The Reconfiguration of the Archive as Data to Be Mined

MICHAEL MOSS, DAVID THOMAS, AND TIM GOLLINS[1]

ABSTRACT   This article discusses changing practices brought about by the move to online digital records, the impact these are having on the way history is written, and the way in which archivists are responding (and will need to respond in the future). We argue that digital administrative records are surrounded by other sources – online newspapers and social media – and that the huge volume of digital records alters the way historians read material. This will require a shift in approach from archivists, who will need to view archives as collections of data to be mined and not as texts to be read. Approaches to appraisal will need to be modified, and archivists will need to understand the tools and techniques used to make sense of digital records. While grappling with these issues, archivists will also need to recognize that the future record will be as much about sound and vision as about text.

RÉSUMÉ   Cet article aborde les pratiques changeantes occasionnées par le passage vers les documents numériques en ligne, l'impact de ces changements sur la façon dont l'histoire est écrite, et les façons dont les archivistes réagissent (et devront réagir à l'avenir). Nous soutenons que les documents administratifs numériques sont entourés d'autres sources – quotidiens en ligne et médias sociaux – et que la quantité énorme de documents numériques modifie la façon dont les historiens lisent le matériel. Ceci nécessitera un changement dans l'approche des archivistes qui devront voir les archives comme collections de données à exploiter et non comme textes à lire. Les approches en évaluation archivistique devront être modifiées et les archivistes devront bien saisir les outils et techniques qui servent à comprendre les documents numériques. Tout en faisant face à ces questions, les archivistes devront aussi reconnaître que le document d'archives de l'avenir sera autant une question de son et d'images que de texte.

Introduction

The move to the digital environment is having a profound impact on the archive. Digital records are voluminous and are increasingly surrounded by clouds of other resources, including social media postings and digitized newspapers. This move to the digital environment is having a profound impact on the ways history is written and texts are read. If the evolving needs of historians and other scholars working in this environment are to be met, archives will have to adjust their approach: they will have to regard their own content (i.e., the archives) as data to be mined and work collaboratively with other organizations (e.g., specialist web archives, newspaper archives, and data archives). Approaches to appraisal will need to be modified, and tools will need to be developed to process and make sense of the mass of heterogeneous data that will comprise the modern archive. Additionally, archivists will need to recognize and acknowledge the threat to the textual tradition as records are increasingly created in other formats.

In this article, we discuss the development of the documentary approach to history and the emergence of archives. We then consider the development of online catalogues and digital records and the new resources that these have created for the study of history. As we demonstrate, the enormous scale of digital records has changed the way scholarly resources are read; increasingly, the emphasis is on being able to study large volumes of material rather than on being able to study individual texts. These radical changes mean that, in future, archives will no longer be conceived of as collections of texts, but as *data to be made sense of*. Clearly, this has implications for the way appraisal will be conducted in the future, and there will be a greater need to take account of the requirements of users and to be aware of the way they will "read" digital records in the future. Finally, we discuss the shift from a purely textual archive to one incorporating sound and vision, which is surely the most significant and radical challenge facing archives in the next decade.

The Development of the Documentary Approach to History

Since the early 19th century, British, European, and North American history has been dominated by the European or German school of historiography – often referred to as the scientific or documentary method – known as *Wissenschaft*

and pioneered by Leopold von Ranke. In the United Kingdom, one of the leading scientific historians with a deep interest in archives was Mandell Creighton, the first Dixie Professor of Ecclesiastical History at Cambridge, who shared Ranke's interest in the history of the papacy. He was the first editor of the *English Historical Review*, the first edition of which opened in 1886 with an essay by Lord Acton, himself partly German, on "German Schools of History." Although Acton was known to be skeptical of the German school, he believed that "[h]istory is better written from letters than from histories."[2] Although Creighton abandoned the academy for an episcopal see, which at the time was much better compensated, his memorial in London's St. Paul's Cathedral bears the legend, "He tried to write true history." Richard Evans, the recent Regius Professor of History at Cambridge, explains that Creighton's magnum opus, *History of the Papacy during the Period of the Reformation*, "constitutes one of the first great attempts to introduce the British to explicitly modern and European history"[3] – by which he means history rooted in documentary evidence. Ranke and Creighton were pioneers in an approach to history that relied on the close reading of records and the systematic examination of all available documentation, rather than on a haphazard antiquarian perspective often based on hearsay. This new approach to history relied on access to the national archives that were then opening across Europe, making it possible, as Lord Acton stated in his celebrated inaugural lecture as Regius Professor of Modern History at Cambridge in 1895, "to penetrate the mysteries of state."[4]

The foundation of national archives in the United Kingdom in 1838 was followed by efforts to rescue the records of the great families through the formation of the Royal Commission on Historical Manuscripts (HMC) in 1869. Toward the end of the century, the Duke of Northumberland and Creighton led a shift to focus on diocesan and local administrative records. Although Creighton admitted privately that "local archives are hardly my business," he was appointed chairman of the local records committee in 1899. He was assisted by a

---

2   Online Library of Liberty, John Emerich Edward Dalberg, Lord Acton, Acton-Creighton Correspondence, 1887, Acton to Creighton, Letter II., n.d., http://oll.libertyfund.org/titles/2254.

3   R.J.W. Evans, "The Creighton Century: British Historians and Europe, 1907–2007," *Historical Research* 82, no. 216 (2009): 320–39.

4   Online Library of Liberty, John Emerich Edward Dahlberg, Lord Acton, The Best of the OLL No. 36: Lord Acton, Inaugural Lecture on the Study of History, 1895, 6, http://oll.libertyfund.org/titles/acton-boll-36-lord-acton-inaugural-lecture-on-the-study-of-history-1895.

committee that was arguably the most influential group ever to address records provision in the UK: Sir Francis Mowatt (1837–1919), Permanent Secretary of the Treasury; Sir Henry Churchill Maxwell Lyte (1848–1940), Deputy Keeper of the Records; Sir Courtenay P. Ilbert (1841–1924), Parliamentary Counsel to the Treasury and later Clerk to the House of Commons; and S.E. Spring Rice (1856–1902), Principal Clerk in the Treasury.[5] Creighton died suddenly before the committee could report and was replaced by the historian and politician James Bryce, who oversaw the publication of a far-sighted report based largely on Creighton's draft, which called for the creation of local record offices across England and Wales.[6] Nevertheless, it took many decades to fulfill this vision. As local record offices fitfully emerged, mostly after the Second World War, their lists and catalogues were added to the National Register of Archives (NRA), which was set up in 1945 by the HMC in the wake of heroic efforts by members of the British Records Association to rescue records at risk in wartime.[7] The NRA indexed the lists by subject and by the names of those included in the *Dictionary of National Biography*. This approach had its deficiencies; there were very few women or individuals from business backgrounds in the dictionary.[8] The indexes could only be consulted in the HMC's offices in central London.

Online Catalogues: Weaknesses and Possible Solutions

Since the 1980s, access to archives has been progressively expanded as catalogues have been made available online. This has made locating records relevant to a particular question quicker and more straightforward, particularly in large, complex collections or where material can turn up in multiple locations. However, the danger of online catalogues and the search engines that support them is that, regardless of their claims or the perceptions of researchers, they

---

5   Hannah Mary Little, "Genealogy as Theatre of Self-Identity: A Study of Genealogy as a Cultural Practice within Britain Since c. 1850" (PhD diss., University of Glasgow, 2010), 98–99.

6   Ibid., 99.

7   Roger H. Ellis, "The Historical Manuscripts Commission, 1869–1969," *Journal of the Society of Archivists* 2, no. 6 (1962): 233–42, doi:10.1080/00379816009513740; Ian Mortimer, "ARCHON and the Indexes to the National Register of Archives: Past, Present and Future," *Program* 33, no. 3 (1999): 225–33, doi:10.1108/EUM0000000006918.

8   These omissions have since been corrected by the *Oxford Dictionary of National Biography*.

are neither objective nor complete.[9] Just as Creighton's scholarship reflected his interest in the "Englishness" of British institutions, contemporary online catalogues reflect the interests and preoccupations of those who compile them.[10] Inevitably, there will be more in any collection than is described in the catalogue, and, for the most part, researchers will need to examine the documents, obtain copies, and/or download images (if they have been made available digitally) in order to discover precisely what the records cover. At Newcastle University, researchers led by Tom Schofield conducted a study to understand the users of the archive of the poetry publisher Bloodaxe Books. Their initial research showed that "virtually all of the aspects identified as interesting by the participants were not intended to be described in metadata in the forthcoming catalogue, and thus would not be represented in future interfaces to the archive." The inherent bias of catalogues has been the subject of a good deal of recent scholarship. Schofield and his team called attention to the problem of claiming neutrality:

> One of the motivations for maintaining an ostensibly neutral catalogue based on international standards for record keeping is to defer the interpretative moment outside of the catalogue and its interface to allow for a plurality of approaches later on. We sympathise with this position but it is hugely problematized of course by the nature of any claims to neutrality. As Drucker has pointed out, the design of our interfaces already embodies an epistemological commitment inherited from particular cultural traditions and emphasising hierarchy, standardization, and objectivity.[11]

9   James Currall, Michael Moss, and Susan Stuart, "Privileging Information is Inevitable," *Archives and Manuscripts* 34, no. 1 (2006): 98–122.

