After the Digital Revolution: Working with Emails and Born-Digital Records in Literary and Publishers' Archives

> *'there lie in his hoards many records that few now can read, even of the lore-masters, for their scripts and tongues have become dark to later men'.*
> J.R.R. Tolkien, *The Fellowship of the Ring*

While we still have letters, manuscripts and other physical documents from the past centuries, we are in danger of losing digital documents created in the last decade. Literary scholars rely on the traces left by writers – from correspondence to drafts – which now take the form of born-digital records. Publishing historians also need access to the records left by publishing companies. Emails and other digital forms of communication have largely replaced letters and memos, and yet, safeguarding digital archives remains an enduring challenge for archivists. Electronic records risk becoming unreadable due to rapidly changing formats and technologies. Even when digital archives are actively preserved, they are often closed to researchers due to data protection and other issues. To paraphrase Tolkien, the scripts and tongues of our digital age risk becoming dark to later men. As late as 2010, a report from the American library community OCLC declared: 'Management of born-digital archival materials is still in its infancy'.[1] What progress has been made to preserve digital archives? How can we improve access to born-digital collections? And how can scholars produce new research using emails and other born-digital records?

This special issue of *Archives and Manuscripts* finds its origins in 2017-18, when I was awarded a British Academy Rising Star Engagement Award for my project 'After the Digital Revolution: Bringing together archivists and scholars to preserve born-digital records and

produce new knowledge' (#AfterDigRev). The project has sought to answer the problem of collaboration on both sides of the reading room. Indeed, archivists and literary scholars rarely 'sit at the same table', and this lack of dialogue has an impact on issues of access, particularly in the case of born-digital materials. For example, the archive of the poetry publisher Carcanet in Manchester contains hundreds of thousands of emails, but it is currently treated as a dark archive closed to researchers.

There are three main reasons for closing email archives: privacy concerns, copyright and technical issues. British and European archives need to comply with the new General Data Protection Regulation. Repositories often prefer to restrict access, instead of sharing potentially confidential and sensitive information. The US has a different approach to privacy. American archivists will often have you sign some paperwork to make sure that if you find anything sensitive, you will refrain from publishing it without permission. Researchers are treated as responsible adults. And yet, getting access to email archives in the US is still not easy. Even when an institution wants to share digital files, it cannot put everything online for copyright reasons. Researchers still need to travel to the archival repository to consult documents. And few institutions have solved all the technical issues specific to born-digital archives, including designing an appropriate interface to make these documents available to researchers.

What do we mean exactly by 'born-digital'? In 2010, OCLC listed various kinds of born-digital resources, broadly defined as 'items created and managed in digital form'.[2] This includes digital photographs; digital documents such as PDFs; harvested web content; the digital manuscripts of noteworthy individuals; the electronic records of institutions; static data sets generated by researchers; dynamic data such as Facebook and Twitter accounts; digital art; and digital media publications – music and movies, for example. In this editorial, I

will focus mostly on materials immediately relevant to literary scholars: web content, including the websites and blogs of writers; and digital manuscripts (or personal digital archives), defined broadly to include not only working files but also email accounts.

How can archivists and scholars collaborate to create a better future for digital collections? The first section presents the 'After the Digital Revolution' project and the discussions at the two workshops in Manchester (September 2017) and London (January 2018). Workshop 1 looked specifically at the problems of preservation and access. Workshop 2 focused on producing new knowledge based on archival emails and other born-digital documents. But this essay goes well beyond summarising a project and its findings. More generally, it emphasises the transformative impact of born-digital records on literary scholarship, and particularly on graduate training in research methods. We continue to train graduate students to do 'traditional' archival work in paper archives. With the shift to digital, we need to better prepare our students to interrogate born-digital data using a wide range of methods – including Artificial Intelligence.

The second section gives an historical overview of the debates over born-digital documents. In the mid-1990s, the archival community started devising strategies to preserve endangered materials. The visibility and impact of these initiatives increased in the early 2000s, through a wave of multi-institutional projects funded by public and private organisations in the US and the UK. In the late 2000s, collaborations between archivists and scholars resulted in the creation of open-source digital library tools for content curation such as BitCurator. Despite these pioneering efforts, libraries are still falling dangerously behind in acquiring and processing information received in digital form.[3]

Within this wider context, the special issue of *Archives and Manuscripts* brings together archivists and scholars to discuss how born-digital materials are preserved, accessed

and used today in major archival repositories in Britain, North America, Australia and New Zealand. In the conclusion of this editorial, I give a summary of the articles included here and I suggest three directions to move the debate forward.

I) 'After the Digital Revolution'

Origins of the project

My first encounter with the field of Digital Humanities came during my doctoral years at the University of British Columbia in Vancouver (2009 to 2013). As a book historian and modernist scholar, I became involved in the Modernist Versions Project at the University of Victoria, contributing to the TEI mark up of Joseph Conrad's *Nostromo*. But my main allegiance was to book history, not DH. I was much more interested in the glamour of archival work – going to far-away places, spending weeks shifting through old documents. I understood, of course, that such work was not an option for many scholars who lacked adequate funding and time. I admired the digitalisation projects that made archives more accessible to those who could not travel. Yet, I also suspected that the partial digitisation of collections did not address a fundamental aspect of archival work: the discoveries that result from *unrestricted* access to a mass of documents. For example, when I was working in the Random House archive at Columbia University, I discovered a series of letters exchanged between William Faulkner and a young would-be writer. These letters were in the uncatalogued part of the collection, which is stored off-site. If Columbia decided to digitise the Random House records, they would certainly start with the catalogued part – not with the uncatalogued boxes full of rejection letters and other documents of little scholarly value (or so it seems).

To make serendipitous discoveries, unrestricted access is key. Let's take the example of the Carcanet Press archive at the Rylands Library in Manchester. When I was working as

Research Associate at the John Rylands Research Institute, I came across the Carcanet Press E-mail Preservation Project (2012–2014), made possible by a grant from the UK funder Jisc. Fran Baker, the archivist in charge of the project, made the sensible choice to focus on the preservation of these emails. Born-digital records are extremely fragile. Leave old letters unattended for decades, and there is a good chance that they will still exist for the next generations to read. But the same is not true of digital correspondence: commercial providers can close down, leading to the deletion of millions of emails; files downloaded on personal computers can become unreadable over time; external storage can become obsolete. The Carcanet Project resulted in the successful rescue and preservation of 215,000 e-mails and 65,000 attachments generated by Carcanet Press.[4] The project's work was recognised in 2014, when it received the Digital Preservation Coalition's prestigious biennial award for 'Safeguarding the Digital Legacy'.

Preserving email collections is one thing, making them accessible is another. The Carcanet email collection is still being treated as a dark archive, embargoed to researchers and with access restricted to a small number of staff only. This was the starting point of the 'After the Digital Revolution' project: bringing together archivists and scholars to find solutions to the problem of access. The next part looks at specific case studies of collaboration, starting with the Harry Ransom Center in Texas.

