ARCHIVES AND MANUSCRIPTS
2019, VOL. 47, NO. 3, 391–403
https://doi.org/10.1080/01576895.2019.1640753

Routledge
Taylor & Francis Group

ARTICLE

Negotiating the born-digital: a problem of search

Jane Winters[a] and Andrew Prescott[b]

[a]School of Advanced Study, University of London, London, UK; [b]School of Critical Studies, University of Glasgow, Glasgow, UK

ABSTRACT

Contemporary approaches to the investigation of digital resources are dominated by the paradigm of free-form natural language search, popularised by Google. The Google form of searching has shaped our view of digital possibilities and profoundly affects our search and research habits. Yet in early pioneering work which led to the digital revolution of the 1990s, search was not a major consideration and there was a stronger emphasis on linking files. With the rise of very large born-digital resources such as e-mail archives, Wikileak dumps and web archives, the limitations of Google-type searching are becoming more evident. This paper reviews the limitations of search in exploring born-digital archives and starts to sketch out possible approaches to an alternative. It is suggested that a return to digital roots, by renewing the interest of pioneers such as Vannevar Bush and Ted Nelson in the linking of files, may provide one approach to born-digital archives.

KEYWORDS

Web archives; email archives; digital humanities; born-digital; search

In 2010, Wikileaks sprang to international prominence when it released two tranches of war logs documenting American military action in Afghanistan and Iraq. This was the first time data had been made available which documented the scale of American military action in those countries and the numbers of civilian casualties. The Afghan war logs comprised 91,000 military records, while the Iraqi files were even larger, containing 391,000 records. The huge number of documents posed great problems for the journalists from *The Guardian* and *New York Times* who worked on them. One journalist said that the experience was 'like panning for tiny grains of gold in a mountain of data'.[1] The Afghan logs were initially loaded into Microsoft Excel. One of *The Guardian* journalists recalled that:

> When I first got access to the database, it felt like being a kid in a candy shop. My first impulse was to search for "Osama bin Laden", the man who had started the war. Several of us furiously inputted the name to see what it would produce (not much as it turned out).[2]

However, Excel had serious limitations. After a while, it was realised that the spreadsheet had automatically truncated the import of the Afghan war logs after 66,000 records, so that a third of the records were missing from the journalists' initial searches.[3] A different approach was needed.

CONTACT Andrew Prescott andrew.prescott@glasgow.ac.uk

Alastair Dant, *The Guardian's* data visualiser, explained that he could create a bespoke interactive visual display of the statistics. He used as a template an interactive map of the Glastonbury music festival previously produced by *The Guardian*.[4] This visualisation enabled journalists to follow day by day and year by year the struggle of the US Army to deal with thousands of improvised explosive devices in Afghanistan. It showed how ordinary civilians were the principal victims of these devices and vividly illustrated the ebb and flow of these incidents in response to political developments. For the first time accurate statistics of the death toll in Iraq could be produced. In addition to 3,771 dead US and allied soldiers, the war logs recorded 109,032 deaths of civilians, members of the Iraqi security forces and people classed as 'enemy'.[5]

The way in which these journalists worked with this first tranche of Wikileaks material anticipated the methods that historians will in future need to adopt as they deal increasingly with born-digital historical records. The whole episode also served to illustrate the importance of who has access to what data. These logs were only available to journalists because they were leaked, an action which circumvented existing legal and national security frameworks. The stakes are not always so high, but barriers to and inequalities of access affect researchers working with all kinds of born-digital materials, and shape the kinds of analysis that can be undertaken, the types of people whose voices and stories may be represented. This is apparent from the over-representation of Twitter in social media studies, for example. Many more people use Facebook than Twitter, but only Twitter allows access to some of its data through APIs. We study Twitter because it is a fascinating source for politics, culture and society, but perhaps more importantly because we can. In many countries, the archiving of national web spheres is enabled by electronic legal deposit legislation, but this comes with more or less stringent restrictions. In the UK, access is limited to what can be viewed page by page in a reading room in one of six legal deposit libraries. The legal, commercial and security imperatives which determine access interact with the technical and methodological challenges of working with born-digital data.