10   Robert Harrison, Aled Jones, and Peter Lambert, "Methodology: 'Scientific' History and the Problem of Objectivity," in *Making History: An Introduction to the History and Practices of a Discipline*, ed. Peter Lambert and Phillipp Schofield (London: Routledge, 2004), 46.

11   Tom Schofield, David Kirk, Telmo Almaral, Marian Dörk, Mitchell Whitelaw, Guy Schofield, and Thomas Ploetz, "Archival Liveness: Designing with Collections Before and During Cataloguing and Digitization," *Digital Humanities Quarterly* 9, no. 3 (2015), accessed 17 December 2017, http://www.digitalhumanities.org/dhq /vol/9/3/000227/000227.html.

As long ago as 1987, Bruce Dearstyne, writing in the *American Archivist*, voiced a very similar criticism.[12] He quoted Roy Turnbaugh, who had suggested that "archivists produce finding aids that are either ignored or are difficult to use and that archivists cling to outdated concepts inappropriate for modern researchers' approaches and needs."[13] Heather MacNeil and Jennifer Douglas are more muted but echo Schofield and his colleagues in calling for consultation with users to get a better sense of their needs and expectations.[14] Sociologist Deb Verhoeven emphatically endorses the co-production of knowledge.[15] More recent criticisms of archival cataloguing were discussed by Stacy Wood and colleagues in 2014.[16]

Online catalogues, on the whole, have been produced to describe physical records and are often simply the translation of haphazard paper catalogues with little attempt at normalization. However, in the past ten years or so, archives have begun to acquire increasing quantities of digital records, which are often made available online.

Online Records

Digital copies of records can exist in two forms: either as simple electronic images, which can be read but not manipulated; or as electronic text, which can be manipulated. At present, many electronic texts (for example, the billions of records housed by the family history company Ancestry) are created by human transcription of handwritten or typewritten material; but in the future, it is possible that even handwritten material could be scanned and interpreted by machines.[17] At present, the written record is also supplemented by a plethora of

---

12   Bruce W. Dearstyne, "What is the Use of Archives? A Challenge for the Profession," *American Archivist* 50, no. 1 (1987): 76–87.

13   Roy C. Turnbaugh, "Living With a Guide," *American Archivist* 46, no. 4 (1983): 451.

14   Heather MacNeil and Jennifer Douglas, "Generic Evolution and the Online Archival Catalogue," *Archives and Records* 36, no. 2 (2015): 107–27.

15   Deb Verhoeven, "Doing the Sheep Good: Facilitating Engagement in Digital Humanities and Creative Arts Research," in *Advancing Digital Humanities Research, Methods, Theories*, ed. Paul Longley Arthur and Katherine Bode (Basingstoke, UK: Palgrave Macmillan, 2014).

16   Stacy Wood, Kathy Carbone, Marika Cifor, Anne Gilliland, and Ricardo Punzalan, "Mobilizing Records: Re-framing Archival Description to Support Human Rights," *Archival Science* 14, no. 3–4 (2014): 397–419.

17   See for example Adam Matthew, "Artificial Intelligence Transforms Discoverability of 17th and 18th Century

digitized printed material which, because it has been further processed into a textual form, can be interrogated by search engines. Although the quality of the machine-readable digital surrogates has been criticized,[18] examination of such content often produces references in surprising places or so deeply buried that finding them in the physical world happens largely by accident, if at all. Who would ever have thought there was such a periodical as *The Catholic Presbyterian* or that it might contain references to a very Anglo-Catholic cathedral on the Island of Cumbrae in the Firth of Clyde on the west coast of Scotland?[19]

An impression of how printed material can be used to supplement (or even substitute for) records in the digital age can be seen in the online material relating to the notorious 1938 Haw Bridge murder in Gloucestershire, England; researchers interested in what is often referred to as the Cheltenham torso mystery (where the headless body of an unknown man was found in the river Severn) can find the information they need about the case simply by Googling. There, available to all, are photographs of the scene, accounts of the event, links to a book about the crime, speculation as to the identity of the victim and his killer, as well as a link to The National Archives of the UK (TNA), where the relevant records are closed until 2029.[20] The long closure can be explained by government policy to neither confirm nor deny press reports, even if they are self-evidently correct.

It is not only scanned images of books and newspapers that now crowd around digital records; it is social media postings as well. This poses a radical

Manuscripts Using Handwritten Text Recognition," news release, 14 September 2017, accessed 10 December 2017, http://www.amdigital.co.uk/m-news/htr/.

18   Simon Tanner, Trevor Muñoz, and Pich Hemy Ros, "Measuring Mass Text Digitization Quality and Usefulness: Lessons Learned from Assessing the OCR Accuracy of the British Library's 19th Century Online Newspaper Archive," *D-Lib Magazine* 15, no. 7–8 (July/August 2009), accessed 8 December 2017, http://www.dlib.org/dlib/july09/munoz/07munoz.html.

19   Michael Moss, "The High Price of Heaven – The 6th Earl of Glasgow and the College of the Holy Spirit on the Isle of Cumbrae," *Architectural Heritage* 22, no. 1 (2011): 77–98; William Garden Blaikie, "Presbyterian Church," *Catholic Presbyterian* 8 (1882), 132.

20   Richard Whittington-Egan, *Great British Torso Mystery* (Liverpool: Bluecoat Press, 2002); national archives file: "Suspected Murder at Haw Bridge, Cheltenham, and Recovery of Human Remains from River . . ." National Archives, accessed 13 December 2017, http://discovery.nationalarchives.gov.uk/details/r/C1257137; local archives file: "Haw Bridge Murder Enquiry, 1938," Gloucestershire Archives, accessed 13 December 2017, http://discovery.nationalarchives.gov.uk/details/r/8ffce2be-75bc-40bd-a365-1b6057f26291; records from local police: "Haw Bridge Murder," Gloucestershire Police Archives, accessed 13 December 2017, http://gloucestershirepolicearchives.org.uk/content/moving-with-the-times/crime-and-criminals/haw-bridge-murder.

challenge to archives. When TNA decided to archive the records of the 1982 Falklands Conflict, it was comparatively easy; it simply acquired the relevant official records from the Ministry of Defence, the Foreign and Commonwealth Office, the Cabinet Office, the Prime Minister's Office, etc. These included a range of files concerned with media coverage of the war, which entirely focused on the mass media of the day: newspapers and television.[21] Now, however, war is not only recorded in official files and conventional mass media outlets but also captured in real time in a range of other media. Anna Reading explored this transformation by comparing the London underground bombing of 2005 with the London underground bombing of 1897. News of the 1897 bombing spread slowly, not reaching one New Zealand newspaper for two months. Reading argues that "digital media technologies have not simply collapsed the event and its memory into one another, as, perhaps, it may seem at first sight. Rather, events are witnessed in time and people's mediated witnessing, including mobile witnessing of events is articulated, re-articulated and disarticulated through intersecting temporalities."[22] Andrew Hoskins refers to what he terms the "connective turn": "the sudden abundance, pervasiveness and immediacy of digital media communication networks and archives [, which] forces a view unprecedented in history. This turn forces a shift in what memory is and what memory does."[23] What both Reading and Hoskins argue is that instant imaging and recording of events has turned the media and the archive on its head and has collapsed or at least reconfigured what we mean by time. This is contested territory, but one the archivist cannot ignore as it lies at the heart of the digital revolution and engages reflection by prominent thinkers, such as Bruno Latour and Wolfgang Ernst. It also raises important questions about the persistence of many resources held behind commercial firewalls, which will never be captured by the Internet Archive.

William Merrin, a lecturer in media studies, demonstrates how social media has changed the nature of the record and even the nature of warfare:

---

21   The Falklands War press files can be identified by searching the UK National Archives Discovery catalogue at http://discovery.nationalarchives.gov.uk/.

22   Anna Reading, "The London Bombings: Mobile Witnessing, Mortal Bodies and Globital Time," *Memory Studies* 4, no. 3 (July 2011): 298–311.

23   Andrew Hoskins, "The Restless Past: An Introduction to Digital Memory and Media," in *Digital Memory Studies: Media Pasts in Transition*, ed. Andrew Hoskins (New York: Routledge, 2017).

With the missile strike on Al-Jazeera's Kabul office on 13th November 2001 the US made it clear that their concept of "full spectrum dominance" developed in Joint Vision 2010 (1996) included the broadcasting as well as the enemy electromagnetic spectrum. As they'd demonstrate in Iraq in 2003, media were now either imploded into the military system or were legitimate targets of information warfare. With this division the US authorities thought they had perfected their system of media-management introduced in the 1991 Gulf War. And then came personal digital media and the end of informational control. It began with Abu Ghraib's digital cameras and military and civilian blogs from Iraq and within a few years mobile technologies, ubiquitous connectivity, and popular Web 2.0 platforms had transformed the informational ecology. From now on we had full spectrum access: individual soldiers were their own embedded self-journalists, war-zone civilians began sharing their experiences and images, governments, military, militias and terrorist groups all had to get social-media-savvy with their own channels and modes of distribution, and ultimately anyone, anywhere, could join in, passing on content, commenting and critiquing and adding their own home-made, bottom-up propaganda for any cause or side they wanted.[24]

It is easy to criticize the quality of digitized printed material[25] and, in so doing, to fail to appreciate the way such resources, available from desktops, have transformed research and will continue to do so – even more as machines become more competent at directly processing handwriting.[26] Much of the skepticism about the research value of this material amounts to little more than sour grapes, as Dan Cohen surmises in his blog post "Is Google Good for History?":

24   William Merrin, "Full Spectrum Access: The Problem of #Participative War," paper presented at Archives of War: Media, Memory and History Conference, The National Archives, Kew, 30 November 2015, accessed 14 August 2018, https://www.jiscmail.ac.uk/cgi-bin/webadmin?A2=CULTURAL-MEMORY;cf3ab6f7.1511; Craig Robertson, ed., *Media History and the Archive*, (London: Routledge, 2014).