Preserving born-digital collections and making them accessible

In Summer 2017, I was the first scholar to access archival emails in the Ian McEwan collection in Texas. A couple of years ago, Stephen Enniss, the Director of the Harry Ransom Center, travelled to McEwan's home in London to discuss the acquisition of the collection. Enniss asked McEwan if he would include his email correspondence with Salman Rushdie,

which would make the archive extremely valuable to researchers. McEwan agreed and when he sold his archive to the Harry Ransom Center, he included seventeen years of emails, from 1997 to 2014.

McEwan's email archive is an extremely useful resource. In the early 1970s, the young McEwan drew a lot of encouragement from established figures of the literary world on both sides of the Atlantic: Malcolm Bradbury, Angus Wilson, Philip Roth and Ted Solotaroff. As the editor of the *New American Review*, Solotaroff published several of his short stories between 1972 and 1975. More than three decades later, Solotaroff wrote him an email to congratulate him on his recent novels. McEwan replied: 'I glowed in the face of your praise, and the experience rather took me back to the early 70s when a letter (ah those were the days) from you appeared to me in my little flat in Norwich to have been sent from an Olympian realm'.[5]

What difference does it make for a researcher to work with emails rather than letters? Emails share much more with previous forms of communication than we usually imagine. Email-writing was initially modelled on letter-writing after all; business memos were another source of inspiration, as the subject line reminds us. Publishers' archives are full of letters that mix private and professional matters. Even the speed of emails is not a radical innovation, as letters were delivered several times a day in the early twentieth century. The main difference is the materiality of letters. As Jacques Derrida and others have noted, we continue to fetishise physical documents in archives. The handwriting and signature of the writer carry an aura that emails lack. But there are other ways to inspire this sense in readers – such as emulation. Everybody will soon be able to access McEwan's emails – even though a trip to Texas will still be necessary.[6]

During the first #AfterDigRev workshop, Abby Adams, Digital Archivist at the Harry Ransom Center, and Rebecca Roach from the Department of English at King's College London,

presented a joint paper on J. M. Coetzee. Before he embarked on a career as a scholar and writer, the South African-born writer was a computer programmer in the early years of the industry's development. Coetzee worked on Atlas 2, one of the most advanced programming projects in Britain in the mid-1960s. Few scholars have paid attention to Coetzee's sustained interest in computing throughout his career. This is about to change, Adams and Roach hope, thanks to Coetzee's digital archives being made available to scholars. The printed materials in the Coetzee archive at the Harry Ransom Center are already regularly consulted by researchers. His born-digital materials – including a computer tape reel, over 100 floppy disks, and email correspondence – were opened for research in 2017. In their talk, Adams and Roach discussed the process and decisions entailed in making Coetzee's born-digital materials available, offering the perspective of both the archivist and the researcher. Adams outlined her workflow for data recovery, preservation and description, and the discovery and access methods she employed. Roach mentioned the implications of these decisions for her use of the collection and attempts to document Coetzee's 'other career'. This born-digital archive sheds light on sixty years of digital innovation, from the perspective of a computer programmer who became a major writer.

These examples of born-digital collections at the Harry Ransom Center offer three valuable lessons to solve the problem of preservation and access. First, contemporary archives – whether paper or digital – will always carry some risk regarding data protection. When I looked at Ian McEwan's emails, I discovered some confidential and sensitive information. Most researchers will follow an unwritten ethical code on what qualifies as 'good' research data, and will refrain from using problematic materials. Archival repositories can also ask users to sign legal agreements prior to releasing potentially risky born-digital documents. In short, it is always preferable to release data (or at least a selection) than to

close entire collections in the mistaken hope that, one day, all issues will have been sorted out.

The second lesson is that the technical infrastructure does not need to be perfect. For the McEwan archive, I was not sure if the emails I was reading on the screen were originally in the Inbox, in the Sent folder or in another Folder. It was often difficult to understand the context. A typical experience was to see McEwan's response to a query, and then, after clicking through dozens of other emails, to find the original question. However, these small technical problems did not prevent me from finding relevant data for my research.

The third lesson is that archival repositories should actively involve researchers at the early stage of opening born-digital collections. Stephen Enniss and his colleagues encouraged me to give feedback on my experience using McEwan's archival emails. Similarly, Rebecca Roach was invited to write a blog post about Coetzee's born-digital collection.[7] This empowers researchers to play an active role in shaping access to these collections – rather than passively waiting for archivists to provide finding aids and design access policies. In other words, researchers can be co-creators of archives, instead of passive users.

Like the Harry Ransom Center, the British Library has actively sought to involve researchers through user testing of born-digital records. At the first #AfterDigRev workshop, Eleanor Dickens and Rachel Foss explained that since 2014, over 60% of acquisitions to the Contemporary British Manuscript department have been hybrid collections. In most cases, the digital objects in these collections already outnumber the paper-based material, a trend expected to continue.[8] In October 2016, the British Library launched a pilot project to develop a workflow for processing born-digital archives and making them accessible to researchers in British Library reading rooms. The pilot drew on three hybrid personal archives with significant born-digital content: the Carmen Callil Archive, the Hanif Kureishi Archive and the

Ronald Harwood Archive. The first step was to develop a process of automatically harvesting metadata from digital objects and uploading this to archival software with minimal intervention. This process enabled archivists to catalogue born-digital archives in a matter of weeks or even days. Once the metadata was extracted, this was then published to the British Library (BL) Explore Archives and Manuscripts Catalogue. Extracted captures were migrated as PDF/A to create access copies. The next step was to provide access to born-digital archives in the BL Manuscripts reading room for the first time using a FTP-server to host PDF/A surrogates of the migrated digital objects. In Spring 2017, the British Library held user-testing workshops involving students, early-career researchers, academics and library/archive professionals. Participants were asked to respond to a questionnaire with practice exercises to gauge how easily they could navigate the collections. Roundtable discussions were held to explore researchers' awareness of, engagement with, and expectations regarding born-digital archive collections.

The British Library's efforts to test delivery to users is laudable, but researchers should be aware that not all born-digital documents are accessible. As Eleanor Dickens and Jonathan Pledge put it in their article on the Wendy Cope archive:

> Our access model uses a resource that was originally developed by the BL's Endangered
>
> Archives Programme (EAP). EAP requires a platform to deliver this content to researchers whilst abiding
>
> by standard BL practices in relation to copyright and data protection for the delivery of archives and
>
> manuscripts. In this case they use an FTP server with access restricted solely to researchers using the BL
>
> Asian and African Studies Reading Room. Owing to the *relatively small amount of born-digital material*
>
> *we were able to make available*, we decided to use this model rather than provide access through our
>
> central Library repository.[9]

There are two problems here. The first is the lack of accountability on the application of data protection laws. PDF/A access copies are checked for data protection clearance, and many documents are considered sensitive and closed for access. Like other archival repositories in the UK, the British Library has a strict understanding of the General Data Protection Regulation. The GDPR puts a lot of emphasis on the documentation that needs to be kept and on the right of individuals to access, rectify and erase their data, a right limited by the 'archiving in the public interest' principle. But we need to know more about what is being archived, and when these documents will be made available. In other words, we need more transparency on decisions to preserve or discard born-digital materials, and to provide access or to close collections.

