frontiers in Digital Humanities

HYPOTHESIS AND THEORY
published: 30 August 2018
doi: 10.3389/fdigh.2018.00020

Artificial Fibers—The Implications of the Digital for Archival Access

Michael Moss[1]*, *David Thomas*[1] and *Timothy Gollins*[2]

[1] i-School, Northumbria University, Newcastle upon Tyne, United Kingdom, [2] University of Glasgow, Glasgow, United Kingdom

This article explores how current methods and approaches in archives are under serious challenge because of the changes brought about by the move to the digital. The availability of digital records has meant that new needs and new possibilities have opened up for users, including new ways of reading. The nature of archives themselves are changing—they are moving from being collections of individual texts to be pored over to data to be made sense of. New tools and techniques have emerged and are available now which offer radical new possibilities for research, but these bring new challenges about trust and the sheer volume of records to be handled. The traditional approaches of applying metadata to facilitate the finding of relevant material and of regarding digital documents as something like electronic paper is no long viable. What is needed is a new approach in which archivists and scholarly researchers see archives as collections of data which are capable of analysis by a range of sophisticated tools and which are capable of being interpreted in a range of different ways.

Keywords: email, preservation, access records, archive, metadata, sense-making, appraisal

OPEN ACCESS

Edited by:
Richard Deswarte,
University of East Anglia,
United Kingdom

Reviewed by:
Ian Milligan,
University of Waterloo, Canada
James Baker,
University of Sussex, United Kingdom
R. C. E. Teszelszky,
National Library of the Netherlands,
Netherlands

***Correspondence:**
Michael Moss
michael.moss@northumbria.ac.uk

Specialty section:
This article was submitted to Digital History, a section of the journal Frontiers in Digital Humanities

Received: *08 September 2017*
Accepted: *06 August 2018*
Published: *30 August 2018*

Citation:
Moss M, Thomas D and Gollins T (2018) Artificial Fibers—The Implications of the Digital for Archival Access. Front. Digit. Humanit. 5:20. doi: 10.3389/fdigh.2018.00020

THE CHALLENGE OF THE NEW

Archival practice remains locked in handicraft processes. From only a glance at the random collection of thousands of ill-assorted emails, to be found in Wiki-Leaks, it is clear that access to born digital content cannot continue to be provided through conventional catalogs. Indeed, in addressing their own collections of what the United Kingdom Foreign and Commonwealth Office (FCO) still call nightly telegrams, the FCO is experimenting with new sense-making tools to assist embassies interpret the streams of data they receive (Greenhalgh, 2014). These ideas and applications are successors to the initiatives to understand operational military communication messages instigated at the end of the twentieth century by the US (Grishman and Sundheim, 1996). Together with the even newer modes of digital communication such as interactive text chat (for example WhatsApp) and the increasing uses of both public and semi-private social media platforms (for example Whitehall departments' use of Twitter and the Scottish Government's tentative internal use of Yammer) means that records that the archive is already confronting are huge accumulations of "stuff."

It is wishful thinking to imagine that order can be imposed on all but a fraction of content even at the time of creation. Even for "conventional" digital documents, we know, at least in the UK civil service, registries and file plans have all but vanished (Allan, 2014, 2015). The Enron emails that were made available during the legal investigations into the business consisted of 620,000 assorted emails and the only way that sense could be made of them was by using advanced computational and statistical techniques at the Language Technology Institute at Carnegie Mellon University (Klimt and Yang, 2004b) and in the Department of Computing Science at Columbia University (Prabhakaran et al., 2014).

Such techniques are now becoming common in the digital forensics community (in commercial tools such as Nuix)[1] Their potential for application to the archival material would arguably only be a sophisticated extension in the digital world to the fundamental concept of cataloging (TNA, 2016). Cataloging has always been about *post-hoc* sense-making, even if at times it was open to accusations of "haphazard historical gerrymandering" (Lynch, 2003, 196).

THE CHANGING NEEDS OF ARCHIVE USERS

The archival community has been slow to recognize the challenge of accessing digital content, perhaps thinking naively that the randomness of current search engines will do. Much of the community is failing to appreciate that users will not only need, but be able, to deploy sophisticated tools and services that allow digital content to be interpreted in radically different ways as at the FCO; or even historically in the network of Francis Bacon's relationships, another Carnegie Mellon University project (Rea, 2015). The archival community needs to abandon a pre-occupation with cumbersome metadata in handling digital objects and engage with communities of statisticians, mathematicians and computational scientists. Such teams at Carnegie Mellon and Columbia, are already developing new sense-making tools and services that do not impose unnecessary burdens on content creators.

Such a trans-disciplinary approach must be predicated on continuous interaction between the supply and demand sides; between, on the one hand, the professional concerns and practical issues facing archivists trying to preserve digital records and, on the other hand, the changing needs of researchers. For too long archivists have concentrated on the supply side and neglected the demand side. This must change. This focus on supply is well articulated in an article by Clifford Lynch in which he said: "we should avoid over-emphasizing **pre-conceived notions** [our emphasis] about user communities when creating digital collection, at least in part because we are so bad at identifying or predicting these target communities" (Lynch, 2003, p. 196). Even in the now maturing field of digital preservation the oft cited "OAIS" reference model (Lavoie, 2014) contains the concept of "Designated Community" to provide the basis of a justification (our term) for the costly actions taken in preserving a collection. Such preconceived notions asserting the value and utility of a collection are intrinsically problematic and must be deployed with great care in our view.