The Afghan and Iraqi military logs were small-scale compared with what was to follow on Wikileaks. Later in 2010, Wikileaks released over a quarter of a million US embassy cables, some dating back to 1966. The material subsequently leaked by Edward Snowden was on an even larger scale. In 2015, Julian Assange wrote that:

> Wikileaks has published 2,325,961 diplomatic cables and other US State Department records, comprising some two billion words. This stupendous and seemingly insurmountable body of internal state literature, which if printed would amount to some 30,000 volumes, represents something new. Like the State Department, it cannot be grasped without breaking it open and considering its parts. But to randomly pick up isolated diplomatic records that intersect with known entities and disputes, as some daily newspapers have done, is to miss "the empire" for its cables.[6]

We have grown accustomed, largely because of Google, to simple keyword searching as the primary strategy in investigating online resources. A recent survey of the online practices of humanities researchers in the Netherlands came to the conclusion that 'digital research practices of Humanities scholars in the Netherlands can be condensed to three words: Just Google it'.[7] Probably our initial reaction if offered a mass of data relating to the war in Afghanistan would also be to search for 'Osama bin Laden'. However, as *The Guardian* journalists struggling to digest the first Wikileaks dumps

found, when investigating large quantities of born-digital information, Google-style searching quickly becomes ineffective. It is particularly unsuited for establishing the scope of a dataset or digital archive, as it encourages researchers to look for what they know to be there rather than to seek the unknown or to identify gaps and absences.

This article will explore the challenges posed by existing methods of working with born-digital data, and suggest some alternatives to our current over-reliance on relatively simplistic keyword searching. It will consider, first, the example of email, which has become such a dominant mode of work-place communication in particular. The email archive of the George W. Bush presidency begins to show us the scale of the problem that contemporary historians and political scientists will face. Second, the article will turn to the archived web, which encompasses everything from personal blogs to the records of government, and is characterised by volume, of course, but also by complexity. It will conclude by reflecting on what we may learn from both artificial intelligence and archival science in working towards new methods of discovery which are not bounded by a search box.

For historians and other researchers working with large email archives, corporate electronic records stores and web archives, search-based methodologies have serious limitations and new approaches are required. These approaches will probably involve some form of visualisation, and to deal with increasing amounts of information, more haptic and immersive methods of engaging with vast quantities of information need to be evolved. Probabilistic methods and artificial intelligence also have contributions to make. We are at the earliest stages of developing approaches to large-scale born-digital corpora of primary sources, but it is already evident that we need to move away from a search-orientated approach towards one that reflects classic archival methods, with an emphasis on hierarchy and context. As Assange indicates, cherry picking information from vast born-digital archives by crude free-text searching often produces misleading results. Whatever the hypothesis, it will almost always be possible to find a piece of supporting evidence. In investigating and analysing large born-digital archives, context and interrelationships are critical issues and will be fundamental in future navigation of such corpora. As Lara Putnam notes, 'digital search offers disintermediated discovery', which bypasses 'the hidden benefits of the unsheddable contextualization that makes work with analog [*sic*] sources so inefficient'.[8] In seeking to develop such approaches, we echo the concerns of such pioneers of digital information as Vannevar Bush and Ted Nelson.

Each age has felt overwhelmed by the quantity of information and has sought to develop new tools and methods to assimilate the mass of new data. In the thirteenth century, teams of Dominican friars pioneered the alphabetisation of knowledge by producing the first biblical concordances.[9] Sometime about 1320–1323, Jean de Hautfuney, afterwards Bishop of Avranches, produced an index to the sections of Vincent of Beauvais's *Speculum Historiale*.[10] The introduction of printing also saw the emergence of more consistent practices in page numbering; probably the first printed book with pagination in Arabic numerals on both sides of a page was a 1513 edition of Niccolò Perotti's *Cornucopiae*.[11] The celebrated Venetian printer Aldus Manutius carefully explained to his readers how his index worked and why it incorporated page numbers: 'a very copious index in which each word that is sought can most easily be found, since each half page throughout the whole work is numbered … with arithmetical numbers'.[12]

Ann Blair has described how the explosion of information in the sixteenth and seventeenth centuries drove the development of new scholarly methods and tools, including catalogues, indexes, encyclopaedias and common place books.[13] With the rise of industrial society, the growth in information continued apace. The management of information itself became industrialised through such inventions as duplicating and copying machines, filing cabinets and card indexes.[14] Michel Foucault saw the appearance of the card index as a key intellectual moment: 'Appearance of the index card and development of the human sciences: another invention little celebrated by historians'.[15] Montesquieu had kept notes on playing cards and the historian Edward Gibbon used playing cards for his library catalogue.[16] In the 1780s, playing cards were used to catalogue the court library in Vienna in what has been claimed as the world's first card index. In America, card index systems became very elaborate. Punched cards were used to collate the 1890 US census.[17] Libraries began to use standardised printed cards as a means of recording and sharing information about books in their collections.[18] Library catalogues were among the first information resources to be converted into a machine readable form and to be made available remotely across networks. Although this made it possible to conduct general keyword searching in library catalogues, nevertheless library catalogues remain repositories of highly structured information which is generally in a consistent format. Effective search strategies in library catalogues require a good understanding of the way the data in the catalogue has been entered.

