Journal of Contemporary Archival Studies

Volume 4 *Governance of Digital Memories in the era of Big Data*

Article 3

2017

Our Digital Legacy: an Archival Perspective

Michael S. Moss
Northumbria University, Newcastle, michael.moss@northumbria.ac.uk

Tim J. Gollins
National Records of Scotland, Edinburgh and The University of Glasgow, timothy.gollins@glasgow.ac.uk

Our Digital Legacy: an Archival Perspective

Cover Page Footnote
The authors would like to thank Norman Gray, Andrew Hoskins, Ruth Paley, Iadh Ounis, Craig MacDonald, Graham Mcdonald, Alessia Ghezzi, Tim Ellis, and David Thomas for their assistance and advice.

OUR DIGITAL LEGACY: AN ARCHIVAL PERSPECTIVE

Many have discussed and debated the preservation of traces from our digital world, mostly from a technical perspective. A great deal of this discussion has been predicated on the false assumptions that little will survive (the so-called digital black hole) and that rapidly changing file formats and software upgrades will make what survives difficult, if not impossible, to read. This narrative has been coupled with alarmist stories about the high cost of digital curation in trusted digital repositories. Taken together, all this scaremongering has diverted attention from the other core principles of archival science: appraisal (what to keep), sensitivity review (identifying material that cannot be disclosed for ethical or legal reasons), and access.[1] The way that archival science uses these core principles to respond to the "supernova" of digital material that will actually survive will define our digital legacy.

Working Practice but Not as We Know It

Many in the records management and archival communities have been slow to recognize that the digital world has fundamentally changed working practices, just because of the way it operates. E-mail by default keeps every message we send or receive.[2] It does not store them in anything that resembles a manila file, but it does hold them on the system. Even if they appear to be deleted, they are, more often than not, still there somewhere; in the analog world, a conscious decision would have had to be made to file a copy.

Filing or the registration of documents in the modern era had its origins in the Renaissance and was in some sense systematized by Luca Pacioli in his chapter "De computis et scripturis" (Details of calculation and recording) in *Summa de arithmetica, geometria, proportioni et proportionalita*, published in 1494.[3] The whole purpose was to ensure documents could be easily retrieved whether they were held sequentially in a registry or, from the early nineteenth century in the United Kingdom Civil Service and its dependencies, in files.

As bureaucracies grew, these systems of recordkeeping became larger and more complex.[4] Nevertheless they had strict rules about what should be filed and, importantly, had at least top-level aids to discovery, such as well-constructed series or file plans. There was an established process of registration overseen by clerks who were independent of the core business operation and responsible for checking documents or files in and out. It was extremely difficult for an individual civil servant creating documents to avoid or bypass the system, and its inherent checks on recordkeeping process and procedure. Civil servants knew what was expected of them. Not only were dockets and letters to be prepared and written in a common form, but even informal conversations were to be minuted.[5] Duplication of documents was expensive and limited, which

[1] Rusbridge, "Excuse Me."
[2] Waugh, "Email."
[3] Pacioli, "De computis et scripturis."
[4] His Majesty's Stationery Office, "Notes for the Use Registry Branches"; Foreign Office, "Report on the Reorganization of Foreign Office Registries."
[5] Moss, "Where Have All the Files Gone?"

constrained the way these record systems were designed. Until the advent of the photocopier it was only possible for copying clerks to make three or four legible carbon copies on a typewriter; one of the copies was then filed sequentially in a letterbook. Even the photocopier required the physical activity of photocopying and distribution by a small army of clerks. Notwithstanding these limitations, the restricted capacity of carbon copying and the introduction of the typewriter did at least do away with the need for elaborate cross-referencing and indexing, as copies could be distributed across files.

Paper systems of recordkeeping were long-lived, both with respect to the records themselves and to the administrative structures that created, managed, and kept the records. The death of paper records rarely happened unless an organization ceased operating, for example ministries established to meet the exigencies of wartime. Registries and their structures were amazingly resilient and persistent, sometimes over hundreds of years.[6]

With the advent of networked personal computers and the internet, in many jurisdictions registries and the systems they implemented were swept away. The computers seemed to serve the same purpose, as they too kept records, albeit not in structures that resembled paper files unless someone went to the trouble of creating them.[7] In the same process, and in the name of efficiency, secretarial posts were slashed and managers, most of whom were unfamiliar with filing and registration, began typing their own letters and e-mails. File references vanished almost overnight. All that was left were e-mail headers, which more often than not provide little indication of what the contents related to or how they are connected to any previous exchange. E-mails began to replace telephone conversations, which at least in the United Kingdom Civil Service, if significant, had always been minuted.

Gradually, as budget cuts and the speed of transactions left little time for reflection and reflexivity, officials began to develop policy through e-mail threads rather than through carefully crafted minutes, memos, or letters.[8] Moreover, largely as a precaution, more and more people were copied in and, as a result, the interchange left its footprint on many recipients' servers. The consequence, as we now know, is that servers and hard drives are littered with an enormous quantity of material that we might characterize as a blinding explosion of information, indeed a supernova.[9]

In recent times, the advent of collaboration tools and environments has created an even more fluid process where the very concept of a document and a version has been eroded.[10] E-mails now communicate links to shared workspaces or team sites, and wiki software enables the creation of content where, while it is possible to establish which word was input by which user, the concept of authorship is becoming moot.

---

[6] Foreign Office, "General Correspondence from Political and Other Departments."
[7] Moss, "Where Have All the Files Gone?"
[8] Moss, "The Hutton Inquiry."
[9] Allan, "Records Review." That said, as was shown recently following James Comey's termination as the director of the U.S. Federal Bureau of Investigation, some still engage in the practice of writing memos. See Rosenwald, "James Comey's Memo Has Shaken a Presidency."
[10] Allison et al., "Digital Identity Matters."

This new environment frustrates record managers. "Where is the record kept?" and "How will we ever find the record among all this stuff?" they cry as they mourn the loss of the registry file. But in all this loss and grief something really vital is being missed: the record system has not gone away; it has merely been transformed out of all traditional recognition.

The human-mediated system described above was not, as is often implied, only the system by which records were kept. It was actually the system by which they were created. Once we can begin to recognize this, we may want to reassess our views of e-mail systems and collaboration tools and environments, and analyze what has been lost in the rapid transition to digital.[11] E-mail systems have mechanisms for keeping and finding their content, and collaboration tools and environments even more so. They do not have some *control* features of the registries, but the absence of these may not be catastrophic. They can also keep stuff (actually the stuff they keep is the record—it is more spread out) very reliably. They have the great advantage that they manifestly work as far as businesses are concerned, allowing individuals to create, use, and reuse information to the benefit of the business.

The benefits of Electronic Document and Records Management Systems (ERDMS), the oft-cited essential technocentric replacement for the registry, may be a little harder to find and articulate. It is sufficient for the time being to observe that rather than serving as the place where records are created and give benefit to a business, often EDRMS are the places where records go to die (unlike paper registries). And even worse, these EDRMS require significant effort from every individual user to put the stuff there in the first place.

Too Much Focus on Preservation Already

Since Jeff Rothenberg's iconic 1995 article, the archival community has been fixated on the technical challenges of digital preservation.[12] The development of the Open Archival Information System (OAIS) reference model and other subsequent ISO (International Organization for Standardization) standards has only served to reinforce this technical preservation bias.[13]

More recently the development of the concept of "parsimonious preservation" by Tim Gollins at the National Archives of the United Kingdom, reflecting the little cited work by Chris Rusbridge and based on the ongoing work of David Rosenthal, has begun to demonstrate that many aspects of these technical concerns are misplaced.[14]

While there are many and varied threats to the successful curation of digital material, the impression given by the marketers of many digital preservation systems and by much received

[11] Waugh, "Email."
[12] Rothenberg, "Ensuring the Longevity of Digital Documents."
[13] Consultative Committee for Space Data Systems, "Space Data and Information Transfer Systems—Open Archival Information System"; Consultative Committee for Space Data Systems, "Space Data and Information Transfer Systems—Audit and Certification of Trustworthy Digital Repositories."
[14] Tim Gollins, "Parsimonious Preservation: Preventing Pointless Processes!" in *Online Information 2009*, 2009, 75–78, http://www.nationalarchives.gov.uk/documents/information-management/parsimonious-preservation.pdf; Rusbridge, "Excuse Me . . . Some Digital Preservation Fallacies?"; Rosenthal, "Formats Through Time."

wisdom is that imminent technological (software or data-format) obsolescence is the primary threat. This gives rise to the belief that the only way to successfully start doing digital preservation is to invest in a large, technically complex, expensive, and difficult to operate integrated digital preservation system. Using the principle of parsimony, Gollins argues that, while the threat of technological obsolescence is real in some particular cases, a much more imminent threat is poor capture and inability to achieve safe and secure storage of the original material.[15] By applying the principle of parsimony to digital preservation, institutions can find ways forward that are incremental, manageable, and affordable, and that achieve the goal of securing digital material for the next generation.