More recently, Tom Schofield and others did some work 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 catalog, and thus would not be represented in future interfaces to the archive" (Schofield et al., 2015). As long ago as 1987 Bruce Dearstyne, writing in the *American Archivist* (Dearstyne, 1987) voiced very similar criticism. He quoted Roy Turnbaugh, (Turnbaugh, 1983, p. 451) 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."

We believe, like Lynch and Dearstyne, that an emphasis on "pre-conceived notions" of potential user communities is indeed futile. But most importantly this does not mean an intellectual abandonment of the demand side by archivists. Quite the contrary, the emergence in a number of other information science domains of techniques that can "make sense" "on demand" of undifferentiated collections of information, provides an opportunity for a transformation in archival practice. In Switzerland, for example, Basma Makhlouf Shabou and Maria Sokhn (Shabou and Sokhn, 2017, p. 219) are working to "valorise cultural heritage through a citizen centric design platform" named City-Zen. The idea is to provide the tourist with utilities to make sense of data available in a variety of formats and locations.

Such emerging tools remove the need to emphasize and impose a single structure that, by its very nature, must be biased (Pitti, 2006). These tools can enable the users each to determine their own view of the archive or collection of data on their own terms. Thus, an understanding of this new approach to the demand side, becomes crucial to the understanding of what the archive needs to contain in order to support this emerging community of new users and their demands. These new demands can in turn reinforce a new view of supply, and serendipitously the very same tools the users wish to apply to collections will be required by archivists to understand the new landscape of digital records they find on the supply side.

By switching their epistemological perspective, these new and imaginative users of archives and related resources are beginning to say that by using such and such techniques we can do exciting things with your data. They could for example visualize catalog entries using NGram, or take a detailed list of correspondence and visualize the links, or as in the Swiss example cited above explore the cultural heritage of a city. There has been a good deal of research into visualizing documents and the complex data generated by statistical and mathematical analysis of content (Ahlberg and Shneiderman, 1994). All you need to do is to explore the options for viewing your photograph gallery on your i-Phone. As Kalpesh Padia points out: "Many web archives such as Archive-It (Archive-It)[2], California Digital Library (California Digital Library)[3], Library of Congress (Library of Congress)[4] and Pandora - Australia's Web Archive (Pandora)[5] provide a textual interface

[1] Nuix, http://www.nuix.com/

[2] Archive-It, https://archive-it.org

[3] California Digital Library, http://www.cdlib.org

[4] Library of Congress Archived Web Sites, http://lcweb2.loc.gov/diglib/lcwa/html/lcwa-home.html

[5] Pandora, http://pandora.nla.gov.au

for interacting with the archived collections" (Padia, 2012, p. 17). Further, an MIT project developed a system, "Themail" to visualize the contents of email boxes. The data that Themail visualizes consist of processed email mailbox files. Themail begins with an email archive in the form of one or more mbox files, which are then processed by applying a keyword-scoring algorithm. This application outputs a datafile that can be read by the visualization (Viégas et al., 2006; Themail)[6] These developments will inevitably result in a reconfiguration of practice and will have ramifications for conventional ways of doing things as new tools and services become commercially available.

Archivists have failed to recognize the impact that the move to the digital has had on users because they have focussed more on the supply side than on the demand side. However, researchers, often outside the archival mainstream, are developing new tools to enable individualized searching of archival resources. One significant aspect of the move to the digital is that it has led to a new mode for reading archives which we now describe.

THE CHANGING MODES OF READING ARCHIVES

Until recently, history has largely been text based and relied on books and on documents supplied by archives. Some writers, notably Tim Hitchcock, (Hitchcock, 2015) have suggested that the digital allows history to be broadened to encompass other sources including sound, video, even the haptic. Indeed, some historians have relied on oral history or indigenous traditions. In this article, we are focussing on the textual which, we believe, still forms the bedrock of present day historical research.

In 1938, Cleanth Brooks and Robert Penn Warren published their seminal *Understanding Poetry* (Brooks and Warren, 1938) which shaped the analysis of literary works for the next 60 or 70 years. During that period, the predominant method for the analysis of literary works has been what is called "close reading," a detailed study of texts which focusses on the work of art as an autonomous object that can be analyzed on its own terms (Davis, 2011). This "close reading" approach has become generalized from literature studies to humanities more broadly. In 2000, Franco Moretti, the Italian literary scholar, 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 rather than studying texts, literary history could be searched for themes (Moretti, 2000). Since then, Moretti (2013) has become more concerned with the digital and his views have been more generalized to the humanities.

The historians and digital humanities scholars William Turkel, Kevin Kee, and Spencer Roberts (Turkel et al., 2013, p. 62) have argued that close reading of texts is impossible in the digital age. They quote Cohen (2011) who pointed out that, while a single historian might have been able to read the 40,000 memos issued at the White House by the Johnson administration, they certainly could not handle the four million emails sent out while Clinton was in office. In future, users will need to rely on sophisticated analytical tools to allow distant reading of a large volume of material.