If we regard the digital revolution as encompassing the rise of the personal PC, the growth of networks and the rise of the World Wide Web between 1985 and the early years of the 21st century, the emergence of unstructured free text searching of a wide variety of digital resources is one of the most distinctive features of that revolution. In its earliest days, the World Wide Web was small enough that it could be easily navigated by means of hierarchical directory listings. Indeed, the structure of the World Wide Web recalled that of historical archives and navigating the early web was very similar to using a historical archive. The earliest web portals took the form of guides and directory listings, such as the World Wide Web Virtual Library (vlib.org), established by Tim Berners-Lee himself in 1991. As the web grew, there was increasing demand for the capability to search for sites. The most popular of these early search services was Alta Vista, established in 1995, which pioneered an easy natural language search with a very simple interface.[19] Alta Vista declined in popularity after it became merged with the web portal Yahoo in 1998 and moved away from a streamlined search service.

While Alta Vista was an important pioneer, it was Google which made a simple natural search query using a minimalist search box the default means of interrogating digital resources. We accept Google's findings without much reflection, simply trying to hone our search terms to get the most helpful results. Google's algorithms are famously secret, but are constantly updated and search engine specialists keep a close eye on each release as small changes may have massive commercial implications. For example, a major Google update in August 2018, known as the Medic update, saw radical changes in the rankings of a number of health, medical and finance websites. As a result, traffic on the sites patient.info and prevention.com fell overnight by more than 50%, whereas the number of hits of sciencedaily.com and businessinsider.com increased by over 30%.[20] The regular Google algorithm updates doubtless have a similar effect on the rankings of more scholarly sites.

Searching Google is not like doing a keyword search in a library catalogue. Google does not rely on processing indexes to highly structured information. It uses a variety of measures (including most famously the number of links to a particular site) to rank web resources. Google attempts to provide seamless access to highly heterogeneous and varied data. Above all, in its default interface, Google accommodates highly unstructured free-form search language. You can type in key terms; you can put your query in the form of a question; you can add Boolean operators; you can even make mistakes. Ted Underwood has observed that 'In practice, a full-text search is often a Boolean fishing expedition for a set of documents that may or may not exist'.[21] If we search a library catalogue, we generally know what we are looking for (even if it is a broad category) and the results are manageable. The thousands of results produced by the free text searches of Google have to be ranked in order by complex mathematical models and, as the 2018 Medic update illustrates, we generally do not know or understand how these operate or the effect they have on the results of our search. Underwood points out that too often the results of our searches confirm our initial hypothesis in a form of confirmation bias.[22]

Despite these problems, the ubiquity and ease of use of Google has led us to expect that all digital resources can be interrogated by means of simple unstructured free text searches. Even when such methods produce poor or misleading responses, as in the case of early printed newspapers where the text is not suitable for OCR and the text searched is full of errors, we nevertheless trust in the ability of the free text search to retrieve the information we want. Search has transformed scholarly views of text and research methods. In the past, research was often either based on a comprehensive search of one very small set of primary sources or was a question of branching out from existing knowledge. A free text search enables a much more fluid and rapid form of engagement with both primary and secondary literature, vividly described by Alan Bilansky:

> direct searching, probing, chaining, "netchaining" (a species of chasing citations from one work to another that moves faster and seamlessly because all the texts are on the desktop), scanning, browsing, rereading, reading around, and assessing – often using structural elements like abstracts, conclusions, and pictures to assess without much reading.[23]

Following Renear and Palmer's 2009 study, Bilansky calls this process 'strategic reading'. Renear and Palmer point out how this process is driven by the growing quantity and complexity of information in combination with the limited amount of time for reading.[24] They compare the way scientists search through their literature with a fast-paced video game: 'They sweep through resources, changing search strings, chaining references backward and citations forward, dodging integrator and publisher sites to find open-access copies, continually working to reduce the number of clicks required for access'.[25] These methods are also very similar to James Sosnoski's strategies of hyper-reading, which include filtering, skimming, pecking, de-authorising and fragmenting.[26] Although these strategies of hyper-reading and strategic reading make use of search, they are used to rapidly develop an overview of resources available for skimming. While these strategies have been developed in response to the explosion of scholarly literature, doubt must be felt about their continued viability as the information resources used by researchers continue to grow in size. While strategic reading might enable researchers quickly to review scholarly articles on a particular topic, how

useful a tool is it when confronted with a quarter of a million diplomatic cables or an archive of 1.9 million web pages which refer to the Iraq war?