The second problem is that due to copyright issues, born-digital materials are only available in the British Library reading rooms. The situation is similar in the United States. For example, researchers need to fly to Atlanta and drive to Emory University to consult Salman Rushdie's digital records.[10] As Stephen Enniss puts it, 'files that were once transmitted instantly across great distances are now bound to a reading room desk like some medieval scriptorium'.[11] This restricts access to a minority of researchers who have the time and funding to travel to archival repositories. Stanford University has tried to address this issue with ePADD, a platform to manage the entire cycle of archival emails – from donors to researchers.[12]

The Bibliothèque nationale de France (BNF) also requires researchers to travel onsite to consult web archives. The BNF has been in charge of collecting and preserving the French Internet since 2006, when the DADVSI law on copyright was voted.[13] The robot Heritrix – developed by the Internet Archive – harvests representative samples of websites, following a selection made by librarians and external partners. For example, the BNF Art and Literature

Department has selected more than 5,000 sites about French literature, foreign languages and literatures, art and book history. In addition to websites and blogs, the BNF will soon collect the Twitter accounts of writers, editors and literary periodicals. Pages, images, sounds and videos are collected and stored in digital storerooms. To explore the archives, researchers can type the website name or use the guidance focusing on specific topics. Featured collections include 'Writing oneself on the Web: personal and literary diaries', which examines the transition from paper to digital, and the way blogs have changed personal, literary and critical writing. [14]

Since 2008, researchers have access to the Web archives in the BNF reading rooms or in the regional libraries in charge of Legal Deposit in France. Being able and willing to travel to these libraries is not enough. Users must also justify their need to access these archives for academic, professional or personal research activities. At the BNF, readers' cards are issued after an individual admission interview. [15] It is of course paradoxical that the data collected is freely available on the web, and yet, consultation of the archive is restricted and can only take place within specific libraries. The situation in France is similar to English-speaking countries: restrictive access policies are meant to address copyright and data protection issues, following the recommendations of the CNIL (National Commission on Informatics and Liberty).

How can we make born-digitals more accessible to researchers who are unable to travel to archival repositories? One solution would be to design a protected online environment, available to users after identity verification. Special Collections libraries already ask researchers to provide their ID as part of the registration process. They could do the same online, and issue a password to researchers to access a delivery interface. Users would then be able to see unredacted email correspondence and other born-digital records. This system

would give less control to archival repositories over copyright, and it could potentially lead to leaks if materials are copied and circulated without permission. But this is not a new issue. At the British Library, researchers were prevented from taking their own photographs of materials until 2015, when this restriction was lifted for selected collections. Self-service photography is purely for reference purposes and researchers are asked to be mindful about publishing, sharing or uploading photos as this could breach copyright, data protection or privacy laws. Archival repositories need to share control and to balance risk with greater access and convenience for users.

Making collections easier to find: the question of open data

To fully understand the issue of born-digital records locked in dark archives closed to researchers, or available only to those who can travel to repositories, we need to move away from the specific case of literary collections and literary scholarship. The problem of born-digital literary records is a small aspect of a much broader question, the question of open data. Most scholars will be familiar with the debates over open access: should we continue to lock our articles behind paywalls? Or is it time to embrace open access and make our research more accessible? Many research funders now require free, online access to research outputs. For example, the Wellcome Trust supports 'unrestricted access to the published output of research as a fundamental part of its charitable mission and a public benefit to be encouraged wherever possible'.[16]

Archival repositories are not, or should not be isolated from the open access movement. In his keynote speech at the second #AfterDigRev workshop, David McKnight – the Director of Special Collections at the University of Pennsylvania – made the case for an Open Access, Linked Open Data (LOD) online line repository of publishers' records that will

preserve the records, enrich the data through LOD, and provide opportunities for the creation of new knowledge. Linked data is a process for embedding archival finding aids and other information into the very fabric of the web. It allows computers to read and use archival descriptions, making the information easier to discover. Linked data also empowers archivists to use and incorporate the information of other providers into their local description. For example, images of people and additional biographical details can automatically be added to descriptions to make collections come alive.[17]

Fostering linked open data would also improve ArchiveGrid, the closest thing the US has to a national archival discovery tool. Speaking at Workshop 2, Amy Chen argued that discovery tools often suffer from poor UX (User Experience). This in turn requires more support by archivists and librarians if users ask for help. 'Those of us who work within the archive and library sector may say that we are service-oriented or patron-driven', Chen said, 'but in practice what that usually means is that we value the accuracy and timeliness of our answers and the warmth of our interactions, not how users engage with us online'. Linked open data would make archival descriptions more complete, and information easier to find. Extensive user testing is also necessary to improve the discoverability and usability of archival data.

The notion of 'archiving in the public interest' embedded in the UK General Data Protection Regulation cannot be separated from the questions of open data and discoverability of this data. Preserving materials in the public interest is not enough. What is the point of preserving data if users have no way to discover and access this data? It seems surprising that publicly-funded libraries and archives can hold vast amounts of born-digital records and keep them hidden in dark archives. Even if materials are deemed too sensitive or confidential to be communicated, researchers should at least know when closed archives will

be open. This is, after all, not a new problem: think of military and other sensitive information contained in government archives. The UK National Archives already recognise that archives are unlikely to be closed for more than 100 years, and permanent closure periods will not be accepted.[18] 'Archiving in the public interest' should therefore go hand and hand with transparency and whenever possible, open access.

Producing new knowledge

Improving preservation and accessibility of born-digital records will allow researchers to produce innovative work, with the potential to transform the Humanities into data-rich, impactful disciplines. Take the example of email, a primary source of historical change. As James Baker showed during Workshop 2, email archives tell us a lot about contemporary history, a history now starting to come under scrutiny. To study the 1990s, scholars will need histories of email, methods for working with email, and reflections of whether email-as-archive creates tensions with archival practice and traditions. Baker's talk focused on historical sources of how people encountered and negotiated this form of communication. He gave the example of the Enron email archive. In October 2003, Andrew McCallum, a computer scientist, read that the federal government had more than five million messages from the prosecution of Enron. He paid $10,000 for a copy of the database and made it freely available to researchers within and outside academia. Since then, it has been used extensively – in part because it is difficult to access other large collections of emails with potentially confidential information. 'It's made a massive difference in the research community', McCallum said.[19]

At first sight, the content of the Enron archive seems far from the preoccupations of literary scholars. But the methods used to analyse this collection can be transferred to more

literary archives, especially archives that contain hundreds of thousands of emails. With such a mass of data, keyword search does not work well, as lawyers who work in the field of e-discovery have long known. To find evidence of an executive expunging incriminating evidence, it is not enough to type 'delete' or 'erase', because crooks often use code words to disguise their actions. In the case of Enron, these code words were often Star War references. 'What reasonable attorney would have thought to use 'Millennium Falcon' or 'Chewbacca' in a keyword search of an energy company's transactions?', asks J. R. Jenkins in an article on the rise of analytics in e-discovery.[20]

To solve this problem, lawyers use data visualisation and analytics software. Visualisations present data graphically, helping attorneys determine what is or is not relevant to the case. Analytics tools can identify words used by dishonest executives to cover their traces – for example, 'obliterate' instead of 'delete' or 'erase'. Predictive coding is another useful tool to identify relevant documents, particularly in cases with very large data sets. The tool relies on two learning methods: supervised learning and active (unsupervised) learning. With *supervised learning*, a lawyer chooses a subset of documents, and this selection enables the analytics system to rank the remaining documents in the collection based on their similarity with the initial subset. In the case of *unsupervised learning*, the machine selects a subset of all case documents using sophisticated algorithms. The attorney reviews this subset to determine its relevancy, and submits it to the system. The machine then analyses the selected documents to identify and code key trends or patterns, before turning to the rest of the collection.