Such tools should not be confused with an obsession with "search" which only enables the finding of the thing expected by the researcher. In the context of the internet, users' pre-conceptions, no matter how extreme or abstruse, can always be justified and reinforced somewhere in the enormity of that gargantuan resource–any larger counter evidence results are simply not seen. We must not allow such a mode of interaction to be the only one we offer to our collections. We should, of course, support search, but by providing other analytical tools we can encourage the taking of a wider view, and the discovery of patterns and understandings currently hidden in plain sight. This is exactly what the aforementioned City-Zen aims to do.

David Weinberger in *Too Big to Know* characterizes this as "long-form reading" which according to some commentators "enables and encourages long-form thought" (Weinberger, 2011, p. 99). He suggests that neither long-form nor close reading were: "a good match to the structure of the world. Perhaps intertwining networks reflect the world more accurately" (Weinberger, 2011, p. 115). He identifies five properties of the networked world we encounter through our browsers: "abundance, links, permission-free, public, and unresolved" which taken together negate traditional practice (Weinberger, 2011, p. 174). These require users to rely on sense-making and delivery tools that enable them to work in new ways. Tim Hitchcock in his Historyonics blog talks of the historian's need to develop a metaphorical macroscope—a device that makes it possible to see both large objects and small ones at the same time (Hitchcock, 2014). He cites Jo Guldi and David Armitage's *History Manifesto* (Guldi and Armitage, 2014), which argues 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' (Hitchcock, 2014). The authors of *Exploring Big Historical Data: The Historian's Macroscope*, Shawn Graham, Ian Milligan and Scott Weingart, further argue:

"We are not implying that this is the way historians will "do" history when it comes to big data; rather, it is but one piece of the toolkit, one more way of dealing with "big" amounts of data that historians are now having to grapple with. What is more, a "macroscope," a tool for looking at the very big, deliberately suggests a scientist's workbench, where the investigator moves between different tools for exploring different scales, keeping notes in a lab notebook" (Graham et al., 2015, p. xvi).

The move to the digital has seen a change in the way in which archives are read. Because of the vast scale of the resources available, researchers are now moving from close reading of documents to distant reading, from a microscope to a macroscope. At the same time, the whole nature of the archive itself is changing.

[6]Themail, http://alumni.media.mit.edu/~fviegas/projects/themail/study/index.htm

THE CHANGING NATURE OF THE ARCHIVE

These processes are changing the nature of the archive, it is coming to be reconceptualised as data to be made sense of. This is not just the official record, but the whole mass of accompanying social media and user input (Merrin, 2014, p. 152). It was certainly the case that the archive as a whole was always difficult to grasp. Indeed access to associated records in the paper world was not as seamless as the internet has made the digital world. For example, news reports could be found in newspapers, private papers in library manuscript collections, archives and family papers in business houses, and so on. Moreover, the archive is no longer static. Not only are bit patterns inherently difficult to authenticate (Allison et al., 2010), the archive is constantly being added to by user comments, re-cataloged and copied, and made available in various locations, which are usually public and "unresolved." An example of this public input and engagement is the National Archives and Records Administration recently launched: "History Hub - A support community for history enthusiasts, researchers, citizen archivists, family historians, archival professionals and open government advocates" (History Hub, 2015). Michelle Caswell makes a similar argument in a recent polemic: "archivists should invite users, as well as outsiders to the archival process, to participate in archival description using language, categories, systems, and standards that are meaningful to them" (Caswell, 2016).

The US State Department has a large archive of telegrams, but copies of this archive are also held by Wikileaks and by others who have downloaded and analyzed it. In so doing, long cherished shibboleths of archivists, such as hierarchies, original order and provenance, are rendered impossible to establish with any authority. Very little born digital textual material has entered the public domain by due legal process anywhere in the world yet, with the notable exception of legislation in the UK (http://www.legislation.gov.uk/) with judgements of the UK Supreme Court (https://www.supremecourt.uk/news/latest-judgments.html) and lesser courts (https://www.judiciary.gov.uk/judgments/) now available publically online. Further, as far as we are aware, the Enron corpus remains the only large-scale collection of authenticated emails available to researchers. There are a large number of leaked corpora on WikiLeaks, but these have not been authenticated in the same way as Enron (Klimt and Yang, 2004a,b). Therefore, it is very difficult for the archival community to gauge just how dramatic the reformation in practice will be, or what the user needs will be, when they deploy new analytical tools. However, anyone who regularly uses large collections of digitized material, such as Ancestry.com, Trove in Australia, British Newspapers online, or Google Books, will recognize almost instinctively what the aforementioned David Weinberger is getting at in proposing entwined networks. Family history websites, such as Ancestry.com, already offer users utilities that link data to be found in the various collections to which they hold rights, such as the census, registers of births, marriages and deaths, and newspapers.

Not only is the archive being read in new ways, and its very nature changing from texts to data to be analyzed, but as we show in the next section, the move to the digital has made possible a whole range of new non-textual ways of experiencing archives.

NEW MODES OF EXPERIENCING THE ARCHIVE

A recent report by Ian Chowcat to JISC (formerly the Joint Information System Committee) that supports UK post-16 and higher education indicates that the generation of young people who will enter universities from 2020 will be used to interfaces based on touch or gesture. They will see online and offline experiences as seamlessly blended and accordingly seem to have high visual preferences (Chowcat, 2015, p. 5). This gives a few clues as to the way forward in thinking about interface design and service delivery.