The first wave of archives have now had the experience of actually curating digital records. They have discovered that the challenges are not in the deep aspects of file format obsolescence (if indeed they ever were), or in debates between emulation and migration, but in the bits and pieces and trivia of human inconsistency in the use of systems that created the records.[16] These variations in human use render the records sufficiently variable in structure to break or clog idealistically constructed automated preservation workflows.

Robust and variation-tolerant workflows are hard to construct. Their success often crucially depends on simplification—of the tasks, of the metadata (discussed later), and of the assumptions about the records being processed. In this domain, less is more, and the parsimonious concept of only doing the minimum necessary for immediate stewardship must come to the fore.[17] We should particularly understand that most of the difficulties do not arise from the preservation aspects of the workflows but from the aspects that address other archival challenges such as describing and presenting the materials for use. Once again it is not the container that presents the challenge but the contents.

The traditional response to such challenges from many in the archival community has been the call for (and generation of) metadata, with little thought of how it can be embedded effectively and effortlessly in the day-to-day work of busy front-office staff.[18]

Metadata

What is metadata? The classic definition asserts that it is "data about data." In some sense this is of course a misnomer: data is just data—stuff to be processed by some sort of computing or information system. Indeed, the material that is regarded as metadata by one system may be the core data of another (for example, e-mail headers, of no interest to the average user, are the core data created and processed by the underlying e-mail communication systems, the message content being but a single field of many).

To be clear, the metadata we are concerned with here is termed by some "descriptive metadata," data that describes or augments individual instances of the core data content (an archival catalog

[15] Gollins, "Parsimonious Preservation"; Gollins, "Putting Parsimonious Preservation into Practice."
[16] Granger, "Emulation as a Digital Preservation Strategy."
[17] National Archives, "The National Archives—Our Role—Digital Preservation."
[18] Currall et al., *"No Going Back."*

description, a finding-aid entry, or exposure data for digital photographs). Traditionally in both archival and library communities, such descriptive metadata has been produced by hand. This is the source of the challenge facing the archival community in the digital transition; hand methods will not scale.

Throughout we need to focus on what might be best described as industrial scale processes, as we are dealing with data generated by machines with industrial scale capacity. It is easy when you see one e-mail printed out on paper and a typewritten letter side by side to think that it might be possible to curate e-mail at volume in the same way you do a collection of typewritten letters. This is a profound mistake, but as we hope to demonstrate throughout this essay, it does not mean that some skills learned in the analog cannot be carefully translated into the digital environment. Computing scientists, in thinking about efficient retrieval systems, use concepts they term "features" or "significant properties" of an object; it comes as a surprise to many in computer science and archives that these are equivalent to many concepts in archival diplomatics.[19] We believe that openness to such multidisciplinary connections will help us find solutions to many of the digital archival challenges we discuss here.

There is also a danger in the assumption that descriptions must be in the same form or structure as those for paper records. Archive users will certainly wish to continue to engage with individual texts of records. However, we should not assume that they will find these texts by following the same kinds of trails through the archive using traditional tools or practices. In addition, there are research questions that only a wholly born-digital collection can answer (where the full texts of the records are available for computer or "natural language" processing) that it would not be feasible to answer from an equivalent paper collection.

The combination of the need to industrialize and the potential need for different types of archival description offers an opportunity rather than a threat. Certain facts may in principle be automatically extracted from digital records, including dates, personal names, and places. Summaries could, theoretically, be created using tools developed from the decades of research into text processing and information retrieval. None of these issues is trivial, and the challenges of applying these techniques to the heterogeneous and messy data that form future archival records are significant. However, our own work suggests that such tools are within reach if archivists and computer scientists can collaborate.[20] Investment in such pragmatic and practical research and development will undoubtedly succeed in providing archivists with the essential tools they need. To benefit from these developments, the archival community needs to radically change its way of thinking.

What Ever Shall We Keep?

In recent times, many in the archival community have fondly imagined that they can influence the management of information and mandate good practice from, as it were, the cradle to the

[19] Underwood, Isbell, and Underwood, "Grammatical Induction and Recognition of the Documentary Form of Records"; Duranti, ed., *Records Management Journal*.
[20] McDonald et al., "Towards a Classifier for Digital Sensitivity Review"; Gollins et al., "On Using Information Retrieval for the Selection and Sensitivity Review of Digital Public Records."

grave. They insist that organizations adopt electronic document records management systems (EDRMS).[21] Leaders of organizations resent such interference unless it can be shown to add value in the front office without additional cost, or at least to payback any investment. This is almost impossible to achieve as the archival community is not experienced in systems design or the needs of the business (their perspective is naturally historical). Also, though they may be invited to make suggestions, archivists' concerns will have relatively little influence as fundamentally such systems exist to augment the effective running of an organization. It was ever thus, but increasingly so as organizations become asset light. This article argues that the archive has to take what it is given, from the context in which the users have chosen to use it.[22]

Once it becomes apparent that large swathes of digital content survive, and that it remains readable, decisions, however ad-hoc, have to be taken about what to keep. This can range from everything, technically practical although still costly and fraught with legal risks, to a selection of some sort. If a choice is to be made, it should be on some rational basis; however, appraisal emphatically cannot be the same as before. In the paper world, where records were held in files, this was relatively easy, as the structures used by organizations to navigate their records could be used as the basis for selection (for example, registry index books and file-plans).[23] Private papers were more difficult as their organization often lacked structure, but the volume was not large and they came primarily from high-profile individuals from all walks of life.

The first problem is that records, if they are known to survive, are legally discoverable. This is less of a concern in the public sector, which usually enjoys the protection of some form of official indemnity. However, even here there are concerns, shared with the private sector, when records that are inappropriate for disclosure have passed to the archive.[24]

Reviewing heterogeneous stuff for what is termed "sensitive content" is time-consuming and expensive. Moreover, storing it for a long time is also costly. Cost-benefit analysis is an essential part of the equation in any consideration of preserving and curating our digital memories.[25] In many areas, the digital has by stealth forced the creation of new business models. It is one of the reasons we have big banks and chains of shops, as small, independent bodies simply cannot afford the capital investment or pricing advantages of the behemoths.

The second problem is that records have fundamentally changed in their nature. The term "ephemera" has often been used to describe items of information that are intended to only be of use for a short time and thus, in the long term, are materially insignificant. In many of the new digital environments these separate concepts of short time-frame of utility and long-term material significance have collapsed. For instance, Donald Trump's tweets offer a useful example of digital objects originally intended as short-term that must now be considered to have huge long-term material significance. We believe that this collapse is one of many unexpected consequences of the emergence of new digital forms of information. Moreover, the reason such

[21] Public Record Office, "Requirements for Electronic Records Management Systems."
[22] Verbeek, *What Things Do*, 11.
[23] Grigg, "Report/Committee on Departmental Records"; Wilson, "Modern Public Records Selection and Access."
[24] Campaign against Arms Trade, "Al Yamamah Documents."
[25] Rumsey, ed., "Sustainable Economics for a Digital Planet."

consequences are unexpected derives from an attribution error that sees digital materials as digital surrogates of paper equivalents. The very word "document" illustrates this in the term "electronic document records management systems." Tweets, Facebook posts, instant messages, YouTube videos, and websites simply do not behave like, or have many of the properties of, "documents."[26]

Given that the economic cost of keeping everything *a fortiori* is unacceptably high, how are we going to choose what to keep?[27] To do this, as with much to do with digital, we are going to have to trust people with sophisticated skills that only a few of us understand. When we put our bank card in a hole in the wall far from home, we trust some complex mathematical systems to check that we have enough credit to meet the transaction and that the bank which owns the ATM will not defraud us. Remarkably it is much the same sort of mathematical modeling we need to trust when we enter the world of digital archiving. We need to model the stuff using tools that are only now in course of development.[28]

We can use digital forensic tools to disentangle genres, as components of these tools were built for that purpose.[29] We should be able to use graph and network analysis techniques developed by the police and intelligence communities to analyze the links and discard trails that lead nowhere (for example, in e-mail to identify people who were copied in to a message but did not need to be).[30] We should be able to use emerging federated services to identify duplicates and discard copies that can safely be deleted, that is, those with no annotations.[31]

From the archival perspective, urgent thought needs to be given to the criteria we use to select material for permanent preservation. In the past archivists have hidden behind a cloak of supposed objectivity, but in reality there is a tendency not to keep content that no one is ever likely to use. The criteria previously used were traditionally focused on records that related to policy and strategy, in other words core business, but this is no longer feasible even in the analog world. The surge in interest in family history means that a large community of users want content overflowing with names. In other research contexts, and at another level, the interest in relationship networks, and the availability of powerful tools for network analysis and mapping, even at a meso-level, should influence appraisal decisions.[32] Names and addresses present a problem. Trying to identify policy and strategic information among the grains of sands is going to be difficult, but digital forensic tools are beginning to be able to make some distinctions, and features such as the length of a document or the number and types of words used may offer some clues (although, in the case of 140-character tweets, in counterintuitive ways). All this, combined with the replacement of memoranda with e-mail threads, court transcripts with digital video, and considered political policy documents with tweets, means we are going to keep much more than before, possibly as much as 20 percent rather than the 5 percent that traditionally is the case in

[26] Merrin, *Media Studies 2.0*.
[27] Rumsey, ed., "Sustainable Economics for a Digital Planet."
[28] Klimt and Yang, "The Enron Corpus."
[29] BitCurator Project, "BitCurator"; AccessData, "Forensic Toolkit."
[30] IBM, "IBM I2 Analysts Notebook White Paper."
[31] Still, it may be interesting to note that the mere existence of a copy under different management control could be considered an annotation in itself in many circumstances.
[32] IBM, "IBM I2 Analysts Notebook White Paper."

analog contexts. This, in itself, will add to the cost of processing and storage.