These tools could be applied to email archives that present confidentiality issues. For example, the team working on the Carcanet Press E-mail Preservation Project used data visualisations. 'In light of access restrictions on the full content of e-mail archives', Fran Baker

wrote, 'these high-level representations, based on metadata, could form the basis of fruitful research in their own right: for tracing literary networks, the evolution of literary movements, or behavioral trends and changes in the way people communicate'.[21] But as with predictive coding, visualisations rely on the availability of data sets.

It is not enough to have access to a few emails pre-selected by archivists, and to analyse them manually as we would do with letters. Close reading is not an efficient method to make sense of vast quantities of data. Paradoxically, it was developed by New Critics in the 1940s and 1950s, at a time when archivists were already complaining about the deluge of paper they had to process and make available. Confronted with a world of big (paper) data, the New Critics zoomed in on a few canonical texts, analysing them through close reading. They repeatedly criticised historical research and turned away from the archive. It was not until the 1980s that New Historicism opened the door for the archival turn that has since shaped literary studies. Yet, the profession has been largely reticent to move away from qualitative research methods to embrace mixed approaches common in social sciences.[22] As literary scholars, we need to learn to engage more actively with big data – paper and digital. But of course, to use archival data, we need to have access to well-preserved records.

II) Born-Digital Records: A Short History

The debates over the preservation of emails and other born-digital materials started in the mid-1990s, following an important legal case. In August 1993, the US Court of Appeals for the District of Columbia Circuit affirmed a decision that federal agencies, including the White House and Congress, must retain all official emails that exist within their computer systems.[23] It was not enough to print certain messages out to paper, the court said. Hard copies often lacked contextualising information, such as the sender, recipient and time of

transmission. Electronic versions therefore had to be retained to satisfy recordkeeping requirements.

For the archival community, this represented a formidable challenge. In a 1994 article on 'Managing Electronic Mail', David Bearman wrote: 'The question of how to manage electronic mail as a record is one that will confront management in every contemporary organisation within the next few years'.[24] The issue was not only to develop technical solutions, but also to adapt mindsets to the new digital world. 'We have paper minds trying to cope with electronic realities', Terry Cook argued.[25] The role of the archivist was to preserve documents and to help users find the correct information. If we are unable to do that, Cook said, 'we will be replaced by software packages that can handle facts, and data, and information very efficiently, without any mediation by archivists or anyone else'.[26] Fast-forward twenty-five years or so, and Cook's words seem particularly prescient. For many records, Google and other search engines play the role that archivists used to play: sorting out information, and guiding users to what they are looking for.[27] Yet, the Google era has not led to the disappearance of the archival community. Instead, archivists have become information studies experts, who seek solutions to complex interdisciplinary problems.

Growing awareness of the need to preserve born-digital records also led to the creation of the first web archives. In 1996, the Internet Archive, a US-based nonprofit foundation, started broadly archiving the web, and it has remained one of the most ambitious initiatives. In 2018, it held 279 billion archived web pages, as well as audio recordings, videos, images, software programs and other records. The Pandora and Tasmanian web archives from Australia, and the Kulturarw3 web archive from Sweden, were also created in 1996. Whereas the Internet Archive aims to preserve the web from all over the world, other initiatives often focus on preserving the parts most relevant from their own perspectives. For

example, Kulturarw3 harvests the Swedish web, including servers located under the top-level domain 'se'. Web archives do not contain sites prior to 1996 except for some pages recovered from backups stored in floppy disks or CDs.[28]

In the late 1990s, the British Library commissioned a report on the rescue of digital materials. The resulting study, by Seamus Ross and Ann Gow, examined the approaches to accessing digital materials in three target areas: 1. the media had become damaged (through disaster or age); 2. the media contained material in an unknown format; 3. the hardware or software was obsolete. In particular, the report examined the role of emulation of both hardware and software to access digital records. Ross and Gow's report exemplifies a slow (and still incomplete) transition from preservation to access.[29]

At the turn of the twenty-first century, it became clear that not enough was being done to safeguard endangered digital records, let alone to make them available to researchers. The immensity of the task required collaborations across various institutions, at the national and international level. In 2002, the Digital Preservation Coalition (DPC) was established as a partnership between several agencies operating in the UK and Ireland. It has since become a global organisation for digital archivists and other people involved in digital preservation. From 2005 to 2007, the UK funder Jisc supported the PARADIGM (Personal Archives Accessible in Digital Media) project, undertaken by the Bodleian Library in Oxford and the John Rylands Library in Manchester. The overall aim was to examine the issues in preserving personal digital materials, and to produce best-practice guidelines.[30]

In 2007, the Arts and Humanities Research Council (AHRC) funded the two-year Digital Lives project, led by the British Library in partnership with University College London and the University of Bristol. The primary aim of the project was to develop ways to secure the personal archives of individuals in the digital era, in order to enable sustained access. With

this move from preservation to access, curators had to address confidentiality and data protection requirements, copyright issues, and the authenticity and provenance of all files. The 261-page report published in 2009 stressed the need to work with a wide range of actors – donors, policy makers and users – to preserve and make accessible personal digital archives. Yet, due to limited resources, the priority was still to work upstream rather than downstream, i.e. to secure the materials and retain them first, before liaising with users and researchers.[31]

In the late 2000s, the number of web archiving initiatives was growing rapidly all over the world.[32] Three organisations – Jisc, the Digital Preservation Coalition and the now-defunct UK Web Archiving Consortium – organised the conference 'Missing links: the enduring web' at the British Library in 2009. 'I came away feeling confident that web archiving has finally "arrived"!', Cathy Smith of the UK National Archives said after the event.[33] The report that Smith presented at the conference focused on delivering UK web archives to user communities. What will the web be like as an historical source, and what use will be made of archived web sites by future generations?, Smith asked.[34] Historians in the room were asking the same question. For Jane Winters of the School of Advanced Study, the event was the starting point of a series of collaborations between historians and archivists. 'It was hard to demonstrate the value of web archiving, and to justify the resources devoted to it, if historians and others were unaware of and apparently unwilling to use web archives in their research', Winters later said.[35]

In the US, the Library of Congress started its Web Archiving program in 2000 and first offered personal digital archiving guidance on its website in 2007. It has continued to provide advice on The Signal, a blog devoted to digital preservation, and in various reports and online resources. It also acquired the Twitter archive – a clear signal that social media was worth preserving. American funding agencies, both private and public, started supporting major

initiatives focused on born-digital records. In 2008, the Andrew Mellon foundation funded the futureArch project at the Bodleian Library to find solutions to the problem of born-digital but also hybrid archives (composed partly of paper materials). In particular, Bodleian Electronic Archives and Manuscripts (BEAM) worked on digital preservation infrastructure and researcher interfaces for hybrid archives.