However, it is important not to see this as something which is **going to** happen; it is already happening **now**. If you have not, go to the Virtual St. Paul's Cross website, which makes it possible to *experience* John Donne's "Gunpowder Day" sermon of 1622 (Virtual St. Paul's Cathedral Project)[7] The site combines the text of the sermon with the use of architectural modeling software and acoustic simulation software, so you are there on that cold November day in the reign of James I when Donne gave his sermon. This is not an isolated example. There are many other examples that reflect the way in which the digital confuses temporality. However, as Dave Nicholas has shown this is how contemporary users work. They jump from one thing to another in real time (Nicholas, 2007, p. 125). They may watch John Donne for some time, but then hop to Tim Hitchcock who is working to recreate a sound scape of the courtroom at the Old Bailey. This sound scape re-creates the aural experience of the defendant–what it felt like to speak to power, and what it felt like to have power spoken at you from the bench (Old Bailey Voices)[8] The Virtual St. Paul's Cross project and Hitchcock's sound scape work take us behind the textuality of the archive toward something approaching the original experience of the audience for Donne's sermon or the judge, jurors, lawyers and defendants in a courtroom. As Holger Schott Syme has argued the witness statements which survive in the form of written depositions and which are a major feature of the textual records of the courts were read out in court by a clerk (Schott Syme, 2003, p. 109). Importantly, as he has emphasized, it was the reading aloud which constituted evidence and not the written text. To this we might add collections of sermons delivered from pulpits. The user will only become fully aware of the purpose of these sites if they can be encouraged to extend "dwell time" on them. This may be, following the Citi-Zen model, by walking between the Old Bailey and St Paul's for example.

The move to the digital which has changed the nature of the archive and allowed for new ways of reading has also enabled it to be experienced visually and acoustically, and, if Tim Hitchcock is

[7] Virtual St. Paul's Cathedral Project, https://vpcp.chass.ncsu.edu/
[8] Old Bailey Voices, https://oldbaileyvoices.org/

to be believed, haptically. Such radical changes open up new and exciting possibilities for research.

NEW POSSIBILITIES FOR RESEARCH

These new technologies make it possible to undertake some remarkable research. Let us take the example of the work at Columbia on the Enron emails. Owen Rambow, one of the investigators, wrote to the authors:

"We see pervasive differences between language use by people in power and people without power, which allows us to predict who has power in a dialog. We have asked how this power-related behavior changes when we incorporate the gender of the discourse participants in the analysis. We have found profound differences in language use between men and women in power, and also in female and non-female gender environments (the gender environment reflects the gender of all discourse participants). We are investigating how the social networks that pre-exist a particular dialog relate to power relations." (Rambow, personal communication).

This is precisely the sort of techniques that Guldi and Armitage are arguing for in the *History Manifesto* (Guldi and Armitage, 2014). Although there are obstacles to accessing and manipulating born digital content, particularly ethical sensitivities, especially data protection, and copyright, these should not stand in the way of experimentation with the growing body of content open to digital exploration that is entering the public domain. The Digital Panopticon project is looking at what visualization techniques can reveal about the overall shape and distinctive patterns in the data, and what does this reveal about the various processes by which the data were created, and their constraints and limitations (Digital Panopticon[9]). As with all technical developments, some techniques will have little utility, but others will allow novel interpretation, for example the visualization of contributors to Wikipedia that confirmed the existence of Wikipedians (Zachte, 2011). However, this will only happen if there is dialogue across the disciplines and in the archive between the supply and demand side.

So far, we have presented a rosy picture of a digital revolution which has changed the nature of the archive and which encourages new ways of reading, new ways of visualizing archives and facilitates new methods of research. However, there are two issues to be considered—trust and volume.

TRUST ISSUES

One of the big challenges facing developers of these new approaches to information is the need to secure the trust of users. Traditionally, users have had to rely on the skills, honesty and breadth of vision of archival cataloguers to provide them with reassurance that all the material of potential interest to them has been described, but as we have shown above, such reliance has sometimes been misplaced. Now there is another trust issue—users of online systems, whether common search tools or much more specific research applications have to trust the technology. The user is, in a very real sense, swapping trust in cataloguers for trust in technologists. Moreover, the trustworthiness of some search engines and social media sites has been questioned. For example, in 2015, researchers from the Harvard Business School, the Columbia Law School and Yelp argued that: "By prominently displaying Google content in response to search queries, Google is able to use its dominance in search to gain customers for this content" (Luca et al., 2015, p. 1).