As yet there is no solution to this conundrum, only an awareness that current appraisal methodologies are hopelessly inadequate. 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 catalog to be made?"[33] The production and definition of the "archive" must become collaborative in a co-creation enterprise or what has been described as a "curated conversation" that extends well beyond the existing customer base.[34] There is considerable interest in the broader concept of co-creation that spills over into concepts of heutagogy and an open-context model of learning.[35]

Making Sense of Sensitivity

Once the archive has made a decision to preserve digital content, it is faced with an even bigger challenge: what can safely be released to users? The archival community has largely overlooked sensitivity and even intellectual property rights (IPR) of digital materials, because archivists have been preoccupied with technical issues. However, in this digital age, sensitivity is far from simple. It includes considerations of the relationship between surveillance and democracy that has so far produced contradictory responses in both policy and action. This is overlaid by further considerations of the identification and preservation of detailed evidence to enable restitution of injustices, while respecting the privacy and information rights of all.

Surveillance Society

During the last ten years, the sensitive nature of personal information has attracted mass-media coverage. The revelations of Edward Snowden in 2013 about the activities of the National Security Agency in the United States and the release of a trove of documents aroused a storm of protest around the world about the way in which security agencies in many countries collect data about individuals. In the wake of Snowden's revelations, the Japanese government passed legislation imposing "draconian penalties on leakers or seekers of information that the government, with no necessarily independent oversight, deems secret, according to standards left undefined." This too sparked off a storm of protest.[36] Such concerns have become more acute with reports of the way in which companies harvest and analyze personal data to manipulate public opinion, notably in the United Kingdom's Brexit referendum and the U.S. elections.[37]

The coincidence of Snowden's revelations of 2013–14 and George Orwell's chilling novel *1984* has not been lost on commentators, who conjure up a nightmare world of Big Brother empowered by quantum computing on a scale Orwell could never have imagined. Orwell has been coupled with William Gibson, who in 1984 wrote *Neuromancer* in which he coined the

[33] Featherstone, "Archive."
[34] Huvila, "Participatory Archive."
[35] Heick, "The Difference between Pedagogy, Andragogy, and Heutagogy"; Garnett, "The Heutagogic Archives."
[36] Hoffman, "Society Struggles to Adapt to Post-Privacy Age."
[37] Flynn, "What Brexit Should Have Taught Us about Voter Manipulation." In the United States the newly appointed special council will undoubtedly comment on such practices. Ford, "What the Special Counsel Appointment Means."

term "cyberspace" and conjured up a dark vision of millions of operators harvesting data from ubiquitous computers. Having spent his career helping security agencies collect such data, Edward Snowden has become an advocate of the right to data privacy and has called for a fundamental rethink of the role of the Internet in our lives—and the laws that protect it.[38]

He is not an unthinking critic of security services. He admits that they do good things that need to be done in an uncertain world, but he is uneasy about the way in which the biggest internet providers have had their arms twisted into allowing access to personal data that they hold as a result of their businesses.[39] Personal data has become a tradeable commodity by analytics companies.

Well before Snowden's revelations, in the United Kingdom the information commissioner, who has oversight of data protection, commissioned a penetrating report on the surveillance society and warned of its dangers. The report to the information commissioner from the Surveillance Study Network began, "We live in a surveillance society. It is pointless to talk about surveillance society in the future tense. In all the rich countries of the world everyday life is suffused with surveillance encounters, not merely from dawn to dusk but 24/7."[40]

However, rather than castigating the surveillance society as "something sinister, smacking of dictators and totalitarianism," the authors characterized it in Weberian terms, "as progress towards efficient administration," a natural extension of "modernity." Such a perspective, in the authors' view, avoids the trap of thinking of surveillance as a token of digital technology. The state has recorded information about people for at least two thousand years. However, the report recognizes that digital technology made possible the rapid interchange of information with the inherent danger of "function creep," as data collected for one purpose can easily be used for another. It recommends that it should not be left to the individual to challenge the inappropriate use of personal data: "The emergence of today's surveillance society demands that we shift from self-protection of privacy to the accountability of data-handlers."[41]

Following the loss of discs containing child benefit data by Her Majesty's Revenue and Customs in the United Kingdom in October 2007, the House of Commons Home Affairs Committee launched an inquiry into the surveillance society. Although there was agreement that a crude description of the United Kingdom as a surveillance society was inappropriate, there was every danger that citizens might reach that conclusion "unless trust in the Government's intentions in relation to data and data sharing is preserved."[42]

Between then and now, and with less drama than the outcry against Snowden's revelations, the United Kingdom Information Commissioner has taken steps to respond to the Committee's concerns by publishing a Privacy Impact Assessment Handbook and encouraging a "Privacy by

[38] Snowden, "Here's How We Take Back the Internet."
[39] McCarthy and Morgan, "Rights and Commons."
[40] Ball and Wood, "A Report on the Surveillance Society for the Information Commissioner," 1.
[41] Ibid., 9, 8.
[42] Home Affairs Committee, "A Surveillance Society?" 5.

Design" approach to building privacy safeguards from first principles.[43] Nevertheless, concern has mounted following widespread allegations of the manipulation of the democratic to process by the use of sophisticated data analytics to target swing voters.[44]

Surveillance and Post Privacy

Much more recently, in response to the outcry following Snowden's revelations, the Intelligence and Security Committee of the British Parliament published a comprehensive review of the full range of intrusive capabilities available to the UK intelligence Agencies, in which, while defending the principles of surveillance in a democracy, they recommended "that the entire legal framework, as it applies to the intelligence Agencies, needs replacing." The purpose of such a comprehensive overhaul of the governance of surveillance would be to improve transparency concerning the work and oversight of the intelligence agencies and thus to improve public understanding and reinforce confidence in their work.[45]

The joint final event of a European Union–funded project, "DEMOSEC: Democracy and Security," took place at the end of October 2014. One of the sessions explored this central theme:

> In the context of surveillance and democracy, the principles of consent, subject access and accountability are at the heart of the relationship between the citizen and the information gatherers. The individual data subject has the right to at least know what data is being collected about them and by whom, how it is being processed and to whom it is disclosed. Furthermore, they have rights to inspect the data, to ensure that it is accurate and to complain if they so wish to an independent supervisory authority who can investigate on their behalf.[46]

The statement may seem uncontentious and accords with the views of Tim Berners Lee, one of the founders of the World Wide Web, who has called for a new model of privacy on the web: "I would like us to build a world in which I have control of my data. I can sell it to you and we can negotiate a price, but more importantly I will have legal ownership of all the data about me."[47]

However, not everyone agrees, and some claim we live in a post-privacy world where anything goes. James Der Derian of the University of Sydney has posited "a much more disturbing picture of a late, a very late modernity, in which the dystopic visions of Orwell and Gibson are converging in a world of über surveillance, diminished privacy and minimal dissent."[48]

The link to modernity echoes the Surveillance Society Network's reference to Weber, who believed modernity was aligned with rationalization on an industrial scale that was reflected in popular disenchantment which is easily manipulated by the unscrupulous.

---

[43] Information Commissioner's Office, "Conducting Privacy Impact Assessments Code of Practice."
[44] Flynn, "What Brexit Should Have Taught Us about Voter Manipulation."
[45] Intelligence and Security Committee of Parliament, "Intelligence and Security Committee of Parliament—Privacy and Security," 8.
[46] Cocq, "DEMOSEC Day 2," 40.
[47] Curtis, "Sir Tim Berners-Lee Calls for New Model for Privacy on the Web."
[48] Der Derian, "Edward Snowden and Cyber-Zombies."

The Privacy and Accountability Dilemma

There is another side to this argument that is equally dark and concerns access to personal records to right wrongs. There are many examples, such as the files of the Stasi, the secret police of the former East Germany, which are being made available online.[49] The website declares on its opening page, "The better we understand dictatorship the better we can shape democracy." This is a bold claim and suggests that if we can comprehend dystopia, there is hope for a better future. It is not far removed from the call of DEMOSEC.[50]

In the United Kingdom the recent revelations about the Hillsborough football disaster in 1989 when 96 people died and 786 were injured has depended on locating evidence from individuals that contradicts the official version of events provided by the police.[51]

Furthermore, the discovery of the scale of abuse and exploitation of children in recent times, whether it be sexual abuse or the forced removal of children from their parents, has rightly led to an expectation of holding institutions and wrongdoers to account (for example, in the United Kingdom, the Independent Inquiry into Child Sexual Abuse—IICSA).[52] Allegations of abuse can only be investigated if records can be located. In Australia, as in many other countries, much of the abuse took place in children's homes. A huge national project—Find and Connect—has identified where records are held that might benefit survivors of wrongs to be redressed.[53] The need to preserve personal records that affect the weak and vulnerable, and at the same time protect those in authority from unwarranted allegations, accords with the concepts of accountability of data owners and governance. It has nothing to do with exaggerated views of a post-privacy world where apparently the ends always justify the means.