It is striking that search does not figure prominently in the early literature that fed into the development of the World Wide Web, such as Vannevar Bush's celebrated 1945 article 'As We May Think', Douglas Englebart's 'Research Center for Augmenting Human Intellect' of 1968 or Ted Nelson's 1965 paper 'File Structure for the Complex, the Changing, and the Indeterminate'.[27] All these papers were driven by the need to deal with increasing quantities of information. Bush declared that 'The investigator is staggered by the findings and conclusions of thousands of other workers – conclusions which he cannot find time to grasp, much less remember, as they appear'.[28] The procedures described by Bush are striking in their multi-media assumptions – the researcher will photograph experiments with a wearable camera, have access to a vast microfilm library in the laboratory, use voice-to-text machines to record observations and write papers, and punch card machines for calculations and storing information. Bush's description of the researcher working with the 'Memex' (a proto-hypertext system) is not of someone primarily carrying out searches. Indeed Bush is dismissive of alphabetical searching:

> Our ineptitude in getting at the record is largely caused by the artificiality of systems of indexing. When data of any sort are placed in storage, they are filed alphabetically or numerically and information is found (when it is) by tracing it down from subclass to subclass … The human mind does not work that way. It operates by association. With one item in its grasp, it snaps instantly to the next that is suggested by the association of thoughts in accordance with some web of trails carried by the cells of the brain.[29]

For Bush, the value of the Memex would be in its ability to record, store and share such trails of association and not in the search. Whereas conventional indexing was available in the Memex, the essential feature was the ability to build trails, whereby 'any item may be caused at will to select immediately and automatically another … The process of tying two items together is the important thing'.[30]

Douglas Englebart's work was directly inspired by Vannevar Bush and was similarly concerned with establishing and documenting links between information. While search was available in Englebart's Research Center for Augmenting Human Intellect, Englebart's guiding principle was the idea that 'the symbols one works with are supposed to represent a mapping of one's associated concepts, and further that one's concepts exist in a "network" of relationships as opposed to the essentially linear form of actual printed records'.[31] Ted Nelson, in introducing the term 'hypertext' to describe the structures facilitated by his Evolutionary File System, was also primarily concerned with links. Nelson summarised the key ideas of his system as: 'the inter-linking of different lists, regardless of sequence or additions; the reconfigurable character of a list complex into any humanly conceivable forms; and the ability to make copies of a whole list, or list complex – in proliferation, at will – to record its sequence, contents or arrangement at a given moment'.[32] Nelson stresses at every point the importance of linking and developing synoptic views. Rather than searching for particular words, Nelson is interested in enabling access to a linked corpus of material:

> … the ELF's cross-sequencing feature – the fact that links ignore permutations – permits the collation of very different cognate textual materials for comparison and understanding.

> In law, this would help in comparing statutes (or whole legal systems); in literature, variorum editions and parodies. Thus such bodies as the Interpreter's Bible and a Total Shakespeare (incorporating Folios, bowdlerizations, satires and all critical commentary) could be assembled for study.[33]

Just as Bush, Englebart and Nelson felt overwhelmed by the growth of information in the middle of the twentieth century, so we today confront what seem unscalable mountains of data. Maybe it is helpful for us to revisit some of the methods and concepts proposed by these pioneers in dealing with these vast information resources. Simple Google-style search won't do much.

Letters and correspondence are a fundamental source for historians and email archives will be a first port of call for historians investigating the twenty-first century. Email messages sent and received by each member of staff of the White House during the Presidency of George W. Bush are stored in the Electronic Records Archive of the US National Archives and form part of the George W. Bush Presidential Library.[34] The system contains over 200 million email messages. The electronic records for Bush's Presidency amount to over 80 terabytes. We can anticipate that the electronic archives of future Presidents may well dwarf that of Bush. There are still many restrictions on access to this website, but the sheer number of emails indicates that the historian who simply searches for 'Iraq' across this archive will retrieve an overwhelming quantity of information which it will be impossible to digest.