VOLUME ISSUES

The fundamental issue facing us all is that of volume. Whether records managers capture everything or engage in a selection process, the inevitable consequence of the digital will be that we acquire many more records, either to meet demand or because trying to disentangle email corpora and other digital datasets is just too difficult. Klimt and Yang only managed to reduce the Enron emails by two thirds after they had cleaned the data set, from about 600,000 emails to 200,000 (Klimt and Yang, 2004a, p. 1). This is an order of magnitude less effective than current appraisal practices. This is because of a rapid reduction in the cost of data creation and storage paralleled by an explosion in internet users from 2 billion in 2010 to 4 billion by 2017 (Internet World Stats, 2017). In the United Kingdom, the Army Historical Branch since 2002 has received 10 million "declared" records and 60 million "undeclared" (Evans, 2015). Viktor Mayer-Schönberger tells the story of his father who, as a teenager, was given a Kodak Brownie camera in the 1930s but was warned that photographs were expensive and should only be used for special occasions. As a result, over the next few years he only took about three dozen photographs of important family occasions and the mountains he climbed (Mayer-Schönberger, 2011, p. 45). Now, cameras come bundled into phones, there is no cost of processing and printing with increasingly capacious portable storage devices. The latest SanDisk memory card for smartphones will store 36,000 photographs—enough even for a tourist with a selfie stick. In addition, free storage is available to consumers on a range of sites—Dropbox, Box, Google Cloud, and so on. Google began to offer a free email account with one gigabyte of capacity on 1 April 2004 with the goal of "free storage so you'll never need to delete another message." More recently, they have advertised their Google Pixel 2 phone with "Unlimited Storage," despite the challenging counter publicity (Smith, 2017). Most new PCs come with a terabyte of storage, something unimaginable only a decade ago. How long will it be before even an exabyte becomes the standard?

Businesses and governments have access to increasingly cheap in-house storage and to high quality Cloud-based solutions. According to Kryder's law (the storage equivalent of Moore's law for computer processing power) the capacity and cost of hard disk space is halved every 18–24 months (Walter, 2005, pp. 32–33). However, this is far from being accepted universally, the recent work of one of the most influential figures in the

[9](The) Digital Panopticon-The Global Impact of London Punishments. 1780-1925, Available online at: http://www.digitalpanopticon.org/

field of digital preservation, David Rosenthal, suggests that such "laws" may not be universal and that expansion of storage space may reach physical limits in the foreseeable future (Rosenthal, 2017).

Further, much of the material that is stored consists of ephemera and of duplicates—copies of emails, photographs and other documents that are automatically stored in multiple copies across numerous portable devices, desktops, servers and back-up systems. While the cost of storage has been falling, the same cannot be said for the costs of records management. Consequently, the cost of "forgetting" data by selective deletion requires more effort and is more costly than having it preserved. This emphatically tilts the default toward preservation or perhaps more accurately "keeping stuff." Moreover, deleting digital records is hard. Copies of emails are stored on the sender's computer, the recipient's computer or a central server and may also be stored in a buffer or temporary store. Pressing the delete key does not necessarily delete them in all their stored or "preserved" copies.

Serious problems arise when humans need to understand the meaning of information in bulk, where the amount of data that provides this information is too much for a person to comprehend or too great for available human resources to read and analyse. One specific issue in this regard is that since a large dataset contains many millions of words, a full text search is likely to yield, at least, a few examples of whatever one wants to search for. Thus, the initial hypothesis is always reinforced. This is a problem that Ted Underwood has defined as "confirmation bias" (Underwood, 2014, p. 66). Underwood gives the example of a researcher who has a hypothesis that the word blushes are symbols of moral consciousness in nineteenth-century poetry. The researcher can go to a database of primary sources and search for poems that contain both "blush" and "conscious." If this is successful, then an article can be written. If not, then alternative word associations can be searched—"blush" and "shame" for example. Given a large enough database, such links will be found. Underwood goes on to state:

> 'It's true that full-text search can confirm almost any thesis you bring to it, but that may not be its most dangerous feature. The deeper problem is that sorting sources in order of relevance to your query also tends to filter out all the alternative theses you didn't bring. Search is a form of data mining, but a strangely focused form that only shows you what you already know to expect.' (Underwood, 2014, p. 66).

Underwood goes on to suggest two alternative approaches. One is to use an algorithm which looks for the words which are most commonly associated with "blush." It turns out that the word is "artless," which undermines the idea that blushes are something to do with moral conscience. The other is to use topic modeling—to allow the computer to organize the language of a collection into clusters of terms that tend to occur in the same contexts (Underwood, 2014, pp. 67–69). This is capable of revealing discursive patterns that the researcher did not necessarily look for. Crucially, all these approaches from simple word searches to topic modeling are not the unconscious product of a black box machine, they all originate with a human hypothesis.

While huge new possibilities are opened up by the digital, there are very real difficulties—can we trust the technology used to search such material and, how can we cope with the gargantuan volumes we are facing. One traditional approach by archivists is to rely on metadata. The validity of this approach is discussed in the next section.

METADATA OR NOT?

Conventional wisdom dictates that to achieve superior search results more and better "semantic metadata" is needed (Duff and van Ballegooie, 2006). However, even in the analog world, metadata is not always the answer. For example, when David Thomas, one of the authors, was responsible for the press release by The National Archives (UK) of files concerning the 1957 Windscale fire, he naturally arranged for copies to be made available of what the catalog description told him was the most useful file. At the end of the press event, a journalist told him that he had got it quite wrong and that the most interesting file was a more obscurely titled one in the collection of the UK Atomic Energy Authority (UKAEA) Northern Production Group, which are generically described as: "Records of the UK Atomic Energy Authority's production divisions touching all aspects of atomic energy research, day to day procedures, administrative functions and industrial relations" (TNA, 2018).