It is, however, not as simple as that. There is an underlying ambiguity in our desire for privacy and the fact that a billion people have signed up for Facebook since it was launched in 2004. There, many happily share personal information. We like our GPS satellite navigation devices to know where we are to get us to where we want to go. We like our mobile devices to identify where we are so we can find food to eat, attractions to see, and so on, but we, perhaps, are prepared to accept that the provider sells this information to someone else to bombard us with information about places to visit, eat, and buy things.

There are many who celebrate what they see as the liberation of the post-privacy world with the tagline, "We are all celebrities in post-privacy age."[54] It may also not occur to some people that less benign organizations are similarly harvesting this information and trying to associate it with other information that might suggest we are linked to terrorist organizations. Even more perversely when a terrorist does break through, as in Boston or London, or girls go to join ISIS,

[49] Federal Commissioner for the Stasi Records, "BStU—Homepage."
[50] Cocq, "DEMOSEC Day 2."
[51] Hillsborough Independent Panel, "Report of the Independent Panel."
[52] IICSA, "Independent Inquiry into Child Sexual Abuse."
[53] Find and Connect Project, "Find and Connect Web Resource."
[54] Auchard, "We're All Celebrities in Post-Privacy Age."

we complain loudly that the security services should have done more to catch them. We are thus presented with a huge and unresolved dilemma: on the one hand, we wish to protect our privacy; on the other, we want trustworthy records to be kept so wrongs can be redressed, innocent protected, and the tractability of the internet can be used to make contact with people we do not know all around the world.

Governance

This dilemma cannot be resolved by addressing issues of surveillance or recordkeeping alone; it is about government and governance. Governments set the boundaries in which our security services operate, and governments should be called to account if terrorists get through. It is governments and in some cases international agreements that mandate the statutory environment in which public and private sector organizations can use personal data. In child abuse cases, it is the fault of those who ran children's homes and allowed pedophiles to go undetected and unpunished who are accountable. The ways in which social media use our data and keep it secure is mandated by the companies that own them in accordance with the regulatory environments in the countries where their services are delivered.[55] Governments and those responsible for good governance must set the necessary criteria for good recordkeeping by the executive, which will make accountability a reality even long after the event. The much-publicized breaches in security and Snowden's revelations have led to a tightening in privacy regulations, including in the European Union "the right to be forgotten"—to be redacted from the pages of history. Regulation places the onus firmly on data providers, and in many jurisdictions comes with heavy penalties for failure to comply.[56]

The Impact on the Archive

When records are transferred to an archive there is a clear expectation that they will be made public. Digitally born records come with the same expectation, such as in the current plans of the National Archives of the United Kingdom.[57] Once online the content will be indexed by ubiquitous web search engines and content will be easily discoverable in a way it was not in the analog world. This places the archive at the center of this privacy debate, whether most archivists have realized this or not. Archives face a major obstacle in granting access to content against a background of tightening privacy regimes and hardening public attitudes toward inappropriate disclosure. The U.S. Council of Library and Information Resources (CLIR) has warned collecting archives not to take digital content unless it has been reviewed for such sensitivities, because once such material is deposited, archives are exposed to contingent liability and can be "discovered" for litigation.[58] The responsibility and accountability for review varies across different jurisdictions with creators and archives being responsible for different aspects of the process. Nevertheless, wherever the responsibility falls, the challenge remains as to how can this be done if all that is available to be transferred to the archive is a large collection of material with

[55] Merrin, *Media Studies 2.0*.
[56] Article 29 Working Party—European Commission—Directorate General for Justice, "Opinions and Recommendations—Justice."
[57] National Archives, "Archives Inspire."
[58] Redwine et al., *Born Digital*.

all sorts of data types?

In the analog world where papers were organized in files and the default was the wastepaper basket, review for sensitivity was done simply by checking content and either redacting offending items or removing pieces, usually a sheet. Only in extreme cases where a great many names were mentioned were whole files closed. Most sensitive content is personal information, which is now closed in most European countries for between 100 and 110 years (less the age of the individual, if known). If the age of the individual is not known, for minors it is closed for the whole period, and for those deemed to be over the age of sixteen, 80 or 94 years.

There are good reasons for such long closure periods. It safeguards the individual, particularly if the material might affect that person's health and well-being, and it helps prevent identity theft used by criminals and unfortunately by a few law enforcement agents. European countries respect reciprocity so such closure applies to personal information about individuals who are not European citizens, as European countries expect other countries to keep data closed for similar periods, particularly if it has been divulged in confidence by a third party. In any event, with the massive global exchange of information, we are moving toward international conventions, which already exist for some categories of data—for example, through the Geneva Convention.

When we shift into the digital realm, the chances of discovery of inappropriate release of personal information is magnified. This may not even be explicit but implicit through inference from a sequence of data sources that can be pieced together in what are termed "mosaics."[59] This practice is the stock in trade of investigative journalists and security services.

What Is Sensitivity?

Somewhat surprisingly for a subject that is so apparently well understood and obvious to us all, sensitivity is a difficult concept to pin down concretely. It can relate to personal, institutional, political, and national security matters and other connotations depending on context. In the United Kingdom governmental context, the Freedom of Information Act attempts to define it by breaking sensitivity down into some twenty-four exemptions ranging from "National Security" (sec. 24), through "Damage to International Relations" (sec. 27) and "Personal Information" (sec. 40), to "Commercial Information" (sec. 43).[60]

This codification does not really get to the heart of sensitivity, as in some cases it appears to suggest that merely the subject of a document is what renders it sensitive. But that is not the only factor.

Consider, for instance, a text appearing to describe the capability of a piece of military hardware. If authored and published by a journalist and copied into a government system for reference this would not be sensitive (despite being remarkably accurate due to informed guesswork by the

[59] Information Commissioner's Office, "Information in the Public Domain."
[60] "Freedom of Information Act 2000." However, we should note that a few of the exemptions are more procedural in nature e.g. "information available by other means" (sec. 21) and "information intended for future publication" (sec. 22).

journalist). However, the same text authored by a civil servant would potentially be extremely sensitive, simply because of the authority imparted by the author. From this we can see that authorship (who said it) is a facet to consider.

Consider another text, this time describing a commercial agreement with significant market consequences for the companies concerned. In this case, prior to the formal announcement, the document is sensitive; following the announcement it is public knowledge. Likewise, in the public-sector context, during its creation, an economic or trade policy could be very sensitive, but once the resulting policy is published or enacted it is not sensitive at all. From this we can see that time (when it was said) is a significant facet.

Consider yet another text, this time relating to an arms deal with another country. If the country is a modern liberal democracy, where there are well-known and existing friendly relations, then this may have little sensitivity. On the other hand, if it is a country ruled by a dynastic and oppressive regime with which a relationship is more of convenience than natural alignment, the nature of the deal may become highly sensitive. From this we can see that other parties involved (to whom it was said) is another facet that must be considered.

Finally, consider a text from the seventeenth century that describes religious or ethnic minorities. Now consider if that identical text were to be authored and published today. In the context of a historical document, the language (even though now reprehensible) would generally not be considered particularly sensitive. In the context of a modern document, the reverse would be true. From this we can see that the zeitgeist of publication (the context in which it was said) is also critical.

One additional factor plays a part in determining sensitivity: the jurisdiction in which the sensitivity review takes place. The sensitivities defined in the United Kingdom Freedom of Information Act represent only one of a large number codifications around the globe.[61] Even in the United Kingdom, the codification of sensitivity in Scotland is distinct from the rest of the country, and while in this case the differences are minor, the distinctions with other jurisdictions are myriad.[62]

Thus, sensitivity can be thought of as depending on who said what to whom, when, in what context, and the jurisdiction of review. As humans, we are easily able to detect and consider these nuances and distributed contexts when we look at a document or text. However, in the case of digital material, computers may not find it so straightforward.

Automating Sensitivity Review

As we have discussed, the future ability of archives to cope with the industrialized volume of born-digital records requires archives to develop industrial processes. As we have also seen, sensitivity review in particular is a peculiarly human matter that does not easily lend itself to industrialization. We believe, therefore, that we must develop assistance tools to enable humans

[61] Ibid.
[62] "Freedom of Information (Scotland) Act 2002."

to deal with the volume of material to be reviewed.