25   See for example Laura Miller, "Is Google Leading an E-book Revolution?" *Salon*, 8 December 2010, accessed 14 August 2018, https://www.salon.com/2010/12/08/google_ebookstore/.

26   Matthew, "Artificial Intelligence."

But then we historians, like other humanities scholars, are natural-born critics. We can find fault with virtually anything. And this disposition is unsurprisingly exacerbated when a large company, consisting mostly of better-paid graduates from the other side of campus, muscles into our turf. Had Google spent hundreds of millions of dollars to build the Widener Library at Harvard, surely we would have complained about all those steps up to the front entrance.[27]

Reading Documents in the Digital Era

Since the publication of Cleanth Brooks and Robert Penn Warren's *Understanding Poetry* in 1938, the predominant method for the analysis of literary works has been what is called *close reading*, a detailed study of texts focusing on the work of art as an autonomous object that can be analyzed on its own terms.[28] By extension, this method of analyzing texts each by each and one by one has been applied to humanities scholarship more generally. In 2000, the Italian literary scholar Franco Moretti became concerned with the idea of world literature. Realizing that there was no possibility of reading more than a tiny proportion of global literature, he proposed a new approach, which he called *distant reading*, in which scholars, rather than studying texts through close reading, could search literary history for themes. As he said, "Distant reading allows you to focus on units that are much smaller or much larger than the text: devices, themes, tropes – or genres and systems. And if, between the very small and the very large, the text itself disappears, well, it is one of those cases when one can justifiably say, less is more."[29]

Moretti's views have shifted over the years as he has become more concerned with digital texts, and his ideas have become generalized beyond literature.[30] There is an increasing consensus that humanities scholars' traditional approaches

---

27   Dan Cohen, "Is Google Good for History?" blog post, 7 January 2010, accessed 12 December 2017, http://www.dancohen.org/2010/01/07/is-google-good-for-history/.

28   Garrick Davis, "The Well-Wrought Textbook: A Look Back at Brooks and Warren's College Classic, *Understanding Poetry*," *Humanities* 32, no. 4 (July/August 2011).

29   Franco Moretti, "Conjectures on World Literature," *New Left Review* 1, (January/February 2000): 54–68.

30   Franco Moretti, *Distant Reading* (London: Verso, 2013).

to texts will not work in the digital era. Canadian historians and digital humanities scholars William Turkel, Kevin Kee, and Spencer Roberts argue that close reading of texts is impossible in the digital age – or, as Roy Rosenzweig has called it, the "age of abundance."[31] Turkel, Kee, and Roberts quote Dan Cohen, who points out that a single historian, who might have been able to read the 40 thousand memos issued at the White House by the Johnson administration, certainly could not handle the four million emails sent out while Clinton was in office.[32]

There are those who regret the decline of close reading, arguing that there is a loss of physical interaction with the materiality of the record and the serendipity that accompanies it.[33] This overlooks the serendipity of deep searching on the Web and the use of analytical tools to help interpret data. As Cohen puts it,

> The existence of modern search technology should push us to improve historical research. It should tell us that our analog, necessarily partial methods have . . . hidden from us the potential of taking a more comprehensive view, aided by less capricious retrieval mechanisms which . . . are often more objective than leafing rapidly through paper folios on a time-delimited jaunt to an archive.[34]

Cohen's criticism of research in the analog paradigm is reminiscent of the German school's criticism of antiquarians.

In the digital order that we now inhabit, everything is available online, and the boundaries between the archive and print culture dissolve; the sheer quantity of data makes close reading impractical, and users of necessity have to rely on search algorithms and sophisticated analytical tools to make sense of the totality of the record. The problem of an increased volume of data and the need to view systems as a whole rather than as a collection of elements led the French molecular biologist and systems thinker Joël de Rosnay to propose the idea of a *macroscope* – a purely theoretical machine that would make it possible

---

31 Roy Rosenzweig, "Scarcity or Abundance? Preserving the Past in a Digital Era," *American Historical Review* 108, no. 3 (2003): 735–62.

32 William J. Turkel, Kevin Kee, and Spencer Roberts, "A Method for Navigating the Infinite Archive," in *History in the Digital Age*, ed. Toni Weller (London: Routledge, 2013), 61–75.

33 This issue is hotly debated by David Bearman and Martin Hall in *Refiguring the Archive*, ed. Carolyn Hamilton, Verne Harris, Jane Taylor, Michèle Pickover, Graeme Reid, and Razia Saleh (Dordrecht, NL: Kluwer, 2002).

34 Cohen, "Is Google Good for History?"

to see both infinitely small and infinitely large things at the same time.[35] De Rosnay's concept was really a metaphor for what might best be described as a humanities workbench. A good example of what a macroscope might look like in reality comes from the field of geography where, according to Jerome E. Dobson from the University of Kansas, "collectively, GIS, GPS, satellite remote sensing, and popular geographics constitute a *macroscope* that allows scientists, practitioners, and the public alike to view the earth as never before." Dobson describes how, by using this technology, two scientists have been able to demonstrate that cattle and deer could sense magnetism – not by measuring the animals' body chemistry or analyzing their brainwaves, but through "analysis of satellite images, field observations, and measuring 'deer beds' in snow."[36]

The possibility of using a macroscope in historical research is explored by Jo Guildi and David Armitage in their *History Manifesto* and described by Shawn Graham, Ian Milligan, and Scott Weingart.[37] Guildi and Armitage argue that, once armed with a "'macroscrope' . . . historians should pursue an analysis of how 'big data' might be used to re-negotiate the role of the historian – and the humanities more generally."[38] A macroscope for modern military historians would have to enable them to read official files, news media, tweets, Facebook pages, YouTube videos, websites, and even historical weather reports.

A simple example of the application of macroscopy to history is the ease with which it is now possible to create visual representations of long strings of numeric data abstracted from analog records using statistical packages. This technique enables the straightforward generation of comparative time series. Research into the import of cinchona bark, from which quinine (a prophylactic against malaria) is extracted, into the UK during the 19th century shows it to be, surprisingly, a predominantly entrepôt trade supplying the German pharmaceutical industry. This analysis could have been done with pen and paper, but it

---

35   Joël de Rosnay, *Le Macroscope: Vers une vision globale* (Paris: Éditions du Seuil, 1975).

36   Jerome E. Dobson, "Through the Macroscope: Geography's View of the World," *ArcNews* (Winter 2011/2012), accessed 2 December 2017, http://www.esri.com/news/arcnews/winter1112articles/through-the-macroscope -geographys-view-of-the-world.html.

37   Shawn Graham, Ian Milligan, and Scott Weingart, *Exploring Big Historical Data: The Historian's Macroscope* (London: Imperial College Press, 2016).

38   Tim Hitchcock, "Big Data, Small Data and Meaning," Historyonics (blog), 9 November 2014, accessed 14 August 2018, http://historyonics.blogspot.com/2014/11/; see also Jo Guldi and David Armitage, *The History Manifesto* (Cambridge: Cambridge University Press, 2014).

would have been hugely time-consuming, involving the laborious entry of each value on graph paper. It now requires only keystrokes: data is entered, assigned a value and manipulated automatically.[39]

Beyond the Administrative Archive

Simply put, the archive has moved from being a collection of largely administrative records to become a collection of data to be made sense of. In a process described by Andrew Prescott and Tobias Blanke as "datafication of the humanities," evidence in all its various formats is being treated as data to be mined.[40] In the process, long-cherished shibboleths of archivists, such as hierarchies and original order, are rendered meaningless. Think of the way that, with a click of a key, the order of a catalogue can be changed. Datafication is ubiquitous. There is hardly a corner of our lives that has not been datafied; everything from our personal preferences, as the recent Facebook scandal has illustrated,[41] to pervasive surveillance by security services. It is essential that archivists reconceptualize the archive in this way.

By reconceptualizing the archive as data 'to be made sense of,' archivists are faced with a huge challenge, however: not only is the quantity of records to be selected for preservation in the digital order likely to increase, but these records are also accompanied by a vast mass of social media data to be addressed. This social media data may or may not have helped shape the records or indeed may be part of the records, but its potential value cannot be ignored.

As we described above, warfare is no longer recorded only in official files but also in real time, and its outputs are broadly spread across the Internet – on YouTube, Twitter, blogs, etc. Capturing this for the archive requires a huge effort and a radical change in perspective. Traditionally, archives have captured

---

39   The records from which the data is extracted are held in the Collectors of customs (CUST) series in The National Archives of the UK (TNA) and are part of a research project on the health of the navy during the 19th century, funded by the Wellcome Trust and led by Professor Laurence Brockliss at Magdalen College, Oxford, from whom further details can be obtained.