That same year, the National Endowment for the Humanities (NEH) funded an interdisciplinary project conducted by Matthew Kirschenbaum and others. The resulting White Paper – *Approaches to Managing and Collecting Born-Digital Literary Materials for Scholarly Use* – explored the challenges faced by archivists, administrators, and scholars at three institutions whose collections include personal digital archives: Emory University in Atlanta, Georgia; the Harry Ransom Center at the University of Texas, Austin; and the University of Maryland College Park. As a book historian, Kirschenbaum was acutely aware of the need to preserve not only the text but also the material context of production. 'Literary scholars are going to need to play a role in decisions about what kind of data survives and in what form', the report declared, 'much as bibliographers and editors have long been advocates in traditional libraries settings, where they have opposed policies that tamper with bindings, dust jackets, and other important kinds of material evidence'.[36] The re-creation of the writer's working environment through emulation would provide valuable information for researchers. Once born-digital materials had been made available, the next step was to produce new knowledge: 'here we may see textual scholarship begin to draw heavily on text mining and visualisation, methods which are specifically aimed at sorting and sifting large volumes of data'.[37] In short, Kirschenbaum and his colleagues proposed to adapt research methods to analyse big data in the humanities. The report looked both towards the past –

with its references to the 'traditional' fields of bibliography and textual scholarship – and towards the future. [38]

Following the NEH White Paper, Emory University created an emulated environment for the Salman Rushdie papers, a project that attracted a lot of attention from mainstream media. Kirschenbaum continued to collaborate with archivists, resulting in the 2010 report *Digital Forensics and Born-Digital Content in Cultural Heritage Collections* published by the CLIR (Council on Library and Information Resources). [39] Positioned within the long history of bibliography, the report cited D. F. McKenzie's emphasis on electronic texts in his 1985 Panizzi lectures. Like the bibliographers of the past who established reliable textual versions, contemporary scholars had to make sure that electronic texts were accurate, particularly in the case of recovered versions that had previously been lost or deleted. This new breed of literary scholars closely resembled computer specialists who used procedures to discover, recover, and present an 'erased' file as trial evidence. Digital forensics, the report argued, could be used to produce innovative new knowledge in literary studies.

Kirschenbaum expanded this argument in his influential article on the '.textual condition', published in 2013 in *Digital Humanities Quarterly*. He once again compared bibliography and textual criticism in the electronic sphere to the field of computer forensics. He also clearly explained why born-digital materials can and should be used in literary analysis:

> a writer working today will not and cannot be studied in the future in the same way as writers of the past because the basic material evidence of their authorial activity—manuscripts and drafts, working notes, correspondence, journals—is, like all textual production, increasingly migrating to the electronic realm. [40]

In other words, the born-digital materials of today are not ephemeral documents, they are the literary records of tomorrow. The article made clear that the transition from print to digital was impacting not only research methods, but also the primary sources themselves. As a textual scholar, however, Kirschenbaum focused mainly on electronic working files rather than email correspondence.

At around the same time, Chris Prom – a University of Illinois professor and archivist – wrote the DPC report on email preservation, which explicitly stressed the importance of electronic correspondence for future scholars. 'Email records can be used alongside other types of records to develop complete and nuanced narratives', Prom wrote,[41] comparing emails to the letters of the past. To illustrate the historical and legal value of emails, Prom gave the examples of several scandals: Enron, WikiLeaks and the Climatic Research Unit email controversy (also known as Climategate). Since the publication of the report, the Clinton email scandal has once again shown the devastating impact of information – or perceived information – contained in email correspondence. Despite the importance of emails as records, few organisations have dedicated programs to preserve them. Benign neglect, or worse, the automatic destruction of emails, are common.[42] 'As a result', Prom said, 'the end users of email systems frequently shoulder the ultimate responsibility for managing and preserving their own email, thus exposing important documentary records to needless and counterproductive risk of loss'.[43] The report outlined technical solution to preserve emails, focusing particularly on the issue of authenticity.

The DPC report attracted the attention of the Mellon foundation, which announced the formation of a Task Force on Technical Approaches for Email Archives in November 2016. Led by Prom and Kate Murray at the Library of Congress, the Task Force has recently completed its report.[44] One of its suggestions is to identify and train personnel who can work

with large-scale email collections. The report also recommends working with donors, for example to explain the functionalities of open-source tools such as ePADD.

To summarise, the 2010s have so far been dominated by two large trends: the development of guidelines to preserve email archives, and to scale up staff training; and the establishment of workflows to engage with donors of personal digital archives. The second trend finds its origins in previous projects, such as the British Library's Digital Lives, and has led to the 2012 AIMS report on Born-Digital Collections. This work, funded by the Mellon foundation, included best practices for working with donors to preserve materials, before making them discoverable and accessible. Like the Digital Lives project, AIMS mentioned the entire cycle of archives – from donors to researchers – but the focus was mainly on preservation. 'Discovery and access are not possible without completion of the preceding steps described in this model', the report declared.[45] What it means in practice is that engagement with end users is postponed to the distant future, when technical and legal issues will have been solved. The same focus on donors and creators of data, rather than users, can be found in the work of Gabriela Redwine – for example, in her 2015 report on Personal Digital Archiving commissioned by the Digital Preservation Coalition.[46]

That archivists should be primarily concerned with preserving records is not surprising. Yet, a better balance needs to be achieved between preservation and use. In his influential textbook *Modern Archives: Principles and Techniques* (1956), T. R. Schellenberg, a pioneer of archival theory, distinguished between the primary value of documents (the function they served when they were initially created) and their secondary value. 'To be archives', Schellenberg wrote, 'materials must be preserved for reasons other than those for which they were created or accumulated'.[47] Since modern archives are created for users who are not the same as creators, the archivist must necessarily pay attention to the needs of these users.

Schellenberg described the archivist as an intermediary between the donor and the researcher, 'preserving records useful for research'.[48]

For Schellenberg, preservation was inseparable from users' needs, and so was appraisal. He was writing at a time when archivists were confronted with a flood of written evidence, due to administrations producing ever greater numbers of paper records. Preserving everything was not only impossible, it would have been a disservice to scholarship. No one could possibly analyse all this data, Schellenberg argued, therefore the archivist's role was to select only the most relevant sections – *in the interest of researchers*. Schellenberg put the needs of the users first, and worked backwards. Half a century later, tech giant Amazon credits its success to customer obsession. Perhaps it is time for the digital archiving community to shift the balance towards end users – which leads me to the 'After the Digital Revolution' special issue.