In general information-processing tools (for example, enterprise search engines, e-discovery tools, and forensic examination tools) process either the textual content of a record or metadata that was created at the time the record was created (e.g., as stored in an EDRMS or recorded in the SharePoint site or with the file properties on a shared drive). We also know that the metadata generated at the time of creation is notoriously unreliable, partial, or often almost entirely absent.

From our thought experiments above we can see that only one of the six aspects that drive the sensitivity of a record is likely to be held explicitly in the text of a record, the what aspect. The who, to whom, and when aspects may be present in original metadata but in general will not be. Finally the context is absent. Our recent work sponsored by the Glasgow University KE fund, the National Archives, the ITAAU (IT as a Utility Network), the National Records of Scotland, and the Welsh government has confirmed these issues and framed the context for research in this challenging domain.[63]

Given this position, how can we envisage an assistance tool for digital sensitivity review?

Text Processing Systems

Before we can think about an assistance tool for digital sensitivity review, we should look at other types of tool that process text to help users find information or make decisions based on the content of documents and what these tools have in common.

Many modern tools of this kind preprocess the documents they are working on into a form known as a "bag of words," where each document is represented as a list of the words in the document with a count of the number of times the word occurs.[64] In addition, the preprocessing may also record where in the document each word occurs (e.g., by counting words or letters from the start of the document). These fundamental representations have emerged from decades of research into search and information retrieval.

The representations used in searching the internet are extensions of this, as in addition to the words and their position (or proximity, which is calculated from position), they also consider how web pages are related to each other in the network of links to other pages.[65] For these representations to work well there needs to be a sufficient density of linking and that linking needs to, in some way, represent a human view of the value of the page linked. This value may be no more than "here is the page about subject X," but nevertheless, the aggregation of these (mostly human-created) links encodes some sense of the value. It is this encoded value judgment that most web search engines exploit. Unfortunately, in the collections of documents that make up the records of most organizations, such links, while they may appear to exist to some extent,

[63] McDonald et al., "Towards a Classifier for Digital Sensitivity Review"; Gollins et al., "On Using Information Retrieval for the Selection and Sensitivity Review of Digital Public Records"; Gollins, "The National Archive, Big Data and Security."
[64] Salton and Buckley, "Term-Weighting Approaches in Automatic Text Retrieval."
[65] Page et al., "The PageRank Citation Ranking."

are not generally present in sufficient density to be exploitable in the same way.

All of these facts about a document—counts of words, position, links, and so on—are known as "features." So we can think of the representations of documents in these tools as being collections of specific features that attempt to capture the significance of the documents in the context of a particular problem (web search or e-discovery).

What almost all of these representations and features have in common is a fundamental focus on the subject matter, or what the text is about. As we have seen above, sensitivity is only partially concerned with this aspect of a document.

These tools do not generally capture or exploit any features from the document or its context that capture the who, to whom, when, or anything about the wider context. As a consequence, if we are to envision tools to assist humans in identifying sensitivity, then we need to consider how we might represent these aspects and what novel features of documents our new tools need to exploit.

Features for Sensitivity

What features might we envision that could go some way to encode the missing information? Perhaps we should start by asking what features human beings examine and use when they make decisions about sensitivity?

This may at first appear a simple matter to understand; we could just ask sensitivity reviewers. Like many people who use their accumulated experience to make nuanced decisions, reviewers can seldom articulate in general what it is that they are focusing on when making decisions about reviewed records.[66] A good deal of detailed and careful anthropological observation, interviewing, and analysis is required to even begin to understand what is going on. Notwithstanding the need for such research, we still may be able to make some independent progress by examining our questions of who, to whom, when, and context.

We can imagine a component of a tool that could estimate the authorship of a document or e-mail by examining patterns in the title, headers and footers, or salutation. While such techniques have only relatively recently been developed in the field of computing science, the concept is familiar to any archivist steeped in the discipline of medieval diplomatics. Similarly, given the notorious unreliability of system-assigned dates in document metadata, a system of automated heuristics could be developed to determine a "most likely" date for a document by examination and comparison of dates in the document with those held in metadata. By a similar approach to analysis, some work has already been done to establish an approach to defining "speech acts"[67] in presidential records in the United States; once this has been done we could imagine further

[66] Gladwell, *Blink*.

[67] A speech act in linguistics and the philosophy of language is an utterance that has performative function in language and communication. Speech acts are commonly taken to include such acts as promising, ordering, greeting, warning, inviting, and congratulating.

automated structural analysis to extract distribution lists and other similarly useful indicators.[68]

Finally, we could establish some comparison of sensitivity at the time of record creation with sensitivity at the time of transfer to an archive by comparing the record with online corpora that are contemporary to the two times. This might go some way to encoding the context or zeitgeist. The semiotics community's study of pragmatics should also provide further insights into decoding the representations of context.

Of course, even with all of these potential sources of features, it is only by careful consideration of the jurisdiction of review, the required sensitivity review process, the humans that do it, and comparison of their results with our imagined tool that we will be able to establish which are the truly useful indicators in each case.

Separating Wheat from Chaff

Notwithstanding the establishment of a set of features, considerable research remains to be done in terms of both the internal representation of the features (see, by contrast, the "bag of words" model mentioned above) and, also, into the intrinsically interlinked question of the best algorithm or approach used to assist the reviewer. A number of techniques could be employed drawing on "ranking" (from classical information retrieval) to "classification" and "clustering." Any of these are also likely to make use of approaches from the discipline of machine learning such as "topic modeling" and "learning to rank."[69]

Recently researchers at the University of Glasgow have been working on the challenges of sensitivity review of records to be deposited at the National Archives of the United Kingdom. Their approach has been to develop elements of assistive technology that can predict the sensitivity of records and use these predictions to make the work of human sensitivity reviewers more efficient and effective. Techniques using machine-learned classifiers have shown that some sensitivities can be predicted, but the perfect prediction of the sensitivity of a record will likely remain a difficult challenge for some time to come.[70] Other work has already been done in the application of "topic modeling" and related artificial intelligence techniques to the challenges of checking the security marking of U.S. State Department cables.[71]

Separate again from the algorithmic challenge is the consideration of the best way to present collections of records to the reviewers to ease their task. If we assume that the imagined tool has the capability to learn from the actions of the user, and also that trust in the machine will be critical to acceptance of the approach, what order should records be offered for review? We are used to search systems presenting documents in a "most likely to be relevant" rank order, and while this might give the user confidence in the machine, the machine will not necessarily

[68] Underwood, "Recognizing Speech Acts in Presidential E-Records"; Underwood, "Speech Acts and Electronic Records"; Jeong, Lin, and Lee, "Semi-Supervised Speech Act Recognition in Emails and Forums."
[69] Börjeson, "Social Network Methods"; Liu, "Learning to Rank for Information Retrieval."
[70] McDonald et al., "Towards a Classifier for Digital Sensitivity Review"; McDonald, Macdonald, and Ounis, "Enhancing Sensitivity Classification with Semantic Features Using Word Embeddings."
[71] Souza et al., "Using Artificial Intelligence to Identify State Secrets"; Emerging Technology from the arXiv, "Machine-Learning Algorithm Can Show Whether State Secrets Are Properly Classified."

improve the throughput of the process if users are merely confirming the obvious for a long time before they reach a point where their insight is truly needed. A similar difficulty can be seen with a "least sensitive first" approach.[72] One system that might be appropriate is a "most unsure first" ordering.[73] Recent work has shown that this method has promise because it chooses a presentation order to optimize the active learning of the specific distinctions that separate sensitive from nonsensitive in a particular collection.[74] In practice of course, this ordering may be much less intuitive to the potential human reviewer.

Given that human reviewers' decisions appear to be dependent on context, and that born-digital records are by their nature more interconnected, further consideration is needed in terms of other approaches to visualization of the collection of records under review. This aspect may be the least well understood of all. Even in areas of study that are much more common, giving a user an understanding of the overall shape and nature of a collection of digital objects remains an extremely difficult area of research. The challenge of presenting large numbers of interconnected documents is familiar to anyone who has tried to choose the best approach to threading an e-mail chain.

Risk

Management of records is essentially about managing risk in the context of legislation, regulation, and reputation while weighing up costs and benefits. The retention and release of records with sensitive content, particularly anything that can be interpreted as personal information, is a risk. Many organizations, especially those in the public sector, are risk averse. Against the background of tightening regulation, already discussed, it is likely that a risk averse organization will close or destroy records as a precaution out of concern for the damage to its reputation that might result from inappropriate disclosure. Such a heavy-handed approach is unnecessary as not every disclosure is equivalent.

Take, for example, the disclosure of trivial personal details of someone who was born in the 1930s. The person may be dead and the personal information might be long out of date. In the paper world, archives regularly made records containing such information public with little thought of the consequences. However, as with so much else, digital changes the paradigm because such data is much more easily discoverable. It is very difficult to envisage how a completely risk-averse stance could be sustained in a digital world if any records are to be made publicly available.