The problem with relying on metadata in the digital world is, apart from ambient automatically generated metadata, how do you create it? The current solution seems to be that users are responsible for devising and applying additional metadata, such as providing new, more "appropriate" titles to emails and giving useful file titles to collections of documents within electronics records document management systems (EDRMS). However, as Michael Moss discovered, "projects which have investigated this have shown that users will only be prepared to adopt such conventions if they can see clear value added to themselves in terms of their business processes which includes compliance with regulations." (Moss, 2005, p. 589). It is simply unreasonable to expect senior managers to act like filing clerks. This has been confirmed recently in the two investigations of UK government record keeping by Allan (2014, 2015). The reports argue:

"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 organized and indexed" (Allan, 2015, p. 1).

Moreover, in the United Kingdom The National Archives has admitted that only 33 percent of data is held in EDRMS, the bulk is held unstructured on shared drives (TNA, 2016, p. 10).

Finally, to illustrate the points we have made about the changing nature of the archive and the issues this poses, we turn to emails which, for most of us are the living embodiment of both the potential of digital records and their challenges.

EMAILS

Without appropriate treatment, emails are invisible as records, but also, as anyone who has looked for the record of a decision in an email chain knows, an individual email has little value on its own. It can only be understood as part of a long and complex string of correspondence with numerous false trails. Threads of many hundreds of email messages to scores of people may contain both context and content that can only be derived by visualizing the entire thread as one big document and extracting context and content across hundreds of cryptic email messages that individually have almost no understandable content or even context.

The philosophy of regarding a small proportion of e-mails as being "substantive" and worthy of becoming "official records" must be questioned. The approach taken by digital forensics teams to e-mail should indicate to archivists where the real evidence of transactions lies (Waugh, 2014). In the forensics world, e-mail is captured as whole collections with no filtering again emphasizing the importance of the supply side. Only once the target of an investigation is determined are filters applied to limit the data to be examined to those that are relevant (demand side). E-mail is naturally rich in what archivists would traditionally describe as ambient metadata, such as addressee, time and date of sending and so on. This, and other considerations have led the National Archives (NARA) in the US to develop CAPSTONE, which came into force at the end of December 2016 in accordance with the Managing Government Records Directive (M-12-18) (Email Management, 2016). This approach suggests that US government agencies should only capture records for permanent preservation from the email accounts of officials at or near the top of an agency or an organizational subcomponent. An agency may also designate email accounts of additional employees as CAPSTONE when they are in positions that are likely to create or receive permanent email records (National Archives, 2013).

Most US federal agencies met the 2016 deadline, but "told NARA it is unclear how they'll measure their success and know that they are compliant with federal records management requirements" (Ogrysko, 2016). The authors do not believe that the approach of only capturing the e-mail of the upper echelons of an organization is sufficient as a record keeping or archival approach. However, we do think that the implicit acknowledgment that archival value lies in whole collections of e-mail rather than in users' arbitrary selection is a very significant step.

Thus, the development of emails as the main business communication tool should lead us to consider whether we are not misunderstanding the nature of digital records. Much of the literature (and there is a lot of it) seems to be based on the assumption that "significant" emails can and should be made to have the same characteristics as paper letters. In other words they are supposedly self-contained, their context is simply evident from their content (or if not it requires a human to provide it in the form of additional semantic metadata), and they can be neatly "filed" with documents in a pre-coordinate structure that has some sort of principle significance for the organization. Thus, they are capable of being accessed through conventional catalogs. Yet, the evidence we have cited suggests that electronic data is not like this at all and there is no prospect it ever will be.

Although we have chosen to illustrate our point with e-mail, e-mail is not a unique case. The advent of Google Docs, SharePoint, intranets, and re-tweeted tweets has created a mesh of further digital data (masquerading as documents) at the heart of most modern organizations. Data today are not homogenous, but heterogeneous with a variety of embedded objects and linkages all of which together constitute a combined record that merits preservation. For 30 years or more, organizations (and indeed individuals) have become more and more co-dependent on this data. These individual types of data (e-mail, Tweets, Google Docs), and the challenges for organizations and archives that they bring, are merely a symptom of the wider condition of "datifiation."

CONCLUSION

Like all medical conditions we are familiar with, treatments that deal only with the symptoms will never provide a cure. We can develop the "e-mail pill" and the "SharePoint pill" but the underlying damage to the body of the archive will continue unless we change our approach. Archivists are in a unique position to observe the progress of the disease and develop a systemic response. Archivists and their users, particularly the scholarly community, need to move to a situation where they see archival collections as online collections of data which have totally different characteristics from traditional analog collections. Their interpretation must be fluid and susceptible to analysis by a new and expanding range of sophisticated tools. The old nostrums of metadata, original order and even original records have lost their power.

AUTHOR'S NOTE

David Thomas was formerly Director of Technology at The National Archives in London and is now a visiting professor at Northumbria University, Michael Moss is Emeritus Professor of Archival Science at Northumbria University and Tim Gollins is Head of Preservation & Information Management, National Records of Scotland. From the perspectives of their differing experience of engaging with technology as users, archivists and technologists, they shared in the development of this paper.

AUTHOR CONTRIBUTIONS

All authors listed have made a substantial, direct and intellectual contribution to the work, and approved it for publication.

ACKNOWLEDGMENTS

A shortened form of this paper was presented at the Activation and impact: the societal role of records and record-keepers (FARMER) Conference 2016 conference of the Forum on Archives and Records Management Education and Research at Dundee University in April 2016. We would like to thank the organizers for inviting them to speak and to Willaim Vinh-doyle and Daniel German for their comments.