III) New Horizons

In the first article, Josh Schneider presents the ePADD software, a project developed at Stanford University. Case studies written by archivists across the world who use the software complete this overview. In the *discovery module*, all messages are redacted to hide identified entities (people, places, organisations) and email addresses. This version is stored in a web server, made available to users with an internet connection, anywhere in the world. Researchers see only a redacted version of the original messages containing extracted entities. The objective is for them to get a sense of the entities present in the archive, and to decide if it is worthwhile travelling to the archival repository to see the full messages without redaction. These messages are displayed in the reading room through the *delivery module* of

ePADD. Researchers can define their own lexicon to analyse the collection, and request copies by flagging the messages they need.[49]

Moving to the specific case of the State Library Victoria in Melbourne, Kevin Molloy gives an overview of the format-mixed archives of four Australian writers: Peter Carey, Sonya Hartnett, Alex Miller and Chinese-Australian writer and translator Ouyang Yu. The article looks at these writers' engagement with the digital, and the impact on their output. How to respond to these hybrid collections is a key question for the State Library Victoria, in terms of preservation but also access. The article thus examines the technologies necessary to deliver born-digital literary content to researchers, without infringing data protection regulations.

In Britain, institutions such as Senate House Library (London) have also responded to the challenge of born-digital records in creative ways. Maria Castrillo takes the example of the *Discworld* computer games based on Terry Pratchett's novels to show that these born-digital records provide unique information on the social, historical and literary contexts which led to their creation. Both the physical format and the content of these computer games will be of relevance to users of the Colin Smythe/Terry Pratchett archive at Senate House Library.

Turning to the example of publishers' archives in the digital age, Samantha Rayner and Jenny Bunn show how to collaborate across the divide between archivists and researchers. Their interdisciplinary article looks at the relationship between the author and the editor within the academic publishing sector. Archival records should not be conceived only as resources for scholarship, they argue. The old way of thinking of archivists as servicing the needs of scholars should instead be replaced by more collaborative models of engagement with born-digital records.

In the next article, Paul Gooding, Justine Mann and Jos Smith respond to Matthew Kirschenbaum's call for new ways to analyse born-digital materials. Based on the example of

the British Archive for Contemporary Writing at the University of East Anglia, their essay uses concepts from the fields of genetic criticism and digital humanities to explore the nature of creativity in the born-digital archive. Playful approaches can lead to completely new ways to engage with writers' archives, leading to innovative scholarly outputs.

In the final article, Andrew Prescott and Jane Winters examine the free-text natural language search query as popularised by Google. Keyword searching is not efficient when dealing with born-digital sources which lack the contextual information that informs searching of the live web. Prescott and Winters draw on three main examples. First, web archives are challenging in their complexity and diversity, and daunting in their size. Second, email archives such as that produced by the White House during the Presidency of George W. Bush contain hundreds of millions of messages. The problems of searching large quantities of born-digital data are also illustrated by the US State Department records leaked by Wikileaks which in 2015 comprised over two million documents. Confronted with these huge quantities of digital information, the problem facing the researcher is not that of finding information at all, but of finding too much. When a search query produces hundreds of thousands of results, ranked only by date, it is difficult to know where to start. This article considers how far other approaches can facilitate easier engagement with large born-digital archives.

Next steps and recommendations

The 'After the Digital Revolution' project started with a simple observation: archivists and scholars need to sit at the same table to solve the problem of born-digital archives. The project has already led to collaborations – for example, Rachel Foss at the British Library and D-M Withers at the University of Sussex are working on an Enhanced Curation initiative that seeks to create new content surrounding acquisitions of contemporary archives. This includes

creating interactive photographs of writers' workrooms (Ted Hughes), recording interviews with archival creators (for example Wendy Cope and Hanif Kureishi) or documenting video-conversational tours of creators' environments. Accounting for the provenance of a collection is a standard archival practice. But for researchers, this process is often hidden. Enhanced curation will change this, by making archivist practices more visible thanks to digital metadata. Enhanced curation generates information in and around the archive, illuminates archival procedures, and enables researchers to read the archive in new ways.

Collaborations between archivists and scholars need to address the entire cycle of born-digital archives – from donors to end users. Archivists have long recognised the importance of involving donors – for example, to understand the organisation and content of the archive. Digital records are still rarely mentioned in these discussions, but this is changing. For example, the University of East Anglia is beginning to collect the work of young and contemporary novelists who work in inventive digital ways.

Following the #AfterDigRev workshops, our first recommendation is to work closely with donors to help them identify and preserve their valuable born-digital records scattered across various platforms: email accounts, Twitter, Facebook, Instagram and the like. The software ePADD already includes a module that allows donors to flag sensitive emails that should remain closed to researchers for a certain period. Donors are better placed than archivists to make important decisions on their own privacy and data protection preferences.

The second recommendation is to engage with policy makers to put digital preservation and open data at the center of the political agenda. In Britain, policy makers have been very keen on growing the Artificial Intelligence industry – the subject of a recent Review by Wendy Hall and Jérôme Pesenti. The development of AI relies on access to data, and yet, data is often unavailable to researchers. To facilitate access to data, Hall and Pesenti

make three suggestions: first, developing data trusts, to improve trust and ease around sharing data; second, making more research data machine readable; and third, supporting text and data mining as a standard and essential tool for research.

As the authors of the Review point out, easing access to data should apply to 'a wider range of sectors'.[50] There is no reason why the Review's recommendations should exclude the archival sector. After all, archival repositories such as the John Rylands Library are largely funded by taxpayers' money. They have a duty to facilitate research rather than hinder it. But what about data protection issues?

Here, it is useful to think of what Google, Microsoft, Amazon, Facebook and Apple are already doing with data. They are 'using the rich, continuous data streams from user interactions continually to train AIs to improve performance in face recognition, language interactions (Siri, Alexa, Cortana etc.), and customer service'.[51] On the one hand, academic researchers are constrained by strict data protection laws and dark archives. On the other hand, researchers working for Internet giants are making huge advances thanks to their access to data. If data is the new oil, these commercial researchers are the new Rockefellers, exploiting their advantage to the detriment of academics. During a recent roundtable, Ulrike Hahn – a professor of psychology at Birkbeck, University of London – denounced a 'farcical situation' where access to data is severely restricted for academics, and easily available for Facebook and the like.[52] While academics need to jump through several loops including approval from Ethics committees, Internet Giants are moving fast and breaking things (to quote Facebook's mantra for developers). Once again, if data is the new oil, universities are at the mercy of data-rich Internet companies. 'Leading players are not just hiring from universities, they are hiring the universities: Amazon, Google and Microsoft have moved to funding professorships and directly acquiring university researchers in the search for

competitive advantage'.[53] To compete in this brave new world, academic researchers need to push policy makers to facilitate access to data – including data locked in dark archives – without infringing on privacy. The Cambridge Analytica scandal has led to a privacy backlash and to calls for tighter regulation of Facebook and other Internet giants. Yet, there is no reason why researchers (including Humanities researchers) could not access anonymised data necessary for large-scale analysis.