40   Tobias Blanke and Andrew Prescott, "Dealing with Big Data," in *Research Methods for Reading Digital Data in the Digital Humanities*, ed. Gabriele Griffin and Matt Hayler (Edinburgh: Edinburgh University Press, 2016), 185–205.

41   Kevin Granville, "Facebook and Cambridge Analytica: What You Need to Know as Fallout Widens," *The New York Times Online*, 19 March 2018, accessed 17 August 2018, https://www.nytimes.com/2018/03/19/technology/facebook-cambridge-analytica-explained.html.

records of official bodies; the UK National Archives captures records of the UK central government, although arrangements for collecting private records that inform the history of the state do exist for other jurisdictions.[42] These legal frameworks permit future collections to be more contemporaneous, capturing records of private organizations' websites and, in the future, possibly also of social media platforms.[43]

Because battlefield technology now captures vast amounts of data, preserving the records of modern warfare requires a broad vision and a willingness to act at speed. It can be done: the 2003 Iraq War has been heavily archived, including by the US Library of Congress and the US National Archives, but much of the most significant material has been made available by independent, not-for-profit organizations including the Internet Archive, with its collection of videos; the independent National Security Archive, with its collection of official and unofficial documents; and Wikileaks, with its leaked material.[44] The year 2003 was a long time ago, however, and before the huge outburst of social media activity. In 2018, to capture the records of a war or similar global event would require a radical change in approach and the reconceptualization of archives as data to be captured on a huge scale. We are not suggesting a hegemonic future for national archives, but rather, the persistence and diversification of the plural provision that characterizes current arrangements.

At present, it seems unlikely that it will be possible to capture anything beyond a small part of the archive of anything as big as a future war. The problem lies partly in the technical complexity of the Internet and of battlefield technology. As David Karpf, of George Washington University, has said,

---

42 *Public Records Act (Northern Ireland) 1923*, http://www.legislation.gov.uk/apni/1923/20/contents; *Public Records (Scotland) Act 1937*, http://www.legislation.gov.uk/ukpga/Edw8and1Geo6/1/43/contents, each as amended.

43 "National Records of Scotland (NRS) Web Continuity Service," 14 August 2018, https://www.nrsscotland.gov.uk/research/researching-online/web-continuity-service; "Selection Policy for the National Records of Scotland Web Continuity Services," National Records of Scotland, accessed 14 August 2018, https://www.nrscotland.gov.uk/files/research/nrs-web-continuity-selection-policy-February-2017.pdf.

44 "Iraq War 2003 Web Archive," Library of Congress, accessed 12 December 2017, https://www.loc.gov/collections/iraq-war-2003-web-archive/about-this-collection/; "War in Iraq," US National Archives and Records Administration, accessed 12 December 2017, https://www.archives.gov/research/alic/reference/military/war-in-iraq.html; "The Iraq War Collection," Internet Archive, accessed 12 December 2017, https://archive.org/details/iraq_war; "The Iraq Project," National Security Archive, accessed 12 December 2017, http://nsarchive.gwu.edu/project/iraq-project; "Iraq War," WikiLeaks, accessed 12 December 2017, https://wikileaks.com/wiki/Category:Iraq_war.

> The Internet is unique among Information and Communica-
> tions Technologies (ICTs) specifically because the Internet of
> 2002 has important differences from the Internet of 2005, or
> 2009, or 2012. It is a suite of overlapping, interrelated technolo-
> gies. The medium is simultaneously undergoing a social diffusion
> process and an ongoing series of code-based modifications.[45]

As Internet philosopher David Weinberger has said, the Internet is both unstable and "consists of billions of sub-networks." It is estimated that, already, 2.3 billion people are creating and sharing personal memories on social media.[46] Andrew Hoskins has argued that "the Internet can hardly be conceived of as a single medium and its transformations are more staccato rather than smoothly evolutionary."[47] Sadly, much of the Internet content that might be of significance for the future record of a major event (for example, content on Facebook, YouTube, and Twitter) is in walled gardens owned by for-profit organizations.

What is needed is for national, local, and specialized archives to accept that their holdings are data to be made sense of and to work with other web archiving organizations to develop specific ad hoc collections that can be brought together online, possibly by third-party aggregators. For example, both local and central government records and the websites and social media records relating to the 2017 Manchester Arena bombings in the UK could be made searchable from the same site. The drawback is that commercial providers jealously guard, aggregate, and sell their data and are likely to be reluctant to share. As Tim Harford argues, the commercial behemoths have become too powerful and are unaccountable for the huge stores of data they hold, analyze, and sell.[48]

Thomas Padilla, Visiting Digital Research Services Librarian at the University of Nevada, Las Vegas, has written and spoken about seeing collections as data. By this, he means that collections need to be transformed into ordered data that is amenable to computation. Without using the word *context*, he makes

45 David Knapf, "Social Science Research Methods in Internet Time," *Information Communication & Society* 15, no. 5 (2012), 640.

46 David Weinberger, *Too Big to Know: Rethinking Knowledge Now That the Facts Aren't the Facts, Experts Are Everywhere, and the Smartest Person in the Room Is the Room* (New York: Basic Books, 2011), 58.

47 Andrew Hoskins, "The Mediatization of Memory," in *Mediatization of Communication*, ed. Knut Lundby (Berlin: De Gruyter Mouton, 2014), 661–80.

48 Tim Harford, "Algorithms of the World, Do Not Unite!" *Financial Times*, 2–3 December 2017.

the important point – very familiar to archivists – that what is on the surface is not all there is. He gives the example of tweets: on the surface, tweets are 140-character communications, but beneath the surface, they might involve geolocation and timestamping; links to webpages and images; and a wide array of language and data that records relationships between Twitter users. He emphasizes the need to pay attention to form, description, discovery, access, and most importantly, partnerships with users. He also emphasizes the need for data integrity and improved documentation, so that it is possible for users to interact not with digital objects but with the data that make up those objects.[49]

Recent discussions about the use of web archive collections by researchers demonstrate how these collections can be treated as data.[50] The potential exists not just in terms of searching – although the challenges of searching an intrinsically duplicative web archive require the development of novel search algorithms – but also in terms of treating the archive as a set of data to be sliced, diced, filtered, and transformed.[51] It is only a matter of time until other digital archive collections, including both digitized and born-digital material, begin to be exposed to algorithmic manipulation. This is already happening on some family history websites, such as Ancestry, and by scholars experimenting with mash-ups.[52]

Implications for Appraisal

Reconceptualizing the archive in this way renders simplistic approaches to appraisal obsolete, particularly as much digital data is stored haphazardly and

49 Thomas Padilla," On a Collections as Data Imperative," UC Santa Barbara, 2017, accessed 12 December 2017, https://escholarship.org/uc/item/9881c8sv.

50 "Web Archives as Scholarly Sources: Issues, Practices and Perspectives," tweets from conference at Aarhus University, Aarhus, Denmark, 8–10 June 2015, Resaw: A Research Infrastructure for the Study of Archived Web Materials, accessed 9 December 2017, http://resaw.eu/events/international-conference-aarhus-june-2015/; "AU4.0: British Invasion! – Archives Unleashed, 4.0: Web Archive Datathon," workshop description, Archives Unleashed, accessed 9 December 2017, http://archivesunleashed.com/au4-0-british-invasion/.

51 Ian Milligan, "Mining the 'Internet Graveyard': Rethinking the Historians' Toolkit," *Journal of the Canadian Historical Association* 23, no. 2 (2012): 21–64.

52 Jeremy Heil and Frank Huntley, "Stones: Using Mashups to Understand a City's Social Tapestry," The Interactive Archivist: Case Studies in Using Web 2.0 to Improve the Archival Experience, 30 June 2009, accessed 20 September 2018, http://interactivearchivist.archivists.org/case-studies/stones/.

certainly not within the familiar structure of a registry file plan.[53] It is one thing to recognize that this is the case and quite another to know what kind of structures and procedures to put in place. As the work of the recordkeeping roundtable in Australia[54] and Sir Alex Allan's report on the situation in the UK makes clear,[55] it is not enough to suggest, as many archivists do,[56] that the focus should be on creation. In this new world, the possibilities of using technology to underpin appraisal and the needs of users of digital records should be thoroughly investigated and assessed.

It is possible to envisage tools that might help with appraisal, and indeed, some such tools are being developed. For example, researchers at the Massachusetts Institute of Technology (MIT) have cleaned the Enron dataset by removing duplicates – data generated by robots and redundancies of the sort that pervade all collections of email.[57] The objective was to generate a stable test collection that could be used to conduct experiments in making sense of such a large corpus.[58] At Columbia University, similar experiments are being conducted on emails released by the State Department and the CIA.[59] Although reducing bulk is not appraisal in the strict sense, it is a step in the right direction. The National Archives of the UK has declared that appraisal of digital records and review of such records for sensitive content will only be possible if carried out

---

53   See for example F. Gerald Ham, *Selecting and Appraising Archives and Manuscripts* (Chicago: Society of American Archivists, 1993).

54   See for example Kate Cumming and Anne Picot, "Reinventing Archival Methods," *Archives and Manuscripts* 42, no. 2 (2014): 117–122.

55   Sir Alex Allan, *Review of Government Digital Records* (London: Cabinet Office, 2015), https://www.gov.uk /government/publications/government-digital-records-and-archives-review-by-sir-alex-allan.