REFERENCES

Ahlberg, C., and Shneiderman, B. (1994). "Visual information seeking: tight coupling of dynamic query filters with starfield displays," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, eds B. Adelson, S. Dumais, and J. Olson (Boston, MA), 313–317.

Allan, A. (2014). *Records Review*. London: Cabinet Office and The National Archives. Available online at: https://www.gov.uk/government/publications/records-review-by-sir-alex-allan

Allan, A. (2015). *Government Digital Records and Archives Review*. London: Cabinet Office and The National Archives. Available online at: https://www.gov.uk/government/publications/government-digital-records-and-archives-review-by-sir-alex-allan

Allison, A., Currall, J., Moss, M., and Stuart, S. (2010). Digital identity matters. *JASIST* 56, 364–372. doi: 10.1002/asi.20112

Brooks, C., and Warren, R. P. (1938). *Understanding Poetry: An Anthology for College Students*. New York, NY: Henry Holt and Company.

Caswell, M. L. (2016). 'The Archive' is not an archives: on acknowledging the intellectual contributions of archival studies. *Reconstruction : Studies in Contemporary Culture* 16.

Chowcat, I. (2015). *Spotlight on the Digital - Spotlight on the Digital Recent trends and Research in Scholarly Discovery Behaviour*. Bristol: JISC. Available online at: https://digitisation.jiscinvolve.org/wp/files/2015/10/spotlight_literature_review_sept2015.pdf

Cohen, D. (2011). *Defining Digital Humanities', Research Without Borders Conference, Columbia University*. Available online at: http://scholcomm.columbia.edu/2011/02/10/defining-the-digital-humanities/

Davis, G. (2011). *The Well-Wrought Textbook, A look back at Brooks and Warren's college classic, Understanding Poetry, Humanities 32*. Available online at: https://www.neh.gov/humanities/2011/julyaugust/feature/the-well-wrought-textbook

Dearstyne, and Bruce, W. (1987). What is the use of archives? a challenge for the profession. *Am. Arch.* 50, 78–87. doi: 10.17723/aarc.50.1.572q383767657258

Duff, W., and van Ballegooie, M. (2006). *Archival Metadata*. Toronto, ON: University of Toronto. Available online at: http://www.dcc.ac.uk/resources/curation-reference-manual/completed-chapters/archival-metadata

Email Management, (2016). *National Archives*. Available online at: https://www.archives.gov/records-mgmt/email-mgmt

Evans, R. (2015). *Archives of War: Media, Memory and History Conference*. TNA, London. 30 November.

Graham, S., Milligan, I., and Weingart, S. (2015). *Exploring Big Historical Data: The Historian's Macroscope*. London: Imperial College Press.

Greenhalgh, M. (2014). *The Data Science in Government Programme: Progress So Far, Examples and Findings*. Policy Lab. Available online at: https://openpolicy.blog.gov.uk/2014/09/18/data-science-2/

Grishman, R., and Sundheim, B. (1996). "Message understanding conference - 6: A Brief History," in *COLING '96 Proceedings of the 16th International Conference on Computational Linguistics*, *Vol. 1*, 466–471. Available online at: http://www.aclweb.org/anthology/C96-1079

Guldi, J., and Armitage, D. (2014). *History Manifesto*. Cambridge: Cambridge University Press.

History Hub (2015). *NARA* Available online at: https://historyhub.history.gov/docs/DOC-1012

Hitchcock, T. (2014). *Big Data, Small Data and Meaning*. Historyonics. Available online at: http://historyonics.blogspot.com/search?updated-max=2015-05-29T02:11:00-07:00andmax-results=7andstart=5andby-date=false

Hitchcock, T. (2015). *The UK Web Archive, Born Digital Sources and Rethinking the Future of Research*. Historyonics. Available online at: http://historyonics.blogspot.com/2015/

Internet World Stats (2017). Available online at: https://www.internetworldstats.com/emarketing.htm

Klimt, B., and Yang, Y. (2004a). "Introducing the Enron Corpus," in *CEAS 2004 - First Conference on Email and Anti-Spam* (Mountainview, CA). Available online at: http://nl.ijs.si/janes/wp-content/uploads/2014/09/klimtyang04a.pdf

Klimt, B., and Yang, Y. (2004b). "The enron corpus a new dataset for email classification research," in *Machine Learning: ECML 2004,* eds J. F. Boulicaut, F. Esposito, F. Giannotti, and D. Pedreschi (Berlin: Springer), 217–224.

Lavoie, B. (2014). *The Open Archival Information System (OAIS) Reference Model: DPC Technology Watch Report 14-02 October 2014, 2nd Edn*. York: Digital Preservation Coalition. Available online at: http://www.dpconline.org/docs/technology-watch-reports/1359-dpctw14-02/file

Luca, M., Wu, T., Couvidat, S., Frank, D., and Seltzer, W. (2015). "Does google content degrade google search? Experimental evidence," in *Harvard Business School Working Paper, 16-035* (Cambridge, MA: Harvard Business School). Available online at: https://www.hbs.edu/faculty/Publication%20Files/16-035_2260fc69-1f63-466f-b4df-7957e77e2a3f.pdf

Lynch, C. A. (2003). "Colliding with the real World: heresies and unexplored questions about audience, economics, and control of digital libraries," in *Digital Library Use: Social Practice in Design and Evaluation*, eds A. Bishop, B. Butterfield, and. N. Van House (Cambridge, MA: MIT Press), 191–216.

Mayer-Schönberger, V. (2011). *Delete: The Virtue of Forgetting in the Digital Age*. Princeton, NJ: Princeton University Press.

Merrin, W. (2014). *Media Studies 2.0*. London: Routledge.

Moretti, F. (2000). Conjectures on World literature. *New Left Rev.* 1, 54–68.

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

Moss, M. (2005). The hutton inquiry, the president of Nigeria and what the butler hoped to see. *Eng. Hist. Rev.* 120, 577–592. doi: 10.1093/ehr/cei121

National Archives (2013). *Bulletin 2013-02: Guidance on a New Approach to Managing Email Records*. NARA Bulletin. Available online at: https://www.archives.gov/records-mgmt/bulletins/2013/2013-02.html

Nicholas, D. (2007). "If we do not understand our users, we will certainly fail," in *The E-Resources Management Handbook*, ed R. Anderson (UKSG). Available online at: https://www.uksg.org/publications/ermh

Ogrysko, N. (2016). *Most Agencies Say They'll Meet Year-End Records Management Deadline*. Federal News Radio. Available online at: https://federalnewsradio.com/agency-oversight/2016/03/agencies-say-theyll-meet-year-end-records-management-deadline/

Padia, K. (2012). *Visualizing Digital Collections at Archive-It.* Dissertation/master's thesis, Old Dominion University, Norfolk, VA. Available online at: https://digitalcommons.odu.edu/cgi/viewcontent.cgi?article=1008&context=computerscience_etds

Pitti, D. V. (2006). Technology and the transformation of archival description. *J. Arch. Organ.* 3, 9–22.

Prabhakaran, V., Reid, E. E., and Rambow, O. (2014). "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, 1965-1976*. Available online at: https://cs.stanford.edu/~vinod/papers/EMNLP_genderpaper_final.pdf

Rea, S. (2015). *Six Degrees of Francis Bacon*. Carnegie Mellon University. Available online at: http://www.cmu.edu/news/stories/archives/2015/october/francis-bacon-launch.html

Rosenthal, D. (2017). *DSHR's Blog*. Available online at: http://blog.dshr.org/search/label/storage%20costs

Schofield, T., Kirk, D., Amaral, T., Schofield, G., and Ploetz,. T. (2015). *Archival Liveliness: Designing With Collections Before and During Cataloguing and Digitization*. Digital Humanities Quarterly. 9:3. Available online at: http://www.digitalhumanities.org/dhq/vol/9/3/000227/000227.html

Schott Syme, H. (2003). Becoming speech: voicing the text in early modern english courtrooms and theatres. *Compar(a)ison, I* 1, 107–124.

Shabou, B. M., and Sokhn, M. (2017). "The new information technologies at the service of historical and cultural heritage and tourism promotion," in *Integrating ICT in Society*, eds Iana Atanassova et al., (Zagreb: FF Press), 219–234.

Smith, C. (2017). *The Pixel 2' Unlimited Photo Storage isn't Exactly Unlimited*. BGR. Available online at: http://bgr.com/2017/10/06/pixel-2-features-unlimited-photo-storage/

TNA (2016). *The Application of Technology-Assisted Review to Born Digital Records, Transfer, Inquiries and Beyond- Research Report*. London: The National Archives. Available online at: http://discovery.nationalarchives.gov.uk/results/r?_q=Records+of+the+UK+Atomic+Energy+Authority%27s+production+divisions+touching+all+aspects+of+atomic+energy+research%2C+day+to+day+procedures%2C+administrative+functions+and+industrial+relations

TNA (2018), *Online Catalogue*. Available online at: http://discovery.nationalarchives.gov.uk/results/r?_q=Records+of+the+UK+Atomic+Energy+Authority%27s+production+divisions+touching+all+aspects+of+atomic+energy+research%2C+day+to+day+procedures%2C+administrative+functions+and+industrial+relations

Turkel, W. J., Kee, K., and Roberts, S. (2013). "A method for navigating the infinite archive," in *History in the Digital Age*, eds Toni Weller (London, Routledge), 61–75.

Turnbaugh, R. C. (1983). Living With a Guide. *Am. Arch.* 46, 449–452.

Underwood, T. (2014). Theorizing research practices we forgot to theorize twenty years ago. *Representations* 127, 64–72. doi: 10.1525/rep.2014.127.1.64

Viégas, F. B., Golder, S., and Donath, J. (2006). "Visualizing email content: portraying relationships from conversational histories," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (New York, NY: ACM).

Walter, C. (2005). Kryder's Law -The doubling of processor speed every 18 months is a snail's pace compared with rising hard-disk capacity, and Mark Kryder plans to squeeze in even more bits. *Sci. Am.* 293, 32–33. doi: 10.1038/scientificamerican0805-32

Waugh, A. (2014). *Email - a Bell Weather Records System.* Recordkeeping Roundtable. Available online at: http://rkroundtable.org/2014/06/30/email-a-bellwether-records-system/

Weinberger, D. (2011). *Too Big to Know*. New York, NY: Basic Books.

Zachte, E. (2011). *Wikipedia Visualisations. Erik Zachte's Wikipedia / Wikimedia Portfolio*. Available online at: http://infodisiac.com/Wikimedia/Visualizations/

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.