchical structure available.

First, the URL provides a natural hierarchy as it maps to a physical folder structure. Second, navigation menus generally form a tree with top-level items and submenus below them. Finally, there are breadcrumb trails placing a web page at the end of a retraceable path through the site.[29] The difficulty that these three sources pose is that they often present a different picture. The URL is influenced by the technical architecture of the website, depending on whether it serves up static or dynamic content, the underlying web framework used, and whether it uses a service-based architecture. As an example, PDF files on gov.uk are found in the sub-domain https://assets.publishing.service.gov.uk/ (not a page in its own right) but they can also be found under https://www.gov.uk/government/publications/ (either a navigational page or redirected to a search page depending on the snapshot). Neither of these provides a meaningful context for analyzing documents. Navigation menus provide more meaning and even though they may differ in form, they tend to be consistently structured across a site. The gov.uk website has presented menus in a number of ways over its eight-year lifespan but generally follows the pattern of providing links to high-level functions (e.g. Business, Driving), or shortcuts to popular services (e.g. Registering a company, driving licence applications) on the home page, and then a more traditional tree-style side menu for the rest of the site. The limitation of the navigation menu is that it does not necessarily lead to every page or resource on the site. So while it may provide a good structure around which to build a macroscope, there is more work to fit all of the pages within that structure. While this could be achieved using the hyperlinks in the WAT file, it becomes difficult when two pages from different branches of the menu link to the same page. Breadcrumb trails, where they exist, could fill this gap since they place a site in context and, if the two align, within the menu structure. Government guidance on applying for leniency for cartel members is found as a page under https://www.gov.uk/guidance, so in this case the URL provides a small amount of meaningful context (it is guidance). The breadcrumb on the 20190102181627 snapshot provides far more context, placing it under "Business and Industry" – "Business Regulation" – "Competition" – "Competition Act and Cartels."[30] Unfortunately, "Business and Industry" is not an option in the main navigation menus elsewhere in the site, which use "Business and self-employed" instead. The page originally belonged under the sub-domain of the Competition and Markets Authority and then moved under a

[29] Breadcrumb trails help users to keep track of their position within a website. They typically appear at the top of a web page, and allow users to retrace their steps within the information hierarchy.

[30] https://webarchive.nationalarchives.gov.uk/20190102181627tf_/https://www.gov.uk/guidance/cartels-confess-and-apply-for-leniency [last accessed: April 2, 2021].

section called Competition (according to the breadcrumb), which could be found under a "topic" menu (according to the URL) which itself could not be navigated to from the home page.

This single example tells a story of government on the web, which needs to be understood by the researcher and should be conveyed through the dashboard. The government web estate is constantly evolving as sites undergo architectural and structural redesigns, and responsibility for government functions moves within and between departments, which themselves may merge, close down, or be created. It also demonstrates that creating a hierarchical structure which would provide a lens through which to zoom in and out is non-trivial, and any attempt to do so will require assumptions to be made, and pre-processing steps which will change the form of the data, all of which must be transparently presented to the researcher. If such a hierarchy can be created then experimentation, including at a small scale with web data, by The National Archives has shown how the hierarchy can be navigated by summarizing the data upwards.[31] Using this approach, a level in the hierarchy is represented by an aggregation of the levels below it, rather than the text of an individual page. Further experimentation is needed to test the efficacy of this approach at scale.

An alternative is to use clustering techniques to group pages according to the similarity of some attributes. In this approach, each page is converted to a numeric form such as word frequency counts or a more sophisticated vectorized form which places each page in a 300 dimensional space.[32] Another method is to use topic modelling to identify a set of topics within the entire corpus and then classify each page according to its topic composition.[33] Pages are clustered based on some measure of similarity (for example, cosine similarity is common).[34] and a suitable number of clusters is defined either by the user or according to some optimality criteria. Again, the numeric representation selected and the clustering methodology need to be explained in a way that is understandable to the researcher. This does not mean explaining the inner workings of the algorithms and the underlying mathematics, but rather giving an understanding of high level concepts and how choices of representation and clustering technique influence the results.