The third #AfterDigRev recommendation is to improve the training of graduate students. Text and data mining are now essential tools for researchers, as Hall and Pesenti note. But before learning to analyse data, students should learn more about data curation. For example, the DH Summer School at the University of Oxford offers a one-week course in data curation, which 'provides the foundation for a range of related activities from analysing and visualising research data to promoting access and reuse across a broader scholarly community'. Oxford also offers a new course on Quantitative Humanities, with an introduction to machine learning to gain insight into historical and literary data. There are of course other summer schools that offer similar training – including at the universities of Leipzig in Germany and Victoria in Canada. But we believe that *all* graduate programs in literary studies should train their students to curate and analyse data, and offer introductions to Artificial Intelligence. Turning our back to AI and other research methods will condemn us to increased marginalisation and further lamentations on the decline of the Humanities.

This editorial started with a call for action: born-digital archives are endangered archives, and we urgently need to preserve these collections, make them available and produce new knowledge. Some progress has been made since 2003, when UNESCO warned that the world digital heritage was rapidly disappearing. Institutions such as the BNF in France are now actively preserving web archives.[54] But public organisations and universities are

moving very slowly. It is astonishing that email and born-digital archives are still treated as a new thing that few archivists really understand. In 2016, Arkivum (a provider of data safeguarding solutions) conducted a survey among professionals in a wide range of global galleries, museums, archives, libraries and other memory institutions. Asked about their digital preservation strategy planning, 67% responded that they were still at the 'information gathering/ just starting out' phase. Only 30% had a process in place and were actively doing digital preservation. This slow rate of progress contrasts with the 'move fast and break things' motto of Internet giants.

We need to accelerate the preservation of born-digital literary archives to avoid losing a large part of our cultural heritage. We also need to push for access to these archives through lobbying for open data respectful of privacy. And we should train the next generation of literary scholars to fully embrace the data revolution. As Bill Gates recently wrote on his blog, data can be used to take a humanist approach, to humanise the work that is ahead of us.[55]

---

[1] Jackie M Dooley and Katherine Luce, *Taking Our Pulse: The OCLC Research Survey of Special Collections and Archives*, OCLC Research, 2010, p. 9, available at

<http://www.oclc.org/research/publications/library/2010/2010-11.pdf>, accessed 18 April 2019.

[2] Ricky Erway, *Defining 'Born Digital'*, OCLC Research, November 2010, available at

<https://www.oclc.org/content/dam/research/activities/hiddencollections/borndigital.pdf>, accessed 18 April 2019.

[3] <https://www.oclc.org/research/themes/research-collections/borndigital.html>, accessed 18 April 2019.

[4] Fran Baker, 'E-Mails to an Editor: Safeguarding the Literary Correspondence of the Twenty-First Century at The University of Manchester Library', *New Review of Academic Librarianship*, vol. 21, no. 2, May 2015, pp. 216–24, doi:10.1080/13614533.2015.1040925.

[5] Ian McEwan to Ted Solotaroff, email, nd [December 2007], Harry Ransom Center, University of Texas at Austin.

[6] See Abby Adams's case study on Ian McEwan's email archive at the Harry Ransom Center in this special issue (Schneider et al., 'Appraising, Processing, and Providing Access to Email in Contemporary Literary Archives').

[7] Rebecca Roach, 'The Computer Poetry of J. M. Coetzee's Early Programming Career', *Ransom Center Magazine*, 28 June 2017, available at <http://sites.utexas.edu/ransomcentermagazine/2017/06/28/the-computer-poetry-of-j-m-coetzees-early-programming-career/>, accessed 18 April 2019.

[8] See Callum McKean's case study on Wendy Cope's email archive at the British Library in this special issue (Schneider et al).

[9] Jonathan Pledge and Eleanor Dickens, 'Process and Progress: Working with Born-Digital Material in the Wendy Cope Archive at the British Library', *Archives and Manuscripts*, vol. 46, no. 1, January 2018, p. 67, my emphasis, doi:10.1080/01576895.2017.1408024.

[10] See Dorothy Waugh's case study on Salman Rushdie's email archive at Emory University in this special issue (Schneider et al.).

[11] Stephen Enniss, 'Objects of Study: Special Collections in an Age of Digital Scholarship', *Forging the Future of Special Collections*, edited by Melissa A Hubbard et al, Neal-Schuman, Chicago, 2016, p. 112.

[12] See article by Schneider et al. in this special issue. More information can be found at <https://library.stanford.edu/projects/epadd>.

[13] DADVSI is the abbreviation for *loi relative au Droit d'Auteur et aux Droits Voisins dans la Société de l'Information* (in English: 'law on authors' rights and related rights in the information society').

[14] Isabelle Le Pape presented an overview of the BNF web archiving strategy at the two #AfterDigRev workshops.

[15] Sara Aubry, 'Introducing Web Archives as a New Library Service: The Experience of the National Library of France', *LIBER Quarterly*, vol. 20, no. 2, September 2010, p. 184, doi:10.18352/lq.7987

[16] 'Open Access Policy', *Wellcome Trust*, available at <https://wellcome.ac.uk/funding/managing-grant/open-access-policy>, accessed 18 April 2019.

[17] Eric Lease Morgan and LiAM, *Linked Archival Metadata: A Guidebook*, 23 April 2014, available at <http://infomotions.com/sandbox/liam/tmp/guidebook.pdf>, accessed 18 April 2019.

[18] UK National Archives, *Principles for Determining the Access Status of Records on Transfer*, 2016, p. 5, available at <http://www.nationalarchives.gov.uk/documents/information-management/principles-for-determining-the-access-status-of-records-on-transfer.pdf>, accessed 18 April 2019.

[19] John Markoff, 'Armies of Expensive Lawyers, Replaced by Cheaper Software', *New York Times*, 4 March 2011, available at <https://www.nytimes.com/2011/03/05/science/05legal.html>, accessed 18 April 2019.

[20] J. R. Jenkins, 'The Rise of Analytics in E-Discovery:' *FTI Journal*, December 2015, available at <http://www.ftijournal.com/article/the-rise-of-analytics-in-e-discovery>, accessed 18 April 2019.

[21] Baker, p. 222.

[22] There are of course many literary scholars who have pushed for a move towards 'distant reading', in Franco Moretti's terms. But as a profession, literary studies remain committed to the small-scale analysis of texts. See James English, 'Everywhere and Nowhere: The Sociology of Literature After "the Sociology of Literature"', *New Literary History*, vol. 41, no. 2, 2010, pp. v–xxiii; and Franco Moretti, *Distant Reading*, Verso, 2013.

[23] Armstrong, et al. v. Executive Office of the President, Office of Administration et al. (Civil Action Nos. 93-5002, etc. etc.), Opinion, 13 August 1993. (1 F.3d 1274)

[24] David Bearman, 'Managing Electronic Mail', *Archives and Manuscripts*, vol. 22, no. 1, 1994, p. 29.

[25] Terry Cook, 'Electronic Records, Paper Minds: The Revolution in Information Management and Archives in the Post-Custodial and Post-Modernist Era', *Archives and Manuscripts*, vol. 22, no. 2, 1994, p. 302.

[26] Ibid., p. 306.