56   Cabinet Office, *Better Information for Better Government*, 18 January 2017, accessed 20 September 2018, https:// assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/589946/2017-01 -18_-_Better_Information_for_Better_Government.pdf.

57   Bryan Klimt and Yiming Yang, "The Enron Corpus: A New Dataset for Email Classification Research," in *Proceedings of the 15th European Conference on Machine Learning: ECML 2004*, Pisa, Italy, 20–24 September 2004, ed. J.F. Boulicaut, F. Esposito, F. Giannotti, and D. Pedreschi (Berlin: Springer-Verlag, 2004), http://nyc.lti.cs.cmu.edu /yiming/Publications/klimt-ecml04.pdf.

58   William W. Cohen, "Enron Email Dataset," accessed 12 December, 2017, https://www.cs.cmu.edu/~./enron/.

59   This research is led by Matthew Connelly, professor of international and global history at Columbia University, from whom further details can be obtained; see David Allen and Matthew Connelly, "Diplomatic History after the Big Bang: Using Computational Methods to Explore the Infinite Archive," in *Explaining the History of American Foreign Relations*, 3rd ed., ed. Frank Costigliola and Michael J. Hogan (Cambridge: Cambridge University Press, 2016), 74–101.

through machine-assisted learning.[60] Ed Summers, a software developer, makes a similar point, arguing that "appraisal is increasingly enacted as part of a human-machine collaboration, where archivists use computers to help select and collect content from the web using automated processes and algorithms."[61] These are aspirations, but research will eventually generate not only new tools[62] but also new methodologies[63] and approaches in using and analyzing archives as data.

The Demand Side

It is fundamental to the design of any appraisal policy, in the digital if not in the analog arena, that it be truly responsive to demand-side concerns. Mike Featherstone invited archivists to rise to this challenge a decade ago when he wrote, "How are decisions on what to collect, what to store, what to throw away and what to catalogue to be made?"[64] The production and definition of the archive will become a collaborative, co-creative enterprise[65] – what has been described as a "curated conversation" that extends well beyond the existing customer base.[66] What is most different in this new world is that disciplines that have

---

60   The National Archives, *The Application of Technology-Assisted Review of Born-Digital Records Transfer, Inquiries and Beyond*, February 2016, www.nationalarchives.gov.uk/documents/technology-assisted-review-to-born -digital-records-transfer.pdf.

61   Ed Summers, "Archival Attention," *Medium*, 5 December 2017, accessed 12 December 2017, https://medium .com/@edsu/archival-attention-df4991aec716.

62   Ed Summers and Ricardo Punzalan, "Bots, Seeds and People: Web Archives as Infrastructure," in *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, Portland, Oregon, 25 February – 1 March 2017 (New York: ACM, 2017), 821–34, doi:10.1145/2998181.2998345.

63   Timothy Gollins, Graham McDonald, Craig Macdonald, and Iadh Ounis, "On Using Information Retrieval for the Selection and Sensitivity Review of Digital Public Records," in *Proceeding of the 1st International Workshop on Privacy-Preserving IR: When Information Retrieval Meets Privacy and Security*, Gold Coast, Australia, 11 July 2014, *CEUR Workshop Proceedings* 1225 (2014): 39, http://ceur-ws.org/Vol-1225/pir2014_submission_9.pdf; Graham McDonald, Craig Macdonald, and Iadh Ounis, "Enhancing Sensitivity Classification with Semantic Features Using Word Embeddings," in *Advances in Information Retrieval*, 39th European Conference on IR Research, Aberdeen, UK, 8–13 April 2017, ed. Joemon M. Jose, Claudia Hauff, Ismail Sengor Altıngovde, Dawei Song, Dyaa Albakour, Stuart Watt, and John Tait (n.p.: Springer, 2017): 450–63, doi:10.1007/978-3-319-56608-5_35.

64   Mike Featherstone, "Archive," *Theory, Culture & Society* 23, no. 2–3 (2006): 591–96.

65   Andrew Flinn, "An Attack on Professionalism and Scholarship? Democratising Archives and the Production of Knowledge," *Ariadne* 62 (January 2010), http://www.ariadne.ac.uk/issue62/flinn.

66   There is considerable interest in the broader concept of co-creation, which spills over into concepts of

not made extensive use of archives are beginning to access them because they are available digitally. Think for example of the way in which the availability of a great mass of born-digital records is likely to transform corpus linguistics – much as Google Books has done – as Dan Cohen and Fred Gibbs have shown in their project Reframing the Victorians.[67] Or consider how Adrian Bingham was able to search the digital archive of the British *Daily Mirror* newspaper to discover the frequency of use of the words *homosexual*, *homosexuality*, and *lesbian* before 1950.[68] Not only will archivists have to engage with users more actively but, because of the nature of digital records, they will also have to engage with user groups that have increased enormously in scale and scope. Moreover, there is evidence that historians are now beginning their research, certainly in the modern period, by consulting online newspapers and other online resources before consulting the archive.[69]

Even within the existing customer base, tastes and interests have changed and are changing with the rise of new concerns that are often not easily addressed by existing catalogues or that were deliberately excluded from them – for example, references to ethnic minorities, Indigenous peoples, homosexuals, and even women. The biggest constituency of users of archives is family historians[70] with their insatiable desire for sources with lots of names, most of which are now accessible through commercial sites – such as Ancestry and Findmypast – that shelter even user-generated content behind paywalls.[71] In a survey of UK archives in 2016, 44 percent of users were undertaking family history research, 30

heutagogy and an open-context model of learning. Fred Garnett, "The Digital Future of Education," The Heutagogy Archives, 24 May 2018, accessed 13 December 2017, https://heutagogicarchive.wordpress.com/.

67 Dan Cohen, "Searching for the Victorians," blog post, 4 October 2010, accessed 14 August 2018, http://www.dancohen.org/2010/10/04/searching-for-the-victorians/; Isto Huvila, "Participatory Archive," blog post, 19 September 2008, accessed 14 August 2018, http://www.istohuvila.se/participatory-archive.

68 Adrian Bingham, "The Digitization of Newspaper Archives: Opportunities and Challenges for Historians," *Twentieth Century British History* 21, no. 2 (2010): 225–31.

69 Ian Milligan, "Illusionary Order: Online Databases, Optical Character Recognition, and Canadian History, 1997–2010," *Canadian Historical Review* 94, no. 4 (2013): 540–569.

70 In the UK, family historians constituted the largest group of users of archives in 2016, at 44 percent. Chartered Institute of Public Finance and Accountancy, *Survey of Visitors to UK Archives* 2016, 2017, accessed 18 August 2018, http://www.archives.org.uk/images/Public_Services_Quality_Grp/Survey_of_Visitors_to_UK_Archives_-_National_Headline_Report_2016.pdf.

71 Wendy M. Duff and Catherine A. Johnson, "Where Is the List with All the Names? Information-Seeking Behavior of Genealogists," *American Archivist* 66, no. 1 (Spring/Summer 2003): 79–95.

percent were researching local history, and 21 percent were academics.[72] A 2007 analysis, by Margaret O'Neill Adams, of the use of the US National Archives and Records Administration (NARA)'s digital records describes approaches being taken by researchers:

> Historians have been a minority of the research users of
> NARA's electronic records. Instead, academics and researchers
> from the government and private sectors, including the
> media, who analyze data using various social science research
> methodologies, continue to prevail as the dominant group
> ordering copies of archival electronic digital data files.[73]

Information about people presents the archival community with formidable problems; as data protection becomes more ubiquitous as the European Union brings in the General Data Protection Regulation,[74] such data is likely be closed for up to a hundred years.[75]

It is also very difficult to predict how elastic the family history market is, as it seems to be a preoccupation of the largely middle-class and over-65 group, which has been the main beneficiary of early retirement.[76] The online family history provider Ancestry, in its 2014 *Global Family History Report*, claimed the market for family history had increased 14 times over the past decade.[77] This growth is largely confined to online resources as visits to archive repositories declined over the same period. A report for the UK National Archives in 2014 asserted, "Since 2005, visitor numbers to local authority archives have declined.

---

72   Chartered Institute of Public Finance and Accountancy, *Survey of Visitors to Archives* 2016.

73   Margaret O'Neill Adams, "Analyzing Archives and Finding Facts: Use and Users of Digital Data Records," *Archival Science* 7, no. 1 (2007): 21–36.

74   "GDPR Portal: Site Overview," accessed 7 December 2017, https://eugdpr.org/.

75   See for example David Neuberger, "Is Nothing Secret? Confidentiality, Privacy, Freedom of Information and Whistleblowing in the Internet Age," Singapore Academy of Law Annual Lecture 2015, 21 September 2015, accessed 12 December 2017, https://www.supremecourt.uk/docs/speech-150921.pdf.

76   The National Archives, *Digital Services and Archive Audiences: Local Authority Archives 2014, A Research Study*, December 2014, accessed 17 August 2018, http://www.nationalarchives.gov.uk/documents/archives/Digital _Services_and_Archive_Audiences_2014.pdf.

77   Ancestry, "Online Family History Research in United States Grows by 14 Times in Past Decade," news release, MarketWired, 19 November 2014, accessed 12 December 2017, http://www.marketwired.com/press-release /online-family-history-research-in-united-states-grows-by-14-times-in-past-decade-1970045.htm.