Previously we suggested the ability to define web spheres around a theme was an important feature of the dashboard. A researcher may wish to define their own

[31] Mark Bell, From Tree to Network: Reordering an Archival Catalogue, in: *Records Management Journal* 30 (3/2020), 379-394, doi:10.1108/RMJ-09-2019-0051.

[32] Tomas Mikolov et al., Efficient Estimation of Word Representations in Vector Space, arXiv:1301.3781 (2013), URL: https://arxiv.org/abs/1301.3781v3 [last accessed: April 2, 2021].

[33] David M. Blei/Andrew Y. Ng/Michael I. Jordan, Latent Dirichlet Allocation, in: *Journal of Machine Learning Research* (3/2003), 933-1022.

[34] https://programminghistorian.org/en/lessons/common-similarity-measures [last accessed: April 2, 2021].

sphere by curating a set of websites which will be the subject of their analysis. This may be a difficult manual task as individual websites may not be relevant but subsections of those sites may be. For example, a study of health advice around salt consumption would include the whole of www.salt.gov.uk but only a small portion of food.gov.uk and the healthy eating section of nhs.uk. A better method would be to seed an automated approach, by compiling a list of keywords, by selecting a few exemplar websites, or perhaps by using the Wikipedia entry for an event of interest. The web sphere idea could mitigate against the complexity identified in the extraction of hierarchies by curating pages around a concept, which could be a government function without explicitly defining its position in a hierarchy.

Specifying the temporal dimension initially appears easy but its impact on results is influenced by the capture process and must be understood. An earlier example suggested selecting snapshots closest to a specific date. This sounds simple but there is nuance, such as whether there should be a limit to how far from the date a snapshot can be. If the date is 1 January 2016, is a snapshot from 2014 still relevant? Should the October 2015 version of a page take precedence over the February 2016 version considering the latter may not have existed in January? What if October in that example was replaced with July? The ability to set rules that answer these questions must be included in the dashboard. They will be informed by graphs such as Fig 2.1, so that the researcher can understand rates of capture and change for sites in their sphere of interest. An analysis of government activity during a specific month, for example, could neglect many sites which were captured at 3 or 6 month periods. Perhaps more complex rules could be defined to select the "best" snapshot from a time period. The rules can then be tested by visualizing coverage of the corpus against the other two dimensions. When working across periods of time, rules are also required for dealing with duplication, which is prevalent in the web archive. The options include selecting a single (first, last, middle) version of a page, removing duplicate versions, removing near duplicates (according to some threshold of nearness) and removing those where content is unchanged. Pages may be removed from the analysis if they are unchanged for more than some period of time, or based on analysis of hyperlinks not navigable to from other pages. These rules can be summarized as those classifying pages as active, static or dormant.

There is a lot of hidden complexity involved in defining the three dimensions, so the dashboard would include default settings which can all be adjusted. The settings used should also be exportable in an open standard so that researchers can not only easily publish them alongside their analysis but also share them with other researchers. Reproducibility should be at the heart of the design. All visualizations, statistical summaries, and intermediate representations, such as word frequency lists and topic models, should also be exportable, but content in its original form may need to be controlled. The UKGWA, as noted above, is fortunate to be an openly accessible web archive. That said, there are still risks in allowing large exports of archival material, particularly related to take down requests. There is also the issue of scale, with sites like gov.uk comprising over 2 million pages. This would be a massive download, and it would be unlikely that any archive would sanction a tool that allows their collection to be extracted at such a scale.