[27] A recent report in the *Economist* shows that the internet was built without a memory. Its base was designed to move data and to publish information, not to record information that had been transmitted previously. This 'created an opportunity for a few firms to become the internet's memory. At its core, Google is a list of websites and a database of people's search histories'. In addition to recording information, Google focused on user experience right from the start, with a clean interface and accurate search results. See 'More Knock-on than Network - The Story of the Internet Is All about Layers', *The Economist*, June 2018, available at <https://www.economist.com/special-report/2018/06/28/the-story-of-the-internet-is-all-about-layers>, accessed 18 April 2019.

[28] Miguel Costa et al., 'The Evolution of Web Archiving', *International Journal on Digital Libraries*, vol. 18, no. 3, September 2017, p. 193, doi:10.1007/s00799-016-0171-9.

[29] Seamus Ross and Ann Gow, *Digital Archaeology: Rescuing Neglected and Damaged Data Resources: A Jisc/NPO Study with the Electronic Libraries (ELib) Programme on the Preservation of Electronic Materials*, Library Information Technology Centre, 1999.

[30] See Workbook on Digital Private Papers, <http://www.paradigm.ac.uk/>.

[31] 'The most crucial requirement for the immediate future is for advocacy on behalf of personal digital archives, both in motivating individuals to keep and develop them and in inspiring decision makers to provide policy and funding support to secure and maximise the effectiveness of this essential resource for research and for individual well being', Jeremy Leighton John, *Digital Lives, Personal Digital Archives for the 21st Century: An Initial Synthesis*, British Library, 2009, pp. x-xi, available at <http://britishlibrary.typepad.co.uk/files/digital-lives-synthesis02-1.pdf>, accessed 18 April 2019.

[32] Miguel Costa et al, 'The Evolution of Web Archiving', *International Journal on Digital Libraries*, vol. 18, no. 3, September 2017, p. 200, doi:10.1007/s00799-016-0171-9.

[33] *Missing Links: The Enduring Web - Conference Report*, Digital Preservation Coalition, Jisc, UK Web Archiving Consortium, 1 September 2009, available at <https://www.dpconline.org/docs/miscellaneous/events/389-missing-links-conference-report/file>, accessed 18 April 2019.

[34] Cathy Smith, *Delivering Coordinated UK Web Archives to User Communities.* Missing Links: The Enduring Web Conference, 21 July 2009, London, available at <https://www.dpconline.org/docs/miscellaneous/events/398-0907smithmissinglinks/file>, accessed 18 April 2019.

[35] 'Open Insights: An Interview with Jane Winters', *Open Library of Humanities*, <https://www.openlibhums.org/news/282/>.

[36] Matthew Kirschenbaum et al, *Approaches to Managing and Collecting Born-Digital Literary Materials for Scholarly Use*, National Endowment for the Humanities Office of Digital Humanities, May 2009, pp. 3-4, available at <http://drum.lib.umd.edu/handle/1903/9787>, accessed 18 April 2019.

[37] ibid., p. 24.

[38] Likewise, Lisa Gitelman has paid attention to 'old' as well as 'new' media. 'Like old art, old media remain meaningful', she argues. Lisa Gitelman, *Always Already New: Media, History and the Data of Culture*, MIT Press, Cambridge, MA, 2006, p. 4.

[39] Matthew Kirschenbaum et al, *Digital Forensics and Born-Digital Content in Cultural Heritage Collections*, CLIR, 2010, available at <https://www.clir.org/pubs/reports/pub149/>, accessed 18 April 2019. This report became a cornerstone of the partnership that developed the BitCurator system now widely used by digital archivists.

[40] Matthew Kirschenbaum, 'The .txtual Condition: Digital Humanities, Born-Digital Archives, and the Future Literary', *Digital Humanities Quarterly*, vol. 7, no. 1, 2013, available at <www.digitalhumanities.org/dhq/vol/7/1/000151/000151.html>, accessed 18 April 2019.

[41] Christopher J Prom, *Preserving Email - DPC Technology Watch Report*, Digital Preservation Coalition, 1 December 2011, p. 5, doi:10.7207/twr11-01.

[42] In 2015, the US National Archives and Records Administration announced a plan for collecting and preserving emails as government records. However, their Capstone approach relies on the preservation of a very small portion of all emails (those generated by officials at or near the top of the organisation). See *White Paper on The Capstone Approach and Capstone GRS*, April 2015, available at <https://www.archives.gov/files/records-mgmt/email-management/final-capstone-white-paper.pdf>, accessed 18 April 2019.

[43] Prom, *Preserving Email - DPC Technology Watch Report*, p. 1.

[44] Christopher J Prom and Kate Murray, *The Future of Email Archives: A Report from the Task Force on Technical Approaches to Email Archives, August 2018*, Council on Library and Information Resources, 2018, available at <https://clir.wordpress.clir.org/wp-content/uploads/sites/6/2018/08/CLIR-pub175.pdf>, accessed 18 April 2019.

[45] *AIMS Born-Digital Collections: An Inter-Institutional Model for Stewardship*, University of Hull; Stanford University; University of Virginia; Yale University, January 2012, p. 46, available at https://dcs.library.virginia.edu/files/2013/02/AIMS_final_text.pdf, accessed 18 April 2019.

[46] Gabriela Redwine, *Personal Digital Archiving*, Digital Preservation Coalition, 14 December 2015, doi:10.7207/twr15-01. See also Gabriela Redwine et al, *Born Digital: Guidance for Donors, Dealers, and Archival Repositories*, CLIR, October 2013, available at <https://www.clir.org/pubs/reports/pub159/>, accessed 18 April 2019.

[47] Theodore R. Schellenberg, *Modern Archives: Principles and Techniques*, University of Chicago Press, Chicago, 1956, p. 13.

[48] ibid., p. 31.

[49] Trevor Owens, 'The EPADD Team on Processing and Accessing Email Archives', *The Signal*, 20 October 2014,

<https://blogs.loc.gov/thesignal/2014/10/the-epadd-team-on-processing-and-accessing-email-archives/>,

accessed 18 April 2019.

[50] Wendy Hall and Jérôme Pesenti, *Growing the Artificial Intelligence Industry in the UK*, 15 October 2017, p. 2,

available at

<https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/652097/Growing_the_artif

icial_intelligence_industry_in_the_UK.pdf>, accessed 18 April 2019.

[51] ibid., p. 25.

[52] Ulrike Hahn, 'Freedom and Responsibilities of Researchers in Society', paper presented at the Humboldt

Colloquium 'Moving Forward – The UK-German Research Network in a Changing World', 17 March 2018,

Oxford.

[53] World Economic Forum, 'Assessing the Risk of Artificial Intelligence', *Global Risks Report 2017*, available at

<http://wef.ch/2izSQRP>, accessed 18 April 2019.

[54] UNESCO, *Charter on the Preservation of Digital Heritage*, 15 October 2003, available at

<http://portal.unesco.org/en/ev.php-URL_ID=17721&URL_DO=DO_TOPIC&URL_SECTION=201.html>.

[55] Bill Gates, 'A Humanist Approach to Teaching Kids [Conversation with Jorge Aguilar, Superintendent of the

Sacramento City United School District]', *Gatesnotes.com*, 13 March 2018, available at

<https://www.gatesnotes.com/Education/A-humanist-approach-to-teaching-kids>, accessed 18 April 2019.