Historians will perhaps analyse email archives by analysing metadata rather than focussing on the detailed text of individual emails. The use of such methods by agencies like the UK's Government Communications Headquarters (GCHQ) and the National Security Agency (NSA) in the US, who scan email and text traffic for indications of terrorist activity, perhaps points the way to the sort of methods historians will use in the future. One of the most important aspects of this will be the address bar. Analysing who corresponded with whom, who was copied into particular emails and how emails are forwarded offers a powerful tool for analysing networks of communication and authority within institutions such as the White House. While a number of tools are available which visualise networks of correspondence in personal email accounts,[35] visualisation of large corpora of emails containing over one year's worth of correspondence is more problematic.[36] Moreover, while analysis of networks of correspondents is fascinating, it would be helpful to combine this with some analysis of the contents of emails. Who were in the inner circle of the Iraq discussions? Which advisors drove forward particular legislation? Subject headings may be useful but extraction of basic topic information from the contents is also required to investigate these issues. For corpora of the scale of the George W. Bush archive, new tools and approaches will be required to address these issues and historians may need to come to terms with the methodological implications of a greater reliance on metadata.

If email archives are problematic enough, they are straightforward compared to web archives. The web grows at an astounding rate and the scale of web archives is intimidating. The first ever UK domain crawl was run in 2013, using 3.8 million seed (or starting) URLs to produce 31 terabytes of data, consisting of 1.9 billion web pages and other assets. Just a year later in 2014, the UK domain crawl began with 20 million seeds and harvested 56 terabytes of data comprising 2.5 billion webpages and other assets (including 4.7 gigabytes of viruses).[37] This archiving activity is essential for the preservation of our most recent history, as the live web is shockingly ephemeral. More than 95% of the UK domain archived in 2004 is either gone or, if a particular URL is still resolvable, the content of the page has changed either

entirely or substantially. Despite the ephemeral character of the web, its importance in contemporary society means that web archives will be a fundamental resource for researchers investigating a wide range of historical, cultural and social issues.

The challenges of using these web archives go beyond their size. There is no single web archive, but rather a patchwork of different archiving activities. One of the oldest is the Internet Archive which has been archiving the web since 1996. Another important archive is the Common Crawl which has been active since 1999 and whose archive comprises petabytes of data. These archives are created using different methods and at different times, so vary significantly in the depth of data they archive. For UK historians, more specialist archives will be of interest, such as the web archive of UK government sites maintained by The National Archives or the UK parliament web archive. Many of these UK-based web archives have been supplemented at various times by data derived from the Internet Archive that relates to the .uk country code Top Level Domain, filling 'gaps' that predate the start of web archiving in the UK. This is true for the UK Web Archive at the British Library, for example. Although the need to create a legal deposit archive of the UK web presence, comparable to the archive of printed output generated by copyright legislation, was recognised early on, the necessary negotiations and legislation were protracted, and legal deposit domain crawls only began in 2013.

All these archives have content stitched together from different sources, collected at different times and in different ways. In the case of conventional archives, we can still inspect and handle the original vellum or paper documents. We can never see the 'original' of web pages; the archiving process transforms a web page into a different artefact, so that it becomes a 'reborn digital' document.[38] The process of capturing web content is very uneven in its nature and much data is often not archived, particularly multimedia content. The crawl processes are unreliable, with domain crawls often failing, so the technical context of various crawls may be different.

Web archives are also subject to change over time: they are not static archives, but transform in front of our eyes. There is an archival exemption for what has become known as the 'right to be forgotten' legislation in Europe, but archival content can and does move in and out of the publicly accessible Internet Archive. An archived website can 'appear' or 'disappear' if there are changes to the robots.txt file on its live version, and take-down notices can result in the immediate removal of material from access. The legislation governing web archiving also imposes artificial national boundaries so that web archives are often based on crawls of particular national domains, but the cross-national nature of the web means that information of interest to future researchers will not necessarily be in tidy national packages. A researcher wanting, for example, to investigate the role of the web and social media in the rise of far right populism will find it difficult to do so from an archive of the web from just one country.