. . . The proportion of adults who had visited an archive (in their own time or for voluntary work) when surveyed was 3%, a significant decrease from the 2005/06 level of 6%."[78] This raises serious questions about the bulk of holdings of record offices that remain accessible only in paper format and can arguably be used to add context to the often rather bald data to be found in family trees.

Perhaps more predictable is the academic market, which is also shrinking as funding for research and doctoral programs is squeezed; at the same time, the market is becoming more fragmented as approaches to the past diversify. It is this emerging plurality, which is characterized by transdisciplinarity, that presents the archival community with the greatest challenge as it becomes harder to predict user expectations related both to what records archivists should keep and to how they should catalogue them. A very obvious issue is the demand from social scientists for access to linked personal data, which can be analyzed to inform policy.[79] This is the same data that family historians want to see retained, and it is subject to the same constraints as data protection regulations globalize, following the cloud.

The range of interest in the archives is so extensive that everything is potentially of interest, even items that archivists destroy. When the Public Record Office in the United Kingdom was set up in 1838, a committee of wise men (and they were all men) reviewed the vast accumulation of records that were scattered around Whitehall in an exercise that could be characterized as co-creation or macro-appraisal.[80] Unlike most archivists, they diligently recorded the series that were consigned to the flames. The registers can be inspected today in the National Archives at Kew at PRO 36. Looking through the registers, we can see series that, using today's criteria, would have merited retention, as well as errors. C.A.F. Meekings describes the indiscriminate nature of the destruction of Court of King's Bench files that, "despised by the Record Office," were systematically destroyed in the early 20th century and turn out to contain valuable information – for example, information on patterns of bail granted to defendants and on the composition of the court.[81] Ashburnham Bulley, the Chief Clerk to the Comptroller of the Exchequer, was sent to examine a "ship load" of old papers in

78   National Archives, *Digital Services and Archive Audiences*, 8.

79   Administrative Data Research Centre and Administrative Data Research Network, https://www.adrn.ac.uk/.

80   John D. Cantwell, *The Public Record Office*, 1838–1958 (London: HMSO, 1991).

81   C.A.F. Meekings, "King's Bench Files," in *Legal Records and the Historian*, ed. J.H. Baker (London: Royal Historical Society, 1978), 110–12.

a cellar at Somerset House. After the inspection, he was of the opinion that they were "of no real value at the present time," and almost nine tons were sold to a fishmonger to wrap his wares. It was not long before antiquaries became aware that the papers used to wrap their fish were of more than passing interest.[82] They are now mostly to be found in library manuscript collections around the world. In Trollope's novel *The Three Clerks*, there was no appraisal conducted when the department of Internal Navigation was closed: "Ruthless men shovelled off as waste paper all the lock entries . . . and the ponderous ledgers . . . were sent away to Cimmerian darkness, and probably utter destruction."[83] These would have been the very records to whet the appetites of the legion of narrow boat enthusiasts on England's canals. Inevitably, there will be such needless loss in the future, especially when departments are closed or their functions privatized, but what concerns us here is the prior knowledge of those who take appraisal decisions or design the underpinning algorithms.

The current advice from The National Archives at Kew is couched in fine words about the "long-term" archival value of digital records, but offers no definition. It is rather like a cookery book with no recipes. Nowhere is there any reference to users or to the fact that they might have at their disposal radically different tools and techniques for interrogating digitally born records. The assumption is that the Grigg method of review defined in 1958 will transfer seamlessly into the digital environment.[84]

The advice from NARA is much more objective; NARA's appraisal policy states, "Records appraisal is not a rote exercise. It requires informed judgments, knowledge of and sensitivity to researchers' interests, recognition of resource considerations, and a willingness to acknowledge and understand comments and suggestions from diverse perspectives." In clause 5 of the appraisal policy, the public is specifically mentioned as having a role in making decisions about selection for permanent preservation. The first permanent record category consists of records that document the rights of citizens.[85] As in the TNA

---

82   Andrew Prescott, "Administrative Records and the Scribal Achievement of Medieval England," *English Manuscript Studies 1100–1700*, 17 (2012): 173–99.

83   Anthony Trollope, "Chapter 45: The Fate of the Navvies," in *The Three Clerks* (1857), http://www.online-literature.com/anthony-trollope/three-clerks/45/.

84   The National Archives Appraisal Policy, (London: The National Archives, 2004), accessed 13 December, 2017, http://www.nationalarchives.gov.uk/documents/information-management/appraisal_policy.pdf.

85   National Records and Archives Administration, *Appraisal Policy of the National Archives and Records Administra-*

guidance, there is no reference to different techniques for interrogation or changing patterns of use.

The advice from Library and Archives Canada adopts a theoretical approach based in macro-appraisal and functional analysis and reflects on the "value" of the records selected for permanent preservation. It places a great deal of emphasis on making decisions based on "analysis and evaluation of government's assigned business functions" and only capturing records of offices of primary interest.[86] This is all very neat, but government no longer operates in the same way, certainly not in the United Kingdom. It is increasingly very flat; many of its functions are either outsourced or privatized, and it is astonishingly complex, particularly in areas of future interest such as counterinsurgency operations, which generate records on a gigantic scale.[87] Moreover, as public enquiries and recent reports have shown, the underlying recordkeeping systems are either non-existent or chaotic. In his report on digital recordkeeping in the UK government, Sir Alex Allan was unequivocal:

> Existing systems which require individual users to identify
> documents that should constitute official records, and then to
> save them into an EDRMS or corporate file plan, have not worked
> well. The processes have been burdensome and compliance poor.
> As a result, almost all departments have a mass of digital data
> stored on shared drives that is poorly organised and indexed.[88]

TNA has admitted that only 33 percent of government data is held in electronic document and records management systems (EDRMSs); the bulk is held on unstructured shared drives.[89] Neither NARA nor LAC's policies demonstrate recognition of the radically changed environment of digital record-making and

tion, (n.d.), accessed 12 December 2017, https://www.archives.gov/records-mgmt/publications/appraisal-policy.pdf.

86 Library and Archives Canada, *The Government Records Disposition Program of Library and Archives Canada: Program Synopsis*, 2017, accessed 17 August 2018, https://www.bac-lac.gc.ca/eng/services/government -information-resources/disposition/Pages/program-synopsis.aspx.

87 See for example United States Government Interagency Counterinsurgency Initiative, *U.S. Government Counter- insurgency Guide*, 2009, accessed 12 December 2017, http://www.state.gov/documents/organization/119629.pdf.

88 Allan, *Review of Government Digital Records*.

89 The National Archives, *The Application of Technology-Assisted Review*, 10.

recordkeeping – nor, for that matter, do appraisal policies at national institutions in Australia or New Zealand. Across the board, we are confronted by a supply-side model that does not reflect either the plurality of the customer base or the power of the analytical tools that are transforming demand.

The Emerging Macroscope

A reconfiguration of the archive as data to be mined requires an understanding of the execution of distant reading and of the emerging tools that are beginning to be employed in many organizations. There are very positive signs that new tools are being developed to facilitate the exploration of the world of digital archives we have described. The UK Foreign and Commonwealth Office (FCO) is collaborating in a project "to help ambassadors understand their physical and digital networks and discover distinct communities in contact circles."[90] The developers think that such tools could be used by the FCO to "help focus their online engagement, and they help policymakers understand social media landscape in real-time to improve engagement."[91] Social media are now deeply embedded in policy making; they are used to crowdsource lobbying and by politicians and bureaucrats both to test opinion and to implement new initiatives. A good example is the widespread use of "nudge" techniques to modify or change behaviour.[92] Whereas in the past, government kept a watchful eye on the press, now it has to be alert to the whole cacophony of noise on social media. This cannot be analyzed by close reading; there is far too much of it in almost any context. It requires distant reading to disentangle significance above the noise. This is happening across government in the United Kingdom and, as William Merrin has pointed out, it disrupts through its very utility accepted practice: creators become users; the role of intermediaries, such as libraries and archives, is colonized by commercial providers; and users become creators. What he has to say about the British media could apply equally well to archivists:

---

90 Madeleine Greenhalgh, "The Data Science in Government Programme: Progress So Far," *Policy Lab* (blog), 18 September 2014, accessed 12 December 2017, https://openpolicy.blog.gov.uk/2014/09/18/data-science-2/.

91 Ibid.

92 "Improving Workplace Performance," BusinessBalls, accessed 12 December 2017, http://www.businessballs.com/nudge-theory.htm.