This leads to the next question: where is the computation performed? The Hathi Trust has an interesting model which balances their requirement to protect copyright with their aim of opening the data to researchers.[35] They make pre-processed representations of their data, such as word lists, openly available but access to the original material is through a protocol known as non-consumptive research. Instead of the researcher taking the data to their tools, they bring their tools to the data. In both cases our dashboard would be initially used to define the scope of the data, by configuring parameters along the three dimensions. In the first case, the pre-processed data could be filtered and then downloaded for further analysis by the user on their own computer. In the non-consumptive case more complicated analysis can be performed against the archive in its original form (the WARC files themselves). This does, however, mean that the computation is on the archive's infrastructure which raises the question of who pays? Charging models for workflows which will probably involve machine learning are difficult to define. In the case of a web sphere defined around a concept, it will not be possible to calculate the size of the data in advance. A deep neural network model may not converge and therefore produce no results, meaning hours of wasted computation. If that algorithm was built by the archive but the data was defined by the user where does "fault" lie?

The non-consumptive model of a researcher running their own code against the archive is an advanced form of the dashboard. What would a first version look like and how much is already available in the web archiving community? Analysis starts with a definition of scope. Depth could be defined in the same way as a web crawl, following links from the home page, then following those links, and so on. Optionally links outside the website of interest could be followed. The process should be incremental so that the researcher can receive feedback on the number of pages they are including in their analysis, and also an indication of how many links were not followed (either because they were outside the site or not archived). This functionality uses the data in the WAT files and is technically well understood. The researcher will also be able to define the elements of interest, which at first will be hyperlinks or text. Initially the breadth would be restricted to selecting individual websites, or sub-domains within sites. The UKGWA already has an A-Z browsable list which could perform this role.[36] The temporal dimension will be defined by either a single point of time, with thresholds for the allowable time periods either

[35] Jacob Jett et al., The HathiTrust Research Center Workset Ontology: A Descriptive Framework for Non-consumptive ResearchCollections, in: *Journal of Open Humanities Data* 2 (2016).

[36] https://nationalarchives.gov.uk/webarchive/atoz/ [last accessed: September 1, 2021].

side, or a date range. Having defined the scope the collection will be visualized so that the researcher can understand the relative sizes of each site, and how often each has been collected (in the case of a date range being specified). A network graph will be viewable either at the page level or website level. The text content of each page will be searchable and keyword searching can be used to filter the collection, with visualizations updating accordingly. This prototype version offers some extra functionality that is not already available but it is really the prompt for a conversation with web archive researchers to understand how they would interact with the archive as data, and with computer and data scientists to tackle the complex challenges of enabling macroscopic analysis at scale.

Thankfully, we are not starting from zero if we want to build a dashboard; there are already great tools available to build on. The Archives Unleashed project has built an open source toolkit which enables the large scale processing of WARC files.[37] The toolkit is aimed at advanced users who are comfortable working with the command line and either of the programming languages Scala and Python. Recognizing that most researchers will not have the necessary programming skills or computing power to handle large scale collections, they also have a cloud service. This service can generate network graphs and summary statistics, but it is currently only available to Archive-It subscribers. The GLAM workbench project has developed a number of Python notebooks for extracting and visualizing data from four web archives.[38] Rather than working with WARC files, they instead use the APIs of the archives to access CDX data, and analyze changes over time using the Memento protocol.[39] While not an integrated tool, the notebooks provide much of the functionality that would form the basis of a dashboard in terms of selecting snapshots, extracting text and visualizing changes over time. They are intended to encourage researchers to explore the possibilities of web archives and to understand the data they contain. Although the developers claim some level of scalability, using an institution's API may have limits and the big data tools of the Archives Unleashed toolkit may be more appropriate for large scale analysis. What the developers have also highlighted is that not every notebook works for each of the four archives, and the implementations of the APIs mean that the data that comes back from a particular query may differ for each archive, so some adaptation may be needed to apply them to the UKGWA. What is key to all of these initiatives, and to the UKGWA's approach to tool design, is the open sharing of code so that the

[37] Nick Ruest et al., The Archives Unleashed Project: Technology, Process, and Community to Improve Scholarly Access to Web Archives, in: *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020 (JCDL '20)*, New York 2020, 157-166.

[38] Tim Sherratt/Andrew Jackson, GLAM-Workbench/web-archives (Version 0.1.1), Zenodo - CERN 2020, doi:10.5281/zenodo.3894079.

