IS DIGITAL DIFFERENT?

How information creation, capture, preservation and discovery are being transformed

Every purchase of a Facet book helps to fund CILIP's advocacy, awareness and accreditation programmes for information professionals.

IS DIGITAL DIFFERENT?

How information creation, capture, preservation and discovery are being transformed

Edited by

Michael Moss and **Barbara Endicott-Popovsky**
With **Marc J. Dupuis**

facet publishing

Published by Facet Publishing,
7 Ridgmount Street, London WC1E 7AE
www.facetpublishing.co.uk

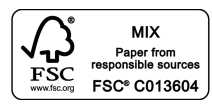
Facet Publishing is wholly owned by CILIP: the Chartered Institute of Library and Information Professionals.

British Library Cataloguing in Publication Data
A catalogue record for this book is available from the British Library.

ISBN 978-1-85604-854-5

First published 2015

Text printed on FSC accredited material.



Typeset from editors' files by Facet Publishing Production
in 11.5/14 pt Garamond and Myriad Pro.
Printed and made in Great Britain by CPI Group (UK) Ltd, Croydon, CR0 4YY.

Contents

Contributors

Emma Bayne

Emma Bayne is Head of Systems Development and Search at The National Archives, UK. She is responsible for leading the design and build of technical solutions to improve access to the vast collection of records. She has worked in a variety of technology roles over the last 15 years. Recently, her focus has been on developing Discovery, an integrated search tool which pulls together archival content and improves findability, through user-centred design, for the collections of over 2500 archives across the UK and beyond.

Ylva Berglund Prytz

Ylva Berglund Prytz works for the University of Oxford and is based within the Academic IT Services, where she manages the RunCoCo service. RunCoCo looks at new ways of working with the public for impact, outreach and engagement and has a particular remit to provide advice, training and support to crowdsourcing and community collection projects.

Ylva's role is to liaise with and support those planning or running crowdsourcing and community collection projects. She also produces teaching and support material and runs training and coaching sessions for project managers and staff. She has been actively involved in a number of local, national and international initiatives using crowdsourcing to engage audiences and enhance or create digital collections. She has also worked on

a range of other digital projects concerned with the creation, use and preservation of digital resources.

Ylva is a member of the Faculty of Linguistics. She also teaches for the English Faculty and serves on the Committee for Library Provision and Strategy in English. She has a PhD in English from Uppsala University, Sweden and has published in the areas of corpus linguistics, English language and computer-assisted language learning.

David Clark

David Clark, Director, CIBER Research, has worked in publishing and related industries for 40 years, as data processor, information manager and analyst. He has a master's degree in Knowledge Engineering. His PhD, from the University of Warwick, concerns the history and establishment of computing as a distinct academic discipline.

Scott David

Scott David works at the intersections of law and technology, where theory informs practice. For 27 years he worked in large law practices (first at Simpson Thacher in New York City and then with K&L Gates in Seattle), with organizations at the front edge of science, technology, business and financial innovation such as Microsoft, TMobile, Gates Foundation, Google, AT&T and many others. Several years ago, Scott joined academia, first as Executive Director of the Law, Technology and Arts programme at the University of Washington Law School, and more recently with the University of Washington Center for Information Assurance and Cybersecurity, creating programmes that provide students and faculty with new opportunities to engage where technology and people meet.

Scott is deeply involved in global data policy work with multiple organizations, including as a member of the World Economic Forum's Global Agenda Council on Data Driven Development, the advisory boards for the MIT Kerberos and Internet Trust initiative and the Open Identity Exchange, and as a senior analyst for the German consultancy KuppingerCole. Through these and other engagements Scott is actively involved with multiple initiatives of various governments, companies, universities and nongovernmental organizations that are developing policy

standards and scalable, distributed legal structures that can help to create unique opportunities for system stakeholders to reduce risk and enhance leverage, security, privacy and value in networked information systems.

Marc J. Dupuis

Marc J. Dupuis, PhD, is a researcher and lecturer with the University of Washington as well as the Director of Human Factors for the Center for Information Assurance and Cybersecurity (CIAC). His main focus is on understanding the information security behaviour of individuals, including issues related to decision making and the user experience. This has included research on the role of trait affect, personality, self-efficacy and risk evaluation on information security decisions made by individuals, as well as an examination of the various security and privacy concerns related to social computing.

In 2014 he started a research group called SPROG – Security and Privacy Research and Outreach Group. The purpose of this group is to provide an environment for innovative research on issues related to security and privacy, as well as to identify opportunities for outreach in these areas with the local community.

Marc earned a PhD and MS in Information Science, in addition to an MPA (Master of Public Administration) from the University of Washington. He has also earned an MA and BA from Western Washington University. He has taught courses on cybersecurity, organizational information assurance, risk management, information assurance strategies, human computer interaction, web design and programming, and research methods.

Barbara Endicott-Popovsky

Professor Barbara Endicott-Popovsky, PhD, is Executive Director of the Center for Information Assurance and Cybersecurity at the University of Washington, designated by the National Security Agency/Department for Homeland Security as a Center of Academic Excellence in Information Assurance Education and Research; Director of the Master of Cybersecurity and Leadership programme; Academic Director for the Masters in Infrastructure Planning and Management in the Urban Planning Department of the School of Built Environments; holds a faculty appointment with the

Institute of Technology in Tacoma; and was named Department Fellow at Aberystwyth University, Wales, (2012). Her academic career follows a 20-year career in industry marked by executive and consulting positions in IT architecture and project management.

Her research interests include enterprise-wide information systems security and compliance management, forensic-readiness, the science of security, cybersecurity education and secure coding practices. For her work in the relevance of archival sciences to digital forensics, she is a member of the American Academy of Forensic Scientists. Barbara earned her PhD in Computer Science/Computer Security from the University of Idaho Center for Secure and Dependable Systems (2007) and holds an MSc in Information Systems Engineering from Seattle Pacific University (1987), an MBA from the University of Washington (1985) and a BA from the University of Pittsburgh.

Tim Gollins

Tim Gollins is currently Head of Digital Archiving at The National Records of Scotland and, as programme director, leads their Digital Preservation Programme. He started his career in the UK civil service in 1987. Since then he has worked on information assurance, user requirements, systems development, systems design, information management, and high-assurance information security, on numerous government IT projects. Tim joined The National Archives in April 2008 and as Head of Digital Preservation led The National Archives' work on digital preservation and cataloguing. He recently worked on the design and implementation of a new digital records infrastructure at The National Archives, which embodies the new parsimonious (digital) preservation approach that he developed.

He recently completed a secondment from The National Archives as an honorary research fellow in the School of Computing Science at the University of Glasgow, where he studied the challenges of digital sensitivity review. Tim remains an honorary research associate in the School of Physics and Astronomy at the University of Glasgow. He holds a BSc in Chemistry (Exeter), MSc in Computing (UCL), and MSc in Information Management (Sheffield). He was a director of the Digital Preservation Coalition for six years and is a member of the University of Sheffield I-School's Advisory Panel.

Norman Gray

Norman Gray (School of Physics and Astronomy, University of Glasgow) studies the development of next-generation scientific data management. He has been directly involved with scientific data-management software since 1997, working with the Euro-VO project (VOTECH) and Astrogrid in Glasgow and Leicester, and has more recently been the Principal Investigator (PI) of a number of (broadly) astroinformatics projects funded by the Engineering and Physical Sciences Research Council (EPSRC) and Jisc in the UK. Between 2012 and 2015 he was the chair of the International Virtual Observatory Alliance's (IVOA) Semantics Working Group and is the co-author of a number of IVOA Recommendations.

Valerie Johnson

Dr Valerie Johnson is the Interim Director of Research and Collections at The National Archives, having previously held the role of Head of Research. Prior to this, she worked on a funded history project based at the University of Cambridge History Faculty. She holds an MA with Distinction in Archive Administration and was awarded the Alexander R. Myers Memorial Prize for Archive Administration. She also has a PhD in History for her thesis, 'British Multinationals, Culture and Empire in the Early Twentieth Century', for which she won the Coleman Prize. She is a Registered Member of the Society of Archivists, a Trustee and member of the Executive Committee of the Business Archives Council, a Fellow of the Royal Historical Society and a Fellow of the Society of Antiquaries. She has worked as an archivist and a historian in the academic, corporate and public sectors.

Gavan McCarthy

Associate Professor Gavan McCarthy is Director of the University of Melbourne eScholarship Research Centre in the University Library, founded in 2007. His research is in the discipline of social and cultural informatics, with expertise in archival science and a long-standing interest in the history of Australian science. He contributes to research in information infrastructure development within the University and his projects highlight strong engagement with community. His distinctive cross-disciplinary research reaches into other fields such as education, social work, linguistics,

anthropology, population health and history. He re-examines theoretical foundations and tests new theories through practical interventions with a focus on public knowledge domains, contextual information frameworks and knowledge archives.

Helen Morgan

Helen Morgan is a Melbourne writer and archivist. She is a research fellow in the area of cultural informatics at the University of Melbourne eScholarship Research Centre, having significant experience of working in collaborative research teams using digital technologies, with particular emphasis on building resilient contextual information frameworks, exploring the challenges and requirements of mapping cultural heritage in digital/networked environments and the transfer of knowledge between researchers, memory institutions and the community. Helen has worked as an information architect and exhibition designer on the Australian Women's Archives Project since its inception in 2000 and is currently a Chief Investigator on the Australian Research Council-funded 'The Trailblazing Women and the Law Project' (2012–15). She is the author of *Blue Mauritius: the hunt for the world's most valuable stamps* (Atlantic Books, 2006).

Michael Moss

Michael Moss is Professor of Archival Science at the University of Northumbria. Previously, he was Research Professor in Archival Studies in the Humanities Advanced Technology and Information Institute (HATII) at the University of Glasgow, where he directed the Information Management and Preservation MSc programme. Prior to being appointed to HATII, he was archivist of the University from 1974 to 2003. He was educated at the University of Oxford and trained in the Bodleian Library. He is a non-executive director of the National Records of Scotland and until 2014 was a member of the Lord Chancellor's Advisory Council on National Archives and Records. In 2015 he was Miegunyah Distinguished Fellow at the University of Melbourne.

Michael researches and writes in the fields of history and the information sciences. His recent publications include: 'Archival research in organisations in a digital age', in David Buchanan and Alan Bryman (eds) *Handbook of*

Organizational Research Methods, Sage, 2009; 'Brussels Sprouts and Empire: putting down roots', in Dan O'Brien (ed.) *The Philosophy of Gardening*, Blackwell Wylie; 'The High Price of Heaven – the 6th Earl of Glasgow and the College of the Holy Spirit on the Isle of Cumbrae', *Journal of the Architectural History Society of Scotland*, 2012; 'Where Have All the Files Gone, Lost in Action Points Every One?' *Journal of Contemporary History*, 2012; and 'From Cannon to Steam Propulsion: the origins of Clyde marine engineering', *Mariner's Mirror*, 2013.

David Nicholas

David Nicholas is a Director of the CIBER research group (http://ciber-research.eu). The group is best known for monitoring behaviours in the virtual space, especially with regard to the virtual scholar and the Google Generation. David holds chairs at the College of Communication and Information Studies, University of Tennessee and at Tomsk University, Russia. Previously he was Head of the Department of Information Studies at University College London (2004–11), and previous to that was Head of the Department of Information Science at City University.

David's interests include the digital consumer, mobile information, e-books, e-journal usage; web analytics and scholarly communication.

David Thomas

David Thomas worked for many years at The National Archives, where he became Director of Technology in 2005. Prior to that he held a variety of posts and from 1999 led The National Archives' pioneering developments of systems to deliver digital copies of records to online users. He is a member of the project board for the Wellcome Trust's digital library and also advised the Globe Theatre on its plans for a digital archive. He is currently a visiting professor at Northumbria University. His current research interests focus on the issues of acquiring, reviewing and delivering digital records.

Introduction and acknowledgements

Michael Moss and Barbara Endicott-Popovsky

The purpose of this book is to introduce students, particularly but not exclusively those on information studies programmes, to the issues surrounding the transition from an analogue to a digital environment. The contributors strip away much of the e-hype that surrounds the digital environment and focus on the opportunities and challenges afforded by this new environment that is transforming the information landscape in ways that were scarcely imaginable even a decade ago. Contributors examine whether analogue practices and procedure that are largely handicraft are still valid and if they shape or distort those in the digital, which can best be characterized as industrial and requiring engineering solutions.

By drawing on examples of the impact of other new and emerging technologies on the information sciences in the past, such as the printing press in the 15th century, the wet-copy process in the 18th century and the typewriter in the late 19th century, the book emphasizes that information systems have always been shaped by available technologies that have transformed the creation, capture, preservation and discovery of content. Whilst seeking to avoid techno-determinism, the contributions illustrate the ways in which the digital environment has the potential to transform scholarship across the disciplines at all levels, even if it has not done so yet, and to break down barriers between the academy and the wider community through social networks and crowdsourcing. There are analogies here with the way in which the reordering of libraries pioneered by Martin Schrettinger

in the early 19th century helped to transform scholarly enterprise that came to be described in all disciplines as 'scientific'.

From the different perspectives of each chapter the contributors explore the role, as they see it, of information professionals in this rapidly changing digital landscape, which is challenging the very existence of the traditional library and archive as more and more resources become available online and as computers and supporting networks become more and more powerful. Users expect to be able to work at their screens from home, however unrealistic this may seem to many traditional curators.

The authors alert the readers to the perils and pitfalls of the digital world with its ever-present risks of breaches in security and unwitting infringement of copyright, data protection and other regulatory constraints. They argue for the need for new ways and models of working and emphasize the importance of information professionals from different disciplinary perspectives working with the grain of societal expectations through a critical encounter with the emerging technologies and mechanisms. Attention is given to the long-term curation and preservation of both born-digital and digitized content and, importantly, to modes of access. Given the broad scope of this book, it has been possible only to introduce the reader to the salient features of the topics covered by each chapter and provide pointers to further reading.

The editors would like to thank all the contributors for their help and support in the preparation of this book, particularly Marc J. Dupuis for formatting the text. Norman Gray would like to thank Susan Stuart for exacting and detailed comments on the drafts of his chapter, members of the semantic-web@w3.org list for 'in-use' references and Chris Bizer for making available an early copy of the 2014 Linked Data cloud. Valerie Johnson and David Thomas wish to thank the following librarians who kindly agreed to be interviewed and provide generous insights for their chapter: Simon Chaplin, Helen O'Neill, Darlene Maxwell, John Tuck, Amy Warner, Dace Rozenberga and Jane Winters.

The editors would like to thank Helen Carley and her team at Facet Publishing for being so understanding. This book has been a long time in the making.

CHAPTER 1

What is the same and what is different

Michael Moss

Screens have become so ubiquitous and so much part of our daily lives that it is easy to forget that they are simply rendering content that is familiar in the analogue world and which still surrounds us. When we buy goods in a store we are usually handed a receipt which is the evidence of a transaction, even if we have paid for the goods electronically with a card. When we use a word processor we render words on a screen in much the same way as we render words on a piece of paper when we use a typewriter or a pen. It is easier, as we can delete and redraft much more readily, but the process is more or less the same. We are rendering or inscribing content on another medium. However, there are things that are different, as processes are happening between the keyboard and the screen which allow the content to be rendered in the typeface and point size we have chosen. With a typewriter we were confined to the typeface and point size provided and the only choice we had was between capitals and lower case. If we wanted the content rendered in another typeface or, for example, in italics, then this had to be done by resorting to a typesetter. What word processing has done is to bring together the typewriter with the skills of the typesetter and cut out many intermediate processes. The quality of the output has improved and a great deal of frustration has been removed, but on the whole content has remained stable. We can do more with the content apart from changing the typeface and point size: we can easily change the layout and we can insert pictures, graphs and tables simply by 'cutting and pasting'. This is a term borrowed

directly from the analogue world of printing and describes precisely what we do in the digital environment. We identify something we want to insert in a text, cut it out using a cursor (not a pair of scissors) and insert it into the text. The final product looks much as before, but we have done it ourselves without resorting to either a designer to create the layout or a typesetter to set the text. It might look better if we had done so, but the final product is essentially the same. It is very easy when addressing the digital, which is nearly always bracketed with the term 'new technology' to assume without thinking that it represents a discontinuity with the past and makes possible radically new ways of doing things. It may allow us to do things more quickly, but it may not necessarily do things differently. This is a theme that will recur regularly in this book.

The speed of the digital depends only in part on the ease with which content can be rendered on the screen, but critically on the tractability of the world wide web and associated communication systems. I can type a message on my computer in the United Kingdom and within seconds it has been delivered to a recipient on the other side of the world. What is different about this is the speed, not the ability to send a message half way around the world. I could do that before by using postal services, which might have taken several weeks but nonetheless depended on technology to get the message there, at first on sailing ships, later by steam ship and finally by aeroplane. Each of these represented a step change in technology and, as a consequence, in the speed of transmission. Even when communication was very slow it was possible to both build and maintain relationships. How otherwise could business have been conducted? From the earliest times all bureaucracies have depended on letter writing. From at least the early Middle Ages the Vatican received a swelling tide of supplications to the Holy See from every Catholic country, which were registered in the registers of the Penitentiary and survive to this days (see, for example, Clarke and Zutshi, 2014). Two great European banking families, the Corsinis of Florence and the Fuggers of Augsburg, have left behind vast collections of correspondence dating from the mid-16th century (Beale, Almond and Archer, 2011)[1]. Almost every archival collection is full of evidence of extensive correspondence and other transactions, particularly accounts, from every part of the known world. From the beginning, efforts were made to speed up communications: by the Romans with their networks of roads and beacons, much later by Napoleon with his manual semaphore telegraph system, in the late 19th century with

the telephone and cable telegraph and in the late 20th century with the fax machine (Coppersmith, 2015). Speed is needed to unify administration and to improve the efficiency of markets. These are the arguments advanced for the high-speed Atlantic cable that will transfer data in nanoseconds. It is claimed that stock trading will get 5.2 milliseconds faster, which allegedly will be worth billions to those who have access to the cable.[2] Much the same could have been said for the Borromeo family's network of post horses that crisscrossed Europe in the 16th century to bring market intelligence to its bank's headquarters in Milan.[3]

What might be missing from such rapid transactions is time for the reflection and reflexivity that characterized communication at a slower pace in the analogue environment. This absence has preoccupied some scholars, who are critical of the digital environment we now inhabit. The neuroscientist Susan Greenfield argues in her latest book, *Mind Change: how digital technologies are leaving their mark on our brains*, that it is changing the way our brains work (Greenfield, 2014); Marc Prensky, the educationalist who coined the terms 'digital natives' and 'digital immigrants', holds much the same opinion but from a different perspective:

> A really big *discontinuity* has taken place. One might even call it a 'singularity' – an event which changes things so fundamentally that there is absolutely no going back. This so-called 'singularity' is the arrival and rapid dissemination of digital technology in the last decades of the 20th century.
>
> (Prensky, 2001, 1)

Andrew Keen, Nicholas Carr and Jaron Lanier all hold similar opinions. Tara Brabazon – whose book *The University of Google: education in a (post) information age* attracted a great deal of attention when it was published in 2007 for its forthright attack on the impact of the digital on higher education – seems to have shifted her ground, partly, it would seem, because most of these authors have extreme views and do not see, as she does, that 'Life and learning are not filed into analogue and digital folders. They spark and dialogue' (Brabazon, 2014). Other authors fail to look back to previous examples of transition in communication technologies, most obviously the coming of printing. This is often mistakenly linked to the Protestant Reformation – as Marshall McLuhan would have had us believe – but in fact evidence suggests that the Counter-Reformation made more effective use of it. The shift from

patient copying by hand to mass production and distribution resulted in precisely the same 'singularity' that Prensky describes, from which 'there [was] absolutely no going back', but it is arguable that it did not change things in the way that some of the alarmist critics of the digital environment suggest, and this is borne out by some of the contributions to this book. Brabazon is nearer the mark when she suggests that the analogue and the digital interact with each other in much the same way that when printing first began it imitated script to ensure a seamless connectivity with the past. Only later were new fonts created that made both printing and reading easier. It was never imagined that printing would suddenly replace script in a binary exchange.

There may, however, be some loss of reflection, that moment staring at the wall or the screen, lost in thought, and of reflexivity; as Julia Gillen put it, 'how we come to interpret and reflect upon our own actions and experiences and communicate these to others in specific language practices' (Gillen, 1999). The screen does not intrinsically prevent or inhibit either of these, apart from its apparent insistence on the shortening of time. There is no reason to strike a deal in nanoseconds or to respond immediately to an e-mail any more than there was to settle accounts in three days or to reply to a letter the day it was received. These are all matters of personal choice and there are many practices designed to inhibit such behaviour by emphasizing the importance of reflection – such as 'mindfulness' – that take us back to that old aphorism, 'stop and think'. What might be different may be the sheer tractability of the web that is the handmaid of the systemic risk that lay at the heart of so much of the trouble in the 2008 financial crisis. It is a truism to say that the insurance group AIG could not have bet the whole of the world's Gross Domestic Product without the internet. The same could be said of the part played by the postal service, or the telegraph or the cable in other financial catastrophes. In fact, when communications were poor or non-existent, greater trust was needed to prevent fraud or ill-timed transactions. When exchange rates were much more unstable than they are today or news took a long time to reach a market, transactions were much riskier, as we know from Shakespeare's play *The Merchant of Venice*. This is why it was not until the coming of the submarine cables to the Far East that futures trading became possible in the commodity markets. Merchants in London need access to local market intelligence before fixing the prices of goods that may take several months to arrive.

Behind all these arguments lies the question that has troubled philosophers since classical times, of how we conceive of technology. Do we conceive of it as neutral, in the Platonist tradition, or do we endow it with 'agency', as McLuhan did in his much-quoted catchphrase 'The medium is the message', a form of technical determinism that shaped 'the scale and form of human association and action' (McLuhan, 1964, 9). In doing so he was endowing inanimate 'objects' with agency in what has come to be known as the 'linguistic turn'. The debate he sparked still has currency. Greenfield echoes McLuhan when she suggests that digital technology can in some ill-defined way shape the pattern of the neurons in our brains. Peter-Paul Verbeek, the Dutch philosopher of technology, has sought to navigate a path between these two opposing perspectives by introducing the concept of mediation, which, he argues, 'helps to show that technologies actively shape the character of human world relations'. He continues: 'Technologies do not control processes of mediation all by themselves, for forms of mediation are always context dependent – otherwise we would be back at the technological determinist view' (Verbeek, 2010, 11). Although there are many other authors who have contributed to this debate, such as Albert Borgmann, Don Ihde and Bruno Latour, Verbeek's analysis helps us to grapple with the question at the heart of this chapter, as all communication, in whatever form, is mediated by technology and, as all archivists know, is 'context dependent'.

For the information manager and the archivist, context in the digital environment is complex, not because it needs to be but because practices that were second nature in the analogue world have been abandoned. For example, a letter written on a piece of paper had a form and properties that unambiguously provided the recipient with context: there was a letterhead declaring where it had been written, a date, the name of the person to whom it was addressed, and a salutation and a valediction. These features often themselves conveyed information; a formal or informal salutation or valediction told a reader something about the relationship of the writer and the recipient. In business correspondence the header often included a reference that located the letter in a file plan that, once the document was filed, would give it further context. All these practices were built up over centuries, beginning with dockets, which were transformed over time into manila files held together with bits of string, known as tags. In many organizations, but particularly government bureaucracies, discussions and correspondence were usually summarized in what were known as minutes,

often developed over time as policy was developed (Moss, 2015). Accounting records mostly followed double-entry principles, first systematized by the Italian Franciscan Luca Pacioli (1494) in the late 15th century in his famous book *Summa de Arithmetica, Geometria, Proportioni et Proportionalita*. Receipts and invoices were referenced and entered carefully into a journal or organizer and posted to a ledger. In the ledger, balances were struck with suppliers and customers and, more often than not, the balance sheet and profit and loss account was calculated. Taken together, all these practices recorded context and provided an unambiguous audit trail which was easy to follow. As technology developed and organizations grew and became more complex these practices were simply transferred to the new environment. The appearance of the typewriter in the early 19th century changed little, except that it made copying more straightforward, as did the invention of the photocopier (Kittler, 1999).

Change came first with the introduction of mechanized accounting systems, which meant there was no longer any need for journals and ledgers, as receipts and invoices could simply be coded and aggregated and disaggregated at will. In other words a balance could be struck at the press of a button. Although the result was exactly the same as it had been in the analogue environment, the working was no longer visible. There was still an audit trail, but the context changed in the transition to the digital environment, leaving a much impoverished record for the archive to capture. It is possible to link individual entries in the accounting database with individual receipts and invoices, but it is unrealistic to expect archives to keep them. Given the way that the digital system works, this transition was probably inevitable. Less to be expected are the consequences of the introduction of the networked personal computer and the emergence of the internet, which coincided with the need to make organizations more competitive by stripping out costs. This resulted in the disappearance of secretaries (who typed letters), managers now typing their own communications (largely e-mails), the closure of registries where files were maintained and stored, the closure of libraries and the takeover of information systems by computer scientists. The outcome is not only the disappearance of systematic filing, but also a fundamental change in the way business is transacted. This can be seen very obviously in the layout and features of e-mails, which have lost much of the form and structure of a letter. The recipient's name is now simply an e-mail address that may or may

not have much to say about the identity of the person, the title of the e-mail is usually less than informative and the date is simply captured from the system. The content of the message usually lacks the salutations and valedictions of a letter and it may only open with 'Hi' or 'Hello' and end with a name or nothing at all. The other feature is the ease with which people can be copied in, either explicitly (visibly) or implicitly (invisibly – blind copying). Moreover, the default of most e-mail systems is to store everything, and, as we all know, even when it has been deleted it can be resurrected. In many organizations the minute or memorandum has disappeared and been replaced by an e-mail thread, the stuff of mosaics. Although digital output is stored in what are called 'files' these bear no resemblance to files in the analogue world. It is the context of the technology that has led to this state of affairs, but the context of the content, which was inherent in analogue practice, has vanished (Moss, 2013).

Information managers and archivists imagined that they could overcome this state of affairs by intervening in the records-creation process. Continuum thinking, which was developed at Monash University in Australia by Sue McKemmish and Frank Upward, is predicated on this assumption:

> Archival documents first and foremost provide evidence of the transactions of which they are a part – from this they derive their meanings and informational value. The effective creation and management of archival documents are critical to their use and the role they play in governing relationships in society over time and space. Their effective creation and management are also preconditions of an information-rich society and underpin the public accountability of government and non-government organisations, freedom of information and privacy legislation, protection of people's rights and entitlements, and the quality of the archival heritage, made up of documents of continuing value. The concept of the archival document can provide a framework for a greater shared understanding of the nature of recorded information, and of the importance of transactional records to the continuing functioning of a society.
>
> (McKemmish, 1997)

This all sounds very neat, but attempts to implement such policies have failed. The underlying premise was an attempt to recreate in the digital environment, through the imposition of what are termed Electronic Document and Records Management Systems (EDRMS), the registry systems that had been

swept away by its advent (see, for example, The National Archives, 2010). Such systems are resented by management, as they impose additional burdens without contributing to efficiency, which from the outset was the main purpose of all registry systems. Busy staff who can see no added value or benefit in following prescriptive rules will not be bothered to file documents, particularly the plethora of e-mails that may or may not be important (Currall et al., 2002). In many organizations IT support is contracted out against tight budgets, and any attempt to add further utilities – and therefore costs – will be resisted. Contrary to continuum thinking, the function of information and records management is not to create an archive, which may or may not be a consequence of the process. Despite evidence that EDRMS have been a failure, there are those who still cling to the idea that records management that will yield records for deposit in the archive can be mandated, particularly across government. They are mistaken and, as such, they divert attention away from the important question of what it is that the archive is likely to receive, which can best be characterized simply as 'stuff' with no discernible structure.

This is very different from the analogue paradigm, where archivists could expect to