Above all, web archives contain none of the contextual information that drives Google search algorithms and they cannot be searched in the same way as the live web. Web archives frequently contain multiple duplicates of web pages, which makes any kind of trend analysis of archived web information difficult. All this makes it very hard to provide meaningful ranking for search results of web archives. Web archives such as the Internet Archive's Wayback Machine, the UK Web Archive or the UK Government Web Archive offer free text search facilities but the extent to which these enable historical or other research to be undertaken is quite limited. It may be necessary

to prepare a subset of relevant sites to undertake linguistic analysis, as Harry Raffal did in his study of the use of the web by the Ministry of Defence and Armed Forces.[39] Alternatively, as with email archives, it may be that archival metadata becomes a key object of study. The British Library, for example, makes available lists of crawled links, an index of which hosts link to which, a format profile of assets in the web archive, and so on. Both Niels Brügger and Harry Raffal have shown how link analysis is potentially enormously valuable for researching web archives.[40] Brügger points out how link analysis can be important in working out the political connections of sites of individual politicians or political organisations, while Raffal shows how link analysis enables the different institutional interconnections involved in recruitment to the armed services to be traced. This kind of digital network analysis may help us to begin to delineate the shape of such a vast born-digital archive, offering a macro-level visualisation of the ecosystem of the archived web.

We are still in the very early stages of developing methods for exploring web archives and other large digital corpora as historical sources. Many of the approaches which are commonly used are limited by their reliance on particular vocabularies and index structures. The effectiveness of topic modelling for example depends on the number of topic words stipulated and it is not clear that it will work effectively with very large corpora. Web annotation has made huge strides recently with the development of stable web annotation standards, and packages such as Hypothesis (*www.hypothes.is*) enable shared work in recording and listing information in web resources. However, such manual annotation is not a very practical approach with very large resources, even if substantial groups of collaborators are assembled. Moreover, these methods are geared to annotating live web pages and their effectiveness in dealing with web archives is less clear.

In recent years, artificial intelligence, based on deep learning techniques in which computers teach themselves using neural networks, has made astonishing advances, facilitating the development of a large number of services ranging from improvements to Google's web searches and text and image recognition in social media such as Facebook through to much more accurate automated translation services and chat bots, self-driving cars and speech assistant systems such as Siri and Alexa.[41] Such new capabilities will obviously play a part in helping future researchers deal with huge digital archives such as the Bush emails, but exactly what that approach might be and how it will relate to search is not yet entirely clear. The process by which AI systems improve themselves through neural networks, although done with great speed and power, is nevertheless still a process of trial and error, and based on probabilistic assumptions. It lacks the ideological, ethical and cultural awareness which play an important part in human decision making, as is illustrated by the fate of Microsoft bot 'Tay' which had to be shut down after it started to spread racist and sexist messages, expressing support for Hitler for example.[42] The methodological and critical issues that will be posed by the use of deep learning techniques to investigate large digital corpora have barely begun to be explored, but one thing that is clear is that use of these tools will require techniques that go beyond the simple free text search.

Some experiments have been done with the use of probabilistic methods in the creation of digital editions and in the presentation of library and archive finding aids. Many linked data packages commonly used by historians, such as *London Lives 1690–1800* (*https://www.londonlives.org/*) and Digital Panopticon (*https://www.digitalpanopticon.org/*), which traces the lives of London convicts in Britain and

Australia between 1780 and 1925, depend on the deployment of probabilistic formulae to identify individuals and suggest links between them. Users are generally unaware of the assumptions lying behind the mathematical black boxes which generate the historical biographies produced by these sites. Another pioneering project has been *Traces Through Time* at The National Archives which has successfully used probabilistic methods to identify individuals in large corpora of online finding aids, providing online cues where there are other possible references to a particular individual.[43] While the initial results of probabilistic methods such as these in dealing with large corpora are promising, the critical implications, and the extent to which researchers can and should define the parameters used in such semi-automated methods, require considerable further discussion.

A more ambitious implementation of AI which offers a good pointer for the type of techniques that historians may wish to use in exploring large email and web corpora is provided by the recent work of the European Holocaust Research Infrastructure project (EHRI). The Holocaust survivor testimonials are a good example of humanities big data. The collection assembled by the Shoah Foundation contained 200 terabytes of data in 2010. The EHRI project used dictionary based approaches to create a simple sentiment analysis model for holocaust survivor testimonials. Using generative Recurrent Neural Networks, the project generated a larger training corpus of positive and negative memories and was able to train a highly accurate neural network that qualitatively and quantitatively improved the baseline dictionary model. These initial experiments were very successful. The major constraint that prevented the project developing this approach further was the lack of access to supercomputing facilities. The EHRI experiments suggest that such approaches might well be fruitful with other corpora.