[39] https://mementoweb.org/guide/quick-intro/ [last accessed: April 2, 2021].

web archiving field as a whole is able to advance, to the benefit of archivists and researchers internationally.

Conclusion

In this chapter, we have tried to address the challenges of making data available to researchers from multiple disciplines who have different levels of exposure to web archives and their complex, multi-layered contexts. Some of the individual tools and methods described above have already been trialled within the UKGWA, while others remain ideas on a whiteboard. Whether its individual components have been realized or not, our prototype dashboard is the result of a process of collaboration and co-design. The stakeholders in web archiving and web archive studies have varied disciplinary interests, work in different sectors (with different cultures and imperatives) and bring different knowledge and expertise. An open exchange of knowledge, leading to the establishment of a common language and shared assumptions, will help to engender trust in web archives, to conceive of tools that are both feasible to develop and of immediate use to researchers, to embed archival expertise in new modes of access, and to plan services that will be sustainable and extensible in the long term. We hope that the way of working we have outlined, and the prototype dashboard that we have begun to specify here, will be the beginning rather than the end of a conversation.

Bibliography

BARONI, Marco, et al., Cleaneval: a Competition for Cleaning Web Pages, in: Proceedings of the Sixth International Conference on Language Resources and Evaluation 2008, 1.

BELL, Mark, From Tree to Network: Reordering an Archival Catalogue, in: Records Management Journal 30 (3/2020), 379-394, doi:10.1108/RMJ-09-2019-0051.

BLEI, David M./NG, Andrew Y./JORDAN, Michael I., Latent Dirichlet Allocation, in: Journal of Machine Learning Research (3/2003), 933-1022.

BRÜGGER, Niels, Website History and the Website as an Object of Study, in: New Media & Society 11 (1-2/2019), 115-132, doi:10.1177/1461444808099574.

COWLS, Josh, Cultures of the UK Web, in: Niels Brügger/Ralph Schroeder (eds.), The Web as History: Using Web Archives to Understand the Past and Present, London 2017, 220-237.

DE ROSNAY, Joël, The Macroscope: A New World Scientific System, New York 1979.

GRAHAM, Shawn/MILLIGAN, Ian/WEINGART, Scott, Exploring Big Historical Data: The Historian's Macroscope, London 2015.

JETT, Jacob, et al., The HathiTrust Research Center Workset Ontology: A Descriptive Framework for Non-consumptive ResearchCollections, in: Journal of Open Humanities Data 2 (2016).

MIKOLOV, Tomas, et al., Efficient Estimation of Word Representations in Vector Space, arXiv:1301.3781 (2013), URL: https://arxiv.org/abs/1301.3781v3 [last accessed: April 2, 2021].

MıLLIGAN, Ian, Web Archive Legal Deposit: A Double-edged Sword, 14.07.2015, URL: https://ianmilli.wordpress.com/2015/07/14/web-archive-legal-deposit-a-double-edged-sword/[last accessed: April 2, 2021].

RUEST, Nick, et al., The Archives Unleashed Project: Technology, Process, and Community to Improve Scholarly Access to Web Archives, in: Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020 (JCDL '20), New York 2020, 157-166.

SCHNEIDER, Steven M./FOOT, Kirsten A., The Web as an Object of Study, in: New Media & Society 6 (1/2004), 114-122, doi:10.1177%2F1461444804039912.

SCHNEIDER, Steven M./FOOT, Kirsten A., Web Sphere Analysis: An Approach to Studying Online Action, in: Christine Hine (ed.), Virtual Methods: Issues in Social Science Research on the Internet, Oxford 2005, 157-170.

SHERRATT, Tim/JACKSON, Andrew, GLAM-Workbench/web-archives (Version 0.1.1), Zenodo - CERN 2020, doi:10.5281/zenodo.3894079.

WINTERS, Jane, Giving with One Click, Taking with the Other: Electronic Legal Deposit, Web Archives and Researcher Access, in: Melissa Terras/Paul Gooding (eds.), Electronic Legal Deposit: Shaping the Library Collections of the Future, London 2020, 159-178.