> The UK Culturalist tradition that . . . became globally famous has
> become obsolete. Reduced to a caricature of its once-political self,
> committed to the performance of its own privileged methodolo-
> gies and viewing reception through the anachronistic prism of the
> "audience," its simplistic humanism uncritically valorises all individual
> behaviour, refusing to think of the user in any more complex way.[93]

As boundaries dissolve, content itself mutates, and it is no longer possible to separate manuscript from print: everything becomes "one kind of archive" feeding the "Internet of things" or perhaps more accurately, linked data.[94] The Internet of things is much more than smart technology that connects our domestic appliances to our smartphones; Samuel Greengard suggests that "smartphones, cloud computing, RFID (radio-frequency identification) technology, sensors, and miniaturization are converging to make possible a new generation of embedded and immersive technology."[95] Such a broad perspective – taking in big data and linked data – immediately raises ethical questions that should concern information professionals and that connect to the long philosophical engagement with technology that led to Martin Heidegger's articulation of thing theory, a theory that resonates with the Internet of things.[96]

It is estimated that by 2020 there will be over 50 billion things connected to the Web and generating data that can be linked and analyzed.[97] Linked data focuses attention on ontologies, which in computing science refer to "the formal representation of knowledge as a set of concepts within a formal domain." The development of domain-specific ontologies is the core of the approach that ensures entities have some chance of being linked with some degree of certainty.[98]

93   William Merrin, *Media Studies* 2.0 (London: Routledge, 2014), 188.

94   See the IOT Council website, accessed 12 December 2017, http://www.theinternetofthings.eu/.

95   Samuel Greengard, *Internet of Things* (Cambridge, MA: MIT Press, 2015).

96   Sally A. Applin and Michael D. Fischer, "Thing Theory: Connecting Humans to Location-Aware Smart Environ-
     ments," paper presented at IEEE Conference on Intelligent User Interfaces, Santa Monica, CA, 19–22 March
     2013, accessed 7 August 2018, http://posr.org/w/images/e/e7/Applin_Fischer_ThingTheoryConnecting
     HumanstoLocationAwareSmartEnvironments_LAMDa13.pdf.

97   Martin Heidegger, *Being and Time* (Oxford: Blackwell, 1978), 135–37.

98   "Ontology – Linked Data," wiki, accessed 7 August 2018, http://ontology-linked-data.wikispaces.asu.edu/Home.

It is straightforward to link data from a known device in a known location, such as a washing machine or an electricity meter; it is much more difficult to link data about people, places, and entities – the foundation of the semantic web that is thought to have potential for the cultural heritage sector.[99] This is Frege's puzzle writ large, and it can only be resolved by human intervention.[100] Archive users, from family historians to ambitious scholars, have always linked data; it is one of their stocks in trade. Data linkage is beset by false positives that can be very difficult to resolve, particularly in communities with few patronymic and first names. Linking data mechanically generates false positives on a large scale that can be seen clearly from the data linkage utilities on the Ancestry website. Another example can be seen in the Humanities Networked Infrastructure federated catalogue developed by Deb Verhoeven, where there was nothing to connect William Moss and James Moss until Michael Moss explained they were brothers.[101]

All of these examples of linked data assume, in essence, that the linkage between "things" is absolute. In the highly contested and uncertain world of archives and history, involving people and places, such an assumption is weak. It is not the number of false positive links to potential ancestors that causes a user of a family history web site confusion but, perhaps, the difficulty in assessing which is most likely. Until the more advanced forms of the resource description framework (RDF) that underlies linked data – forms capable of representing uncertainty and probability of trust[102] – become ubiquitous in archival collections, then these fundamental challenges are unlikely to be resolved.

99 V. Richard Benjamins, Jesus Contreras, Mercedes Blázquez, Juan Manuel Dodero, A. Garcia, Eva Navas, Francisca Hernandez, and Carlos Wert, "Cultural Heritage and the Semantic Web," in *The Semantic Web: Research and Applications*, Proceedings of First European Semantic Web Symposium, ESWS 2004, Heraklion, Crete, Greece, 10–12 May 2004, ed. John Francis Davies, Christoph Bussler, and Rudi Studer (Berlin: Springer-Verlag, 2004), 433–444.

100 Edward N. Zalta, "Gottlob Frege," *Stanford Encyclopedia of Philosophy*, 20 October 2016, accessed 7 August 2018, https://plato.stanford.edu/entries/frege/#FrePuz.

101 Verhoeven, "Doing the Sheep Good."

102 Nuno Lopes, Antoine Zimmermann, Aidan Hogan, Gergely Lukácsy, Axel Polleres, Umberto Straccia, and Stefan Decker, "RDF Needs Annotations," paper presented to W3C Workshop on RDF Next Steps, Stanford, CA, 26–27 June 2010, accessed 12 December 2017, https://www.w3.org/2009/12/rdf-ws/papers/ws09; Octavian Udrea, Diego Reforgiato Recupero, and V.S. Subrahmanian, "Annotated RDF," in *The Semantic Web: Research and Applications, Lecture Notes in Computer Science*, 4011 (Berlin: Springer, 2006), 487–501, doi:10.1007/11762256_36.

The Supply Chain

Although some parts of the scholarly community have for some time recognized this fracturing of what we might call the *supply chain*, they are notable exceptions; the archival community has remained largely locked in to handicraft processes that will not scale to address the sheer volume of heterogeneous digitally born "stuff." As we have seen, there is an implied assumption that business can go on as usual. This cannot be the case even with existing analog holdings, where new and potentially disruptive technologies offer new ways of doing old things. The commercial family history websites are offering utilities and tools that exploit the linking of data across their holdings, together with family tree–building software with a range of visualization options and shoeboxes, which make reading so distant that, for many, it loses contact with the close reading necessary in the archival search room.[103] Some historians are experimenting with network analysis of the sort that is being developed for the FCO, for example, the Six Degrees of Francis Bacon project at Carnegie Mellon University.[104] Others at the History Lab at Columbia are plotting cable and email traffic to pick up aberrations,[105] while other History Lab researchers are using natural language processing to determine gender, social, and power relations in corpora.[106] These techniques have the power to transform research across the humanities and social sciences, but they depend on heavy-duty computational, mathematical, and statistical skills that will scale; they also require skeptical users to trust the underlying algorithms. What is emerging is a transdisciplinary

---

103 See for example "Powerful Free Genealogy Software with Tree Syncing," My Heritage, accessed 12 December 2017, https://www.myheritage.com/family-tree-builder.

104 "Francis Bacon Network," Six Degrees of Francis Bacon, accessed 12 December 2017, http://www.sixdegreesoffrancisbacon.com/.

105 "History as Data Science," History Lab, accessed 7 August 2018, http://www.history-lab.org.

106 See for example the work of Owen Rambow at Columbia: Vinodkumar Prabhakaran and Owen Rambow, "Predicting Power Relations between Participants in Written Dialog from a Single Thread," in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics* (Volume 2: Short Papers), 2014, 339–44, accessed 12 December 2017, http://www.aclweb.org/anthology/P/P14/P14-2056.pdf; Vinodkumar Prabhakaran, Emily E. Reid, and Owen Rambow, "Gender and Power: How Gender and Gender Environment Affect Manifestations of Power," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing* (EMNLP), 2014, 1965–76, accessed 12 December 2017, http://www1.cs.columbia.edu/~vinod/papers/EMNLP_genderpaper_final.pdf; Apoorv Agarwal, Adinoyi Omuya, Jingwei Zhang, and Owen Rambow, "Enron Corporation: You're the Boss if People Get Mentioned to You," in *Proceedings of the 2014 International Conference on Social Computing*, ACM, 2014, 2, accessed 12 December 2017, http://dl.acm.org/citation.cfm?id=2640065.

approach that could only be guessed at even 10 years ago. As Cohen pleads in his blog post "Is Google Good for History?,"

> I would much rather have historians and Google to work together. While Google as a research tool challenges our traditional historical methods, historians may very well have the ability to challenge and make better what Google does. . . .
>
> Because Google Books is the product of engineers, with tremendous talent in computer science but less sense of the history of the book or the book as an object rather than bits . . .[107]

It is for humanists to critically test the results across these multiple platforms.

If we embed such demand- and supply-side reciprocity in the relationship between archivists and users, we not only reconfigure the archive but also advance into new or perhaps old territory where the user participates in the "archival" process and helps shape future discourse. Unlike in the past, when archivists could safely claim to exercise complete fiduciary responsibility (they held the keys), the digital environment implies a three-way relationship among archivists, users and the "engineers, with tremendous talent in computer science," in what we might best describe as a "recursive trans-disciplinary curatorial conversation." As Helga Nowotny explains,

> Transdisciplinarity contributes to a joint problem solving that it is more than juxtaposition; more than laying one discipline alongside another. . . . If joint problem solving is the aim, then the means must provide for an integration of perspectives in the identification, formulation and resolution of what has to become a shared problem.
>
> One virtue you need when working in transdisciplinary research: patience. . . . To understand the language of other disciplines takes time.[108]

---

107 Cohen, "Is Google Good for History?"

108 Helga Nowotny, "The Potential of Transdisciplinarity," in *Discussing Transdisciplinarity: Making Professions and the New Mode of Knowledge Production, The Nordic Reader*, ed. Halina Dunin-Woyseth and Liv Merete Nielsen (Oslo, Norway: Oslo School of Architecture, 2004), 10–19, accessed 12 December 2017, http://www.helga-nowotny .eu/downloads/helga_nowotny_b59.pdf.

There could be no better advice, as this conversation will embrace use, access, and critical appraisal. Archivists and users will be setting engineers challenging problems to resolve, experimenting with the tools that emerge, and then consulting again with the engineers – which will lead iteratively and recursively to further questions and new engineering problems to be solved.