Traditional approaches to archives have relied heavily on administrative hierarchies as a means of understanding the context of documents. The navigation of these hierarchies has been the traditional way to seek information in vast administrative archives which are unlikely ever to be indexed or calendared in detail. Likewise, pioneers of the web such as Vannevar Bush and Ted Nelson saw the most effective way of processing very large quantities of information as seeking and recording links and information. Indeed, the hypertextual structure of the World Wide Web looks very much like a representation of the structure of an administrative archive. Our addiction to search, fed by Google, means that we have become much less interested in and aware of these interrelationships. Paradoxically, the effect of search, which encourages us to focus on the individual document or phrase, has been to cause us to lose sight of the context of documents and, as Julian Assange put it, to 'miss the empire'.

The Google type of search is not a practicable approach to dealing with large collections of emails or web archives. We are still feeling our way for the best methods, but the way in which we have in the past approached large analogue archives offers many indications as to the best way of proceeding. This is to develop the type of 'web of associative trails' of which Vannevar Bush dreamed.

Notes

1. David Leigh and Luke Harding, *Wikileaks: Inside Julian Assange's War on Secrecy*, second ed., *The Guardian*, 2013, p. 105.
2. ibid.

3. Charlie Beckett with James Ball, *Wikileaks: News in the Networked Era*, Polity Press, Cambridge, 2012, p. 52.
4. Leigh and Harding, p. 106.
5. ibid., p. 130.
6. Julian Assange, *The Wikileaks Files: The World According to the US Empire*, Verso, London, New York, 2015, pp. 1–2.
7. Max Kemman, Martijn Kleppe and Stef Scagliola, 'Just Google It', in Clare Mills, Michael Pidd and Esther Ward (eds.), *Proceedings of the Digital Humanities Congress 2012*, The Digital Humanities Institute, Sheffield, 2014, available at <*https://www.dhi.ac.uk/openbook/chapter/dhc2012-kemman*>, accessed 21 March 2019.
8. Lara Putnam, 'The Transnational and the Text-searchable: Digitized Sources and the Shadows they Cast', *American Historical Review*, vol. 121, no. 2, 2016, pp. 377, 393.
9. Richard and Mary Rouse, 'The Verbal Concordance to the Scriptures', *Archivum Fratrum Praedicatorum*, vol. 44, 1974, pp. 5–30.
10. Ann M. Blair, *Too Much to Know: Managing Scholarly Information before the Modern Age*, Yale University Press, New Haven, 2010, p. 44.
11. ibid., p. 49.
12. ibid.
13. ibid.
14. David Shuttleton, '" . . . to whom it will be extremly Usefull." Dr William Cullen's adoption of James Watt's Copying Machine', *Journal of the Royal College of Physicians of Edinburgh*, vol. 46, 2016, pp. 127–33; JoAnnYates, 'From Press Book and Pigeonhole to Vertical Filing: Revolution in Storage and Access Systems for Correspondence', *Journal of Business Communication*, vol. 19, no. 3, 1982, pp. 5–26; Markus Krajewski, *Paper Machines: About Cards and Catalogs, 1548–1929*, MIT Press, Cambridge, MA, 2011.
15. Michel Foucault, *Überwachen und Strafen: Die Geburt des Gefängnisses [Discipline and Punish: The Birth of the Prison]*, twelfth edition, Suhrkamp Verlag, Frankfurt am Main, p. 363, n. 49 (trans. Krajewski, *Paper Machines*, p. 6).
16. Blair, p. 94.
17. Kevin Driscoll, 'From Punched Cards to "Big Data": A Social History of Database Populism', *Communication +1*, vol. 1, no. 4, 2012, pp. 7–11, available at <*https://scholarworks.umass.edu/cpo/vol1/iss1/4*>, accessed 30 March 2019.
18. Krajewski, pp. 87–122.
19. Richard Seltzer, Eric J. Ray and Deborah S. Ray, *The Alta Vista Search Revolution: How to Find Anything on the Internet*, Osborne McGraw-Hill, Berkeley, 1997, pp. 217–37.
20. <*https://searchengineland.com/googles-august-first-core-algorithm-update-who-did-it-impact-and-how-much-303538*>, accessed 21 March 2019.
21. Ted Underwood, 'Theorizing Research Practices We Forgot to Theorize Twenty Years Ago', *Representations*, vol. 127, 2014, p. 64.
22. ibid., pp. 68–70.
23. Alan Bilansky, 'Search, Reading and the Rise of the Database', *Digital Scholarship in the Humanities*, vol. 32, 2017, p. 516.
24. Allen H. Renear and Carole L. Palmer, 'Strategic Reading, Ontologies and the Future of Scientific Publishing', *Science*, no. 325, 2009, p. 829.
25. ibid.
26. James Sosnoski, 'Hyper-Readers and Their Reading Engines', in Gail Hawisher and Cynthia Selfe (eds.), *Passions, Pedagogies, and 21st Century Technologies*, Utah State UP, Logan, 1999, pp. 161–177.
27. Vannevar Bush led the US Office of Scientific Research and Development during the Second World War. In 'As We May Think', he described a proto-hypertext system, the 'Memex'. The work of the internet pioneer Douglas Engelbart was foundational in the field of human-computer interaction; and Ted Nelson, among other things, has shaped the very language we use to describe the digital, coming up with the terms 'hypertext' and 'hypermedia'. For