If we can develop algorithms that rank material according to the probability of sensitivity (and also the differential impacts), then we may be able to persuade those responsible for an organization's risk appetite, usually an audit and risk-management committee consisting entirely of nonexecutives, to adopt a more nuanced approach by downgrading some risks to neutral, with

[72] Notwithstanding this issue, it may emerge that rather than identifying sensitive records, the systems may be more easily tuned to eliminate certain non-sensitive records with greater confidence.
[73] Berardi, Esuli, and Sebastiani, "A Utility-Theoretic Ranking Method for Semi-Automated Text Classification"; Scholer et al., "The Effect of Threshold Priming and Need for Cognition on Relevance Calibration and Assessment."
[74] Berardi et al., "Semi-Automated Text Classification for Sensitivity Identification."

suitable take-down policies to mitigate any limited impacts in the event of objections. Some archival organizations already do this. The National Archives of the United Kingdom put the names of all those who served in the First World War online, even though technically some could still have been alive at the time and thus should have been closed.[75] At that time there were no objections. What is important is that any risks embedded in information holdings are on the risk registers of the institutions, and that those responsible for managing information interact regularly with audit and risk-management committees to achieve a balanced outcome.

The whole topic of sensitivity review is ultimately concerned with risk—the risk of something being released when it should not be and causing real and tangible harm or the risk that an organization's reputation for trustworthiness will be damaged by incorrectly withholding embarrassing but benign material for longer than is permitted. We have seen examples of the realization of these risks for the United Kingdom Cabinet Office, the Foreign and Commonwealth Office, and others within the last few years.[76] For many records and information managers and archivists, at least in Europe, this is unfamiliar territory but is now inescapable, largely because users have very different expectations of the archive in the digital age. The management of risk is a corporate responsibility and can only be delegated within established parameters, leaving records managers and archivists, despite any influence they can bring to bear, virtually no room for discretion.

A Vision of Access to the Digital Archive

The advent of digital text has the potential to completely change the nature of scholarship as all the recent research in the digital humanities has begun to demonstrate. The content of a digital archive is no longer a set of discrete documents with which scholar interacts one by one, but a collection to be examined as a whole, sliced, diced, and analyzed using all the statistical and inferential techniques that have been established by the big data community. New work on graph theory may help, as it may make it possible to identify further and different patterns or hot-spots in content.

The research we describe and envisage on sensitivity review above, will lead to a whole set of new, internal contextual features that can be fed into these techniques to enable users of the archive to explore its labyrinth. These include features such as word length, frequency of certain words or strings, patterns and frequency of certain parts of speech, use of salutations, valedictions, addresses, and dates.[77] If these features enable archivists to discriminate among the nuances of sensitivity, then the same features will most likely be excellent indicators of other properties. For instance, the security services are often described by sobriquets, such as "our friends"; individuals are often hard to identify unambiguously or may only be referred to by their initials or the initials of their job description; and projects can have multiple titles easily taken

[75] National Archives, "Takedown and Reclosure Policy."
[76] Hague, "The Foreign Secretary's Statement to Parliament on the Indian Operation at Sri Harmandir Sahib in 1984"; Allan, "Records Review"; Weiner, "It's Vital We Have Access to the Records on Britain's Colonial Past"; Cook, "The Government Departments Breaching Freedom of Information Law."
[77] McDonald, Macdonald, and Ounis, "Enhancing Sensitivity Classification with Semantic Features Using Word Embeddings."

out of context. By examining the patterns of these features, and the distribution of the properties they represent, novel research questions may emerge or provide pointers to sets of documents that would otherwise have been passed over as of no significance.[78]

As a machine develops and expands its semantic vocabulary, it becomes much easier to link documents together quickly and with greater certainty of the context in which they were produced.[79] Within a body of digital documents that are difficult to attribute, it is already possible using techniques borrowed from the world of linguistics to identify authors and power relationships, which often reveal themselves through salutations and valediction.[80] Much of this is the stuff of diplomatics in the paper world.

It takes a bit of getting used to, but the way we interrogate archives in the digital environment will be very different from the paper world we have left behind. Much has gone. We defy anyone to read all the e-mails collected during the investigation into the Enron scandal publicized in October 2001. There are no longer files, and even where EDRMS systems have been implemented, they are largely empty.[81] As we have emphasized, the archive will on the whole accession stuff with little preprocessing. There will be no conventional catalogs, and users will need to learn how to use such machine-learning techniques to make sense of it.

The tools to interrogate large swaths of digital data are still being developed.[82] Take for example a long run of e-mails. Using sophisticated techniques will allow researchers to discard false trails, but then we will be able to map traffic to detect spikes that might give us clues as to topics that were a focus at a specific time.[83] Tools will be able to learn from our own discriminative decisions and may present some material to us in unusual ways to enable the machine to learn our needs; only later might the machine present the relevant material for us to directly comprehend the content.[84]

We will also be able to use graphs to map networks of who is talking to whom about what.[85] To do some of this we will need access to personal data that could be deemed to be in the public domain, such as the role or job an individual holds. We may also want to map our family trees, and again some data will be in the public domain, such as our birth, marriage, and death certificates, but others will not, such as the census.

The boundary between the public domain and "closed to access in the public interest" is fuzzy, changing, and contested. What we need to do as content providers is be aware and be ready to meet the challenges of continuing to defend the tradition of openness in digital environments. As we have argued, this inevitably requires us as records managers, information managers, and

[78] Bernstein, "Can an Algorithm Do the Job of a Historian?"
[79] Bountouri, *Archives in the Digital Age*.
[80] Prabhakaran, Reid, and Rambow, "Gender and Power."
[81] Currall et al., *"No Going Back";* Klimt and Yang, "The Enron Corpus."
[82] These will be much more sophisticated even than Google.
[83] Klimt and Yang, "The Enron Corpus."
[84] Berardi, Esuli, and Sebastiani, "A Utility-Theoretic Ranking Method for Semi-Automated Text Classification"; Berardi et al., "Semi-Automated Text Classification for Sensitivity Identification."
[85] IBM, "IBM I2 Analysts Notebook White Paper."

archivists to argue for a less risk-adverse attitude to the release of information. If this fails, we run the risk in the public sector of being overwhelmed by freedom of information requests for access against preemptively closed collections. If access is still refused, it will be for the courts to decide, but in the digital age even that is not straightforward, and the courts may also be much less supportive of risk avoidance. Lord David Neuberger, until recently president of the United Kingdom's Supreme Court, compellingly reminded us of this in a lecture on personal information in Singapore in 2015:

> During this talk, I have made much of the point that the far-reaching developments in IT require that steps are to [be] taken to ensure that the right to privacy is appropriately protected. However, we must also bear in mind the possibility, indeed the likelihood, that the relationship between developments in IT and fundamental rights is not a one-way street. It is, I suggest, inevitable that developments in technology that we are witnessing will change our attitude to privacy, and that is essentially for two reasons. First, one only has to consider the way that IT has changed the patterns and character of all aspects of our lives to appreciate that it is very likely to affect our values as well. Secondly, the existence of the Internet inevitably affects what can be practically achieved in terms of enforcement of privacy, and the law should never seek to acknowledge or enforce rights which are in practice unenforceable.[86]

In the United Kingdom, the Court of Queen's Bench, which hears cases against the executive, is already clogged with litigants seeking judicial review. In a post-truth world where even the rule of law is overshadowed by majoritism, this may not be enough to deter unscrupulous politicians from ignoring judicial decisions and ridiculing or sacking members of the judiciary.[87] The information professions must remain steadfast to their core values of preserving and safeguarding "evidence" that can be used unambiguously under the rule of law to call government and the executive to account across the public and private sectors. In the digital environment, this sacred duty, as Sir Hilary Jenkinson saw it, poses huge challenges that requires a sharing of this responsibility with technologists whose craft may be hard to comprehend.[88] There must be greater dialogue with users, wider society, lawyers, and politicians in deciding what should be kept and the terms of access.

Bibliography

AccessData. "Forensic Toolkit (FTK)." 2017. http://accessdata.com/solutions/digital-forensics/forensic-toolkit-ftk.

Allan, Sir Alex. "Records Review." August 2014. https://www.gov.uk/government/publications/records-review-by-sir-alex-allan.

Allison, Arthur, et al. "Digital Identity Matters." *Journal of the American Society for Information*

[86] Neuberger, "Is Nothing Secret?"
[87] Office of the United Nations High Commissioner for Human Rights, "SAHA—South African History Archive."
[88] Jenkinson, *Selected Writings*.

Science and Technology 56, no. 4 (2005): 364–72.

Article 29 Working Party—European Commission—Directorate General for Justice. "Opinions and Recommendations—Justice." European Commission, 2015. http://ec.europa.eu/justice/data-protection/article-29/documentation/opinion-recommendation/index_en.htm.

Auchard, Eric. "We're All Celebrities in Post-Privacy Age." Reuters UK, June 2007. http://uk.reuters.com/article/2007/06/22/celebrities-privacy-idUKNOA24969820070622.

Ball, Kirstie, and David Murakami Wood. "A Report on the Surveillance Society for the Information Commissioner." Surveillance Studies Network, Information Commissioner's Office, September 2006. https://ico.org.uk/media/about-the-ico/documents/1042391/surveillance-society-summary-06.pdf.