Some elements of this interaction are germinating, as we have seen in the examples of the History Lab at Columbia and the Six Degrees of Francis Bacon project at Carnegie Mellon. At the Institute for Advanced Technology in the Humanities at the University of Virginia,[109] the Social Network and Archive Context (SNAC) project employs the affordances of the semantic web "to provide researchers with convenient, integrated access to historical collections held by multiple private and public archives and libraries around the world while also setting the stage for a cooperative program for maintaining information about the people documented in the collections."[110] The Maryland Institute for Advanced Technology in the Humanities, and the School of Information and Library Science at the University of North Carolina have developed "BitCurator Tools for Digital Forensics Methods and Workflows in Real-World Collecting Institutions."[111] At the University of Melbourne, the eScholarship Research Centre's[112] "Social and Cultural Informatics Lab is a development and testing ground for informatics solutions in archival science and digital humanities," with a "broader research agenda including the advancement of knowledge in the discovery, contextualization and interpretation of information in published literature, archival records and research data."[113] The lab has developed the "Online Heritage Resource Manager (OHRM) – a resource discovery and access system that links creators, archival and heritage resources and published materials"

109   "Mission," The Institute for Advanced Technology in the Humanities, accessed 12 December, 2017, http://www.iath.virginia.edu/mission.html.

110   "About SNAC," Social Networks and Archival Context, accessed 12 December 2017, http://snaccooperative.org/static/about/about.html.

111   "BitCurator," Maryland Institute for Technology in the Humanities, accessed 12 December 2017, http://mith.umd.edu/research/project/bitcurator/.

112   "eScholarship Research Centre," University of Melbourne, accessed 12 December 2017, http://www.esrc.unimelb.edu.au/.

113   Teresa Chitty and Donna McRostie, "Better Together: The ESRC in the University Research Library of the Twenty-First Century," *Australian Library Journal* 65, no. 3 (2016), 159.

but without explicit input from users at the ingestion stage.[114] The School of Computing Science at the University of Glasgow and the iSchool at Northumbria University are collaborating in the Abacá project, with the intention to develop technically assisted sensitivity review, which is critical for the release of digital records into the public domain.[115] The Yale University Open Data Access (YODA) Project is developing a new approach to data access and transparency for clinical trials data, which may be applicable in other contexts, using emulation services.[116]

These initiatives are still based mostly in a supply-side paradigm and are often predicated on "harnessing knowledgeable users of archival collections to contribute in the form of new and improved descriptions, translations, summaries, and relationships to other records."[117] Reciprocity is still absent. This worries many commentators who, from a post-modernist perspective, question the way in which appraisal policies and cataloguing conventions portray a particular view of the past.[118] As we have argued, tools are emerging that go a long way to address such concerns. There have been some experiments employing what is termed *participatory design*.[119] Although it is unclear if such an approach will be capable of functioning on the industrial scale required to ingest huge quantities of born-digital records, these initiatives provide useful lessons. Researchers at Newcastle employed "liveness" techniques to process and display "the archive of a publisher of contemporary poetry, Bloodaxe Books, consisting of around 60,000 items"; they concluded that "if we, as a society, are to encounter our cultural heritage in digital forms online and elsewhere it seems appropriate to engage ourselves in a conversation about not only usability but also aesthetics and creativity informed by engagement with future users."[120]

114   "Online Heritage Resource Manager," University of Melbourne, accessed 12 December 2017, http://esrc.unimelb .edu.au/about/tools-and-technology/ohrm.

115   "Tag Archives: Project Abaca," Project Abacá: Technically Assisted Sensitivity Review, accessed 12 December 2017, https://projectabaca.wordpress.com/tag/project-abaca/.

116   "YODA Project," Yale School of Medicine Center for Outcomes Research & Evaluation (CORE), accessed 12 December 2017, http://medicine.yale.edu/core/current_projects/yodap/.

117   Huvila, "Participatory Archive."

118   Rachel Hardiman, "En mal d'archive: Postmodernist Theory and Recordkeeping," *Journal of the Society of Archivists* 30, no. 1 (2009): 27–44.

119   Schofiel et al., "Archival Liveness."

120   Ibid., para. 61.

Sound and Vision

There is one final point to make. The move to the digital environment is having a profound impact on the ways history is written, historical records are accessed, and the records used by future historians are created and captured. This article and most of the sources it cites are concerned with texts because texts are the primary resources acquired and preserved in archives and are relatively susceptible to "datafication." During the course of this century, archivists have, by and large, created an online version of paper archives with better catalogues and some user input. The leading family history site, Ancestry, has shown it is possible to go much further than this by developing links between individual digitized resources and facilitating and promoting user input so that a huge amount of the data it holds is now user generated. However, we believe that the direction of technological travel reaches beyond the textual, with the greatest developments on the Web being related to sound and images. It is hardly necessary to mention YouTube, Pinterest, Flickr, Google Earth, and the rest. Archives are beginning to acquire audio-visual records that replace textual material. The record of the UK Supreme Court, as currently deposited at the UK National Archives, is an explicit example of this tendency. Since its inception in October 2009, the court has recorded and broadcast online all open proceedings at the court. Although other (presumably) textual material is yet to be deposited, these videos currently constitute the public record of the court's activities.[121] The privately generated images of warfare that we mentioned earlier are not notes written from the battlefield but video images and spoken diaries. Somehow, this mass of visual and sound media (which are equally susceptible to digitization and datafication) needs to be incorporated into the collections of archives or, at the very least, be made searchable from them; otherwise, we will continue to face the kind of broken and dispersed record we have of the Iraq War, with videos being held at the Internet Archive and texts being held in more traditional institutions. To quote Tim Hitchcock,

> The digital humanities have a genuine opportunity to create
> something exponentially more powerful than the textual human-
> ities. What the digital side of all this allows, is the removal of the

---

[121] The National Archives of the UK, UKSC – Records of The Supreme Court of the United Kingdom, UKSC 1, Supreme Court: Video Recordings of Court Proceedings, http://discovery.nationalarchives.gov.uk/details/r/C15143467.

> barriers between sound and image and text – between novel, song
> and oil painting. Each of these is no more than just another variety
> of signal – of encoding – now, in the digital, divided one from the
> other by nothing more substantial than a different file format.[122]

Substitute the word *archives* for the word *humanities*, and there you have what must surely be the future direction of travel. It is the pioneers of memory studies who have done the most to explore the complexity of what Hitchcock alludes to and who are undermining the very concept of what we imagine the archive to be. We would do well to take them seriously.

BIOGRAPHY   Michael Moss is Professor Emeritus of Archival Science at Northumbria University, Newcastle. He was previously Research Professor in Archival Studies in the Humanities Advanced Technology and Information Institute (HATII) at the University of Glasgow, where he directed the Information Management and Preservation MSc program. Prior to being appointed to HATII, he was Archivist of the university from 1974 to 2003. He was educated at the University of Oxford and trained in the Bodleian Library. He was Miegunyah Distinguished Visiting Fellow in the eScholarship Research Centre at the University of Melbourne in 2015. He was a member of the Lord Chancellor's Advisory Council on Records and Archives from 2007 to 2015 and a non-executive director of the National Records of Scotland from 2008 to 2018. He researches and writes in the fields of history and the information sciences. His recent publications include "Brussels Sprouts and Empire: Putting Down Roots," in *Gardening: Philosophy for Everyone*, ed. Dan O'Brien (Chichester, UK: Wiley-Blackwell, 2011); "Where Have All the Files Gone, Lost in Action Points Every One?" *Journal of Contemporary History* 47, no. 4 (2012); edited with Barbara

---

122 Tim Hitchcock, "Humanities²," Historyonics (blog), 20 January 2017, accessed 12 December 2017, http://historyonics.blogspot.com/2017/01/.

Endicott-Popovsky, *Is Digital Different?* (London: Facet Publishing, 2015); and "Understanding Core Business Records" in *The International Business Archives Handbook*, ed. Alison Turton (Abingdon, UK: Routledge, 2017). With Professor Laurence Brockliss at the University of Oxford, he has just completed Victorian Professions, a big research project to investigate the professions in Victorian Britain (http://www.victorianprofessions.ox.ac.uk/). A book based on the project will be published in 2019.

BIOGRAPHY   David Thomas is currently a visiting professor in the iSchool at Northumbria University, United Kingdom. Until 2013, he was Director of Technology at the UK National Archives, where he was responsible for the acquisition and preservation of digital records and websites from government departments and for the development of preservation systems. More recently, he has researched and written on the silence of archives – why archives, despite their grandiloquent claims and high status, frequently do not contain the information that people might reasonably expect. He is currently researching the impact of forgeries and fakes on memory institutions as part of a larger project on fake news and archives.

BIOGRAPHY   Tim Gollins is Head of Digital Archiving at the National Records of Scotland, where he leads the Digital Preservation program. Since starting his career in the UK technical civil service in 1987, Tim has worked in information security, information management, systems design, and development on numerous large government information projects. Tim joined the National Archives of the UK (TNA) in April 2008, and as Head of Digital Preservation, led TNA's work on digital preservation and cataloguing. He worked on the design and implementation of a new digital records infrastructure at TNA, embodying his approach to "parsimonious preservation." Tim recently completed a secondment in the School of Computing Science at the University of Glasgow, where he investigated the challenges of digital sensitivity review. Tim holds a BSc in chemistry (Exeter), an MSc in computing science (University College London), and an MSc in information management (Sheffield). Tim was a director of the Digital Preservation Coalition for six years, from 2009, and is a former member of the University of Sheffield iSchool's advisory panel.