convenient copies of these texts, see Noah Wardrip-Fruin and Nick Montfort (eds.), *The New Media Reader*, MIT Press, Cambridge, MA, 2003, pp. 37–47, 134–45, 233–46.

28. ibid., p. 37.
29. ibid., p. 44.
30. ibid., p. 45.
31. ibid., p. 235.
32. ibid., p. 145.
33. ibid., p. 141.
34. <*https://www.georgewbushlibrary.smu.edu/en/Research/Presidential-Records*>, accessed 1 April 2019. George W. Bush's presidency was not, of course, the first to have seen the use of email, but Bill Clinton does not seem to have taken to the medium with enthusiasm. While the William J. Clinton Presidential Library does hold around 40 million emails from White House staff, there are apparently only two from the president himself. Adrienne LaFrance, 'The Truth About Bill Clinton's Emails', *The Atlantic*, available at <*https://www.theatlantic.com/technology/archive/2015/03/the-myth-about-bill-clintons-emails/387604/*>, accessed 12 June 2019. We have yet to see the scale of the email archive generated during Barack Obama's presidency, but it will undoubtedly dwarf what has come before.
35. For example, MIT's Immersion: <*https://immersion.media.mit.edu/*>, accessed 1 April 2019.
36. Tim Repke and Ralf Krestel, 'Topic-aware Network Visualisation to Explore Large Email Corpora', *EDBT/ICDT Workshops 2018*, pp. 104–107.
37. Peter Webster, 'Crawling the UK Web Domain', *UK Web Archive Blog*, available at <*http://blogs.bl.uk/webarchive/2013/09/domaincrawl.html*>; and Sabine Hartmann, '2015 UK Domain Crawl has Started', *UK Web Archive Blog*, available at <*http://blogs.bl.uk/webarchive/2015/09/2015-uk-domain-crawl-has-started.html*>, both accessed 16 May 2019. By 2017, the UK Web Archive comprised around 500 terabytes of data, with between 60 and 70 terabytes being added every year ('Frequently Asked Questions', *UK Web Archive*, available at <*https://www.webarchive.org.uk/en/ukwa/info/faq*>, accessed 16 May 2019).
38. Niels Brügger, 'When the Present Web is Later the Past: Web Historiography, Digital History, and Internet Studies', *Historical Social Research/Historische Sozialforschung*, vol. 37, no. 4, 2012, p. 104.
39. Harry Raffal, 'Tracing the Online Development of the Ministry of Defence and Armed Forces through the UK Web Archive', *Internet Histories*, vol. 2, no. 1–2, 2018, pp. 156–178.
40. Niels Brügger, 'Historical Network Analysis of the Web', *Social Science Computer Review*, vol. 31, no. 3, 2013, pp. 306–21.
41. Stefan Strauß, 'From Big Data to Deep Learning: A Leap Towards Strong AI or "Intelligentia Obscura"?', *Big Data and Cognitive Computing*, vol. 2, no. 16, 2018, p. 6.
42. ibid., p. 9.
43. Sonia Ranade, 'Traces Through Time: A Probabilistic Approach to Connected Archival Data', *2016 IEEE International Conference on Big Data*.

Disclosure statement

No potential conflict of interest was reported by the authors.

Notes on contributors

Jane Winters is Professor of Digital Humanities in the School of Advanced Study, University of London. Her research focuses on the use of born-digital archives, particularly the archived web, for historical research.

Andrew Prescott is Professor of Digital Humanities in the School of Critical Studies at the University of Glasgow. From 2012–2019, he was theme leader fellow for the Digital Transformations strategic theme of the UK Arts and Humanities Research Council.

ORCID

Jane Winters http://orcid.org/0000-0001-5502-5887
Andrew Prescott http://orcid.org/0000-0001-6474-1068