Berardi, Giacomo, et al. "Semi-Automated Text Classification for Sensitivity Identification." In *Proceedings of the 24th ACM International Conference on Information and Knowledge Management*, 1711–14. New York: Association for Computing Machinery, 2015. doi:10.1145/2806416.2806597.

Berardi, Giacomo, Andrea Esuli, and Fabrizio Sebastiani. "A Utility-Theoretic Ranking Method for Semi-Automated Text Classification." In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 961–70. New York: Association for Computing Machinery, 2012. doi:10.1145/2348283.2348411.

Bernstein, Joseph. "Can an Algorithm Do the Job of a Historian?" Buzzfeed, June 2015. https://www.buzzfeed.com/josephbernstein/can-a-computer-algorithm-do-the-job-of-a-historian?utm_term=.bcbPmQBeMz#.acE01OaEo4.

BitCurator Project. "BitCurator." 2011–16. http://www.bitcurator.net.

Börjeson, Love. "Social Network Methods: Topic Modeling and Correspondence Analysis Lab (Results Included)." 2016. https://rpubs.com/loveb/tm_ca_lab_results.

Bountouri, Lina. *Archives in the Digital Age: Standards, Policies and Tools*. Waltham, Mass.: Chandos, 2017.

Campaign against Arms Trade. "Al Yamamah Documents." 2007. https://www.caat.org.uk/resources/countries/saudi-arabia/al-yamamah.

Cocq, Céline. "DEMOSEC Day 2—Panel: The Intersection of Surveillance with Citizens Rights." In *Surveille Deliverable 5.5: Report of the Third Annual Forum for Decision Makers—Joint Final Event*, 40. European University Institute, November 2014. https://surveille.eui.eu/wp-content/uploads/sites/19/2015/06/D5-5-Report-of-the-Third-Annual-Forum-for-Decision-Makers.pdf.

Consultative Committee for Space Data Systems. "Space Data and Information Transfer Systems—Audit and Certification of Trustworthy Digital Repositories." International Organization for Standardization, 2012. http://www.iso.org/iso/catalogue_detail.htm?csnumber=56510.

———. "Space Data and Information Transfer Systems—Open Archival Information System (OAIS)—Reference Model." International Organization for Standardization, 2012. http://www.iso.org/iso/catalogue_detail.htm?csnumber=57284.

Cook, Chris. "The Government Departments Breaching Freedom of Information Law." *Newsnight*, January 2015. http://www.bbc.co.uk/news/uk-politics-31045687.

Currall, James, et al. "*No Going Back": The Final Report of Effective Records Management Project, University of Glasgow*. Glasgow: University of Glasgow, 2001. https://daedalus.lib.gla.ac.uk/retrieve/23/ERM-Final.pdf.

Curtis, Sophie. "Sir Tim Berners-Lee Calls for New Model for Privacy on the Web." *Telegraph*, October 2014. http://www.telegraph.co.uk/technology/internet/11148584/Tim-Berners-Lee-calls-for-new-model-for-privacy-on-the-web.html.

Der Derian, James. "Edward Snowden and Cyber-Zombies—A Host of New Surveillance Dangers." *Australian Book Review*, June–July 2014. https://www.australianbookreview.com.au/abr-online/archive/2014/117-june-july-2014/1995-edward-snowden-and-cyber-zombies.

Duranti, Luciana, ed. *Records Management Journal: Special Issue: Digital Diplomatics Records Management Journal* 25, no. 1 (2015).

Emerging Technology from the arXiv. "Machine-Learning Algorithm Can Show Whether State Secrets Are Properly Classified." *MIT Technology Review*, November 14, 2016. https://www.technologyreview.com/s/602848/machine-learning-algorithm-can-show-whether-state-secrets-are-properly-classified.

Featherstone, Mike. "Archive." *Theory, Culture and Society* 23, nos. 2–3 (2006): 593. doi:10.1177/0263276406023002106.

Federal Commissioner for the Stasi Records. "BStU—Homepage." http://www.bstu.bund.de/EN/Home/home_node.html.

Find and Connect Project. "Find and Connect Web Resource." 2011. http://www.findandconnect.gov.au.

Flynn, Paul. "What Brexit Should Have Taught Us about Voter Manipulation." *The Guardian*, April 2017. https://www.theguardian.com/commentisfree/2017/apr/17/brexit-voter-manipulation-eu-referendum-social-media.

Ford, Matt. "What the Special Counsel Appointment Means." *The Atlantic*, May 2017. https://www.theatlantic.com/politics/archive/2017/05/special-prosecutor-mueller-trump/527130/.

Foreign Office. Series: "General Correspondence from Political and Other Departments," FO Division 1, 1756. National Archives of the United Kingdom, Kew, London.

———. "Report on the Reorganization of Foreign Office Registries." FO366/787, 1918. National Archives of the United Kingdom, Kew, London.

"Freedom of Information Act 2000." http://www.legislation.gov.uk/ukpga/2000/36/contents.

"Freedom of Information (Scotland) Act 2002." http://www.legislation.gov.uk/asp/2002/13/contents.

Garnett, Fred. "The Heutagogy Archives—From Access to Content to Context." 2011–17. https://heutagogicarchive.wordpress.com.

Gibson, William. *Neuromancer*. New edition. London: Harper Voyager, 1995.

Gladwell, Malcolm. *Blink: The Power of Thinking without Thinking*. Reissue edition. London: Penguin, 2006.

Gollins, Tim. "The National Archive, Big Data and Security—Why Dusty Documents Really Matter." In *A Report on the RUSI/STFC Futures Event—Big Data for Security and Resilience: Challenges and Opportunities for the Next Generation of Policymakers*, edited by Jennifer Cole. London: Royal United Services Institute, 2014. https://www.rusi.org/publications/occasionalpapers/ref:O542BD3B7F097E.

———. "Parsimonious Preservation: Preventing Pointless Processes!" In *Online Information 2009*, 2009, 75–78. http://www.nationalarchives.gov.uk/documents/information-management/parsimonious-preservation.pdf.

———. "Putting Parsimonious Preservation into Practice." Presented at Online Information 2012, National Archives of the United Kingdom, Kew, London. http://www.nationalarchives.gov.uk/documents/information-management/parsimonious-preservation-in-practice.pdf.

Gollins, Timothy, et al. "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 (PIR 2014)*, edited by Grace Luo Si and Sicong Hui Yang, 39–40. CEUR-WS.Org, Vol-1225. Gold Coast, Australia, 2014. http://ceur-ws.org/Vol-1225/pir2014_submission_9.pdf.

Granger, Stewart. "Emulation as a Digital Preservation Strategy." *D-Lib Magazine* 6, no. 10

(October 2000). http://www.dlib.org/dlib/october00/granger/10granger.html.

Grigg, James. "Report/Committee on Departmental Records; Presented by the Chancellor of the Exchequer to Parliament by Command of Her Majesty, July 1954." Cmd 9163. London: Her Majesty's Stationery Office, 1954.

Hague, William. "The Foreign Secretary's Statement to Parliament on the Indian Operation at Sri Harmandir Sahib in 1984—Oral Statements to Parliament—GOV.UK." https://www.gov.uk/government/speeches/statement-on-the-indian-operation-at-sri-harmandir-sahib-in-1984.

Heick, Terry. "The Difference between Pedagogy, Andragogy, and Heutagogy." TeachThought. 2017. http://www.teachthought.com/pedagogy/a-primer-in-heutagogy-and-self-directed-learning.

Hillsborough Independent Panel. "Report of the Independent Panel." 2012. http://hillsborough.independent.gov.uk.

His Majesty's Stationery Office. "Notes for the Use Registry Branches." National Archives of the United Kingdom: T1/12334. London: His Majesty's Stationery Office, 1919.

Hoffman, Michael. "Society Struggles to Adapt to Post-Privacy Age." *Japan Times Online*, December 2013. http://www.japantimes.co.jp/news/2013/12/14/national/media-national/society-struggles-to-adapt-to-post-privacy-age/.

Home Affairs Committee. "A Surveillance Society?—Fifth Report of Session 2007–08." House of Commons, May 2008. http://www.publications.parliament.uk/pa/cm200708/cmselect/cmhaff/58/58i.pdf.

Huvila, Isto. "Participatory Archive: Towards Decentralised Curation, Radical User Orientation, and Broader Contextualisation of Records Management." *Archival Science* 8, no. 1 (2008): 32–33. doi:10.1007/s10502-008-9071-0.

IBM. "IBM I2 Analysts Notebook White Paper." January 2014. http://www-03.ibm.com/software/products/en/analysts-notebook.

IICSA. "Independent Inquiry into Child Sexual Abuse." 2017. https://www.iicsa.org.uk.

Information Commissioner's Office. "Conducting Privacy Impact Assessments Code of Practice." February 2014. https://www.igt.hscic.gov.uk/KnowledgeBaseNew/ICO_Privacy%20Impact%20Assessment%20Code%20of%20Practice.pdf.

———. "Information in the Public Domain—Freedom of Information Act and Environmental Information Regulations, Guidance." March 2013. https://ico.org.uk/media/for-organisations/documents/1204/information-in-the-public-domain-foi-eir-guidance.pdf.

Intelligence and Security Committee of Parliament. "Intelligence and Security Committee of Parliament—Privacy and Security: A Modern and Transparent Legal Framework." March 2015. http://isc.independent.gov.uk/news-archive/12march2015.

Jenkinson, Hilary. *Selected Writings of Sir Hilary Jenkinson*. Edited by Robert Ellis and Peter Walne. London: Alan Sutton with the Society of Archivists, 1980.

Jeong, Minwoo, Chin-Yew Lin, and Gary Geunbae Lee. "Semi-Supervised Speech Act Recognition in Emails and Forums." In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, 1250–59. Stroudsburg, Penn.: Association for Computational Linguistics, 2009. http://dblp.uni-trier.de/db/conf/emnlp/emnlp2009.html#JeongLL09; http://www.aclweb.org/anthology/D09-1130.

Klimt, Bryan, and Yiming Yang. "The Enron Corpus: A New Dataset for Email Classification Research." In *Machine Learning: ECML 2004*, edited by Jean-François Boulicaut et al., 3201:217–26. Berlin: Springer, 2004. doi:10.1007/978-3-540-30115-8_22.

Liu, Tie-Yan. "Learning to Rank for Information Retrieval." *Foundations and Trends in Information Retrieval* 3, no. 3 (2009): 225–331. doi:10.1561/1500000016.

McCarthy, Gavan, and Helen Morgan. "Rights and Commons: Navigating the Boundary between Public and Private Knowledge Spaces." In *Is Digital Different?* edited by Michael Moss and Barbara Endicott-Popovsky, 171–88. London: Facet, 2015.

McDonald, Graham, 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, ECIR 2017, Aberdeen, UK, April 8-13, 2017, Proceedings*, edited by Joemon M. Jose et al., 450–63. Cham (ZG), Switzerland: Springer International, 2017. doi:10.1007/978-3-319-56608-5_35.

McDonald, Graham, et al. "Towards a Classifier for Digital Sensitivity Review." In *Proceedings of the 36th European Conference on Information Retrieval*, 500–506. Cham (ZG), Switzerland: Springer International, 2014. http://eprints.gla.ac.uk/93566/1/93566.pdf.

Merrin, William. *Media Studies 2.0*. Abingdon, U.K.: Routledge, 2014.

Moss, Michael. "The Hutton Inquiry, the President of Nigeria and What the Butler Hoped to See." *English Historical Review* 120, no. 487 (2005): 577–92. doi:10.1093/ehr.

———. "Where Have All the Files Gone? Lost in Action Points Every One?" *Journal of Contemporary History* 47, no. 4 (2012): 860–75. doi:10.1177/0022009412451291.

National Archives of the United Kingdom. "Archives Inspire: The National Archives Plans and

Priorities 2015-19." 2015. http://www.nationalarchives.gov.uk/documents/archives-inspire-2015-19.pdf.

———. "The National Archives—Our Role—Digital Preservation." N.d. http://www.nationalarchives.gov.uk/information-management/manage-information/preserving-digital-records/our-role.

———. "Takedown and Reclosure Policy: The National Archives." 2015. http://www.nationalarchives.gov.uk/legal/takedown-policy.htm.

Neuberger, D. "Is Nothing Secret? Confidentiality, Privacy, Freedom of Information and Whistleblowing in the Internet Age." Singapore Academy of Law Annual Lecture, United Kingdom Supreme Court, 2015. https://www.supremecourt.uk/docs/speech-150921.pdf.

Office of the United Nations High Commissioner for Human Rights. "SAHA—South African History Archive—UNHRC Report on Archives and the Right to the Truth." February 2011. http://www.saha.org.za/news/2011/May/unhrc_report_on_archives_and_the_right_to_the_truth.htm.

Orwell, George. *Nineteen Eighty-Four*. London: Penguin, 2013.

Pacioli, Luca. "De computis et scripturis." In *Summa de arithmetica, geometria, proportioni et proportionalita*. Venice: Paganini, 1494.

Page, Lawrence, et al. "The PageRank Citation Ranking: Bringing Order to the Web." Technical Report, Stanford InfoLab. 1999. http://ilpubs.stanford.edu:8090/422.

Prabhakaran, Vinodkumar, 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*, 1965–76. Stroudsburg, Penn.: Association for Computational Linguistics, 2014.

Public Record Office. "Requirements for Electronic Records Management Systems." National Archives of the United Kingdom, 2002. http://www.nationalarchives.gov.uk/documents/requirementsfinal.pdf.

Redwine, Gabriela, et al. *Born Digital: Guidance for Donors, Dealers, and Archival Repositories*. Washington, D.C.: Council on Library Information Resources, 2013. http://www.clir.org/pubs/reports/pub159/pub159.pdf.

Rosenthal, David. "Formats through Time." DSHR's Blog, October 2012. http://blog.dshr.org/2012/10/formats-through-time.html.

Rosenwald, Michael S. "James Comey's Memo Has Shaken a Presidency. Here's Why Memos Have Always Mattered." *Washington Post*, May 2017.

https://www.washingtonpost.com/news/retropolis/wp/2017/05/17/james-comeys-memo-has-shaken-a-presidency-heres-why-memos-have-always-mattered.

Rothenberg, Jeff. "Ensuring the Longevity of Digital Documents." *Scientific American,* January 1995, 42–47. http://www.scientificamerican.com/article/ensuring-the-longevity-of-digital-d.

Rumsey, Abby S., ed. "Sustainable Economics for a Digital Planet: Ensuring Long-Term Access to Digital Information." Blue Ribbon Task Force on Sustainable Digital Preservation, February 2010. http://brtf.sdsc.edu/biblio/BRTF_Final_Report.pdf.

Rusbridge, Chris. "Excuse Me . . . Some Digital Preservation Fallacies?" *Ariadne: Web Magazine for Information Professionals* 64 (2006). http://www.ariadne.ac.uk/issue46/rusbridge.

Salton, G., and C. Buckley. "Term-Weighting Approaches in Automatic Text Retrieval." *Information Processing and Management* 24, no. 5 (1988): 513–23. http://comminfo.rutgers.edu/~muresan/IR/Docs/Articles/ipmSalton1988.pdf.

Scholer, Falk, et al. "The Effect of Threshold Priming and Need for Cognition on Relevance Calibration and Assessment." In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 623–32. New York: Association for Computing Machinery, 2013. doi:10.1145/2484028.2484090.

Snowden, Edward. "Here's How We Take Back the Internet." TED Talk, March 2014. https://www.ted.com/talks/edward_snowden_here_s_how_we_take_back_the_internet.

Souza, Renato Rocha, et al. "Using Artificial Intelligence to Identify State Secrets." *Computing Research Repository (CoRR) Connell University Library,* arXiv:1611.00356 (2016). http://arxiv.org/abs/1611.00356; http://dblp.uni-trier.de/rec/bib/journals/corr/SouzaCSC16.

Underwood, William. "Recognizing Speech Acts in Presidential E-Records." Georgia Tech Research Institute; Georgia Institute of Technology; Computer Science; Information Technology Division Information Technology; Telecommunications Laboratory, 2008. http://www.archives.gov.edgesuite-staging.net/applied-research/gtri/tr_08_03_recognizing_speech_acts.pdf.

———. "Speech Acts and Electronic Records." In *Proceedings of DigCCurr2009 Digital Curation: Practice, Promise and Prospects*, 135–42. Chapel Hill: School of Information and Library Science, University of North Carolina at Chapel Hill, 2009. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.164.54&rep=rep1&type=pdf#page=147.

Underwood, William, Sheila Isbell, and Matthew Underwood. "Grammatical Induction and Recognition of the Documentary Form of Records." In *DigCCurr2007*. Chapel Hill: School of Information and Library Science, University of North Carolina at Chapel Hill, 2007. http://www.ils.unc.edu/digccurr2007/papers/underwood_paper_4-5.pdf.

Verbeek, Peter-Paul. *What Things Do: Philosophical Reflections on Technology, Agency, and Design*. Translated by Robert Crease. University Park: Pennsylvania State University Press, 2005.

Waugh, Andrew. "Email—A Bellwether Records System." Recordkeeping Roundtable, June 2014. http://rkroundtable.org/2014/06/30/email-a-bellwether-records-system.

Weiner, Gaby. "It's Vital We Have Access to the Records on Britain's Colonial Past." *The Guardian*, January 2014. http://www.theguardian.com/uk-news/2014/jan/22/vital-access-records-britains-colonial-past.

Wilson, Duncan. "Modern Public Records: Selection and Access: Report of a Committee Appointed by the Lord Chancellor/Chairman Sir Duncan Wilson; Presented to Parliament by the Lord High Chancellor." Cmd. 8204. London: Her Majesty's Stationery Office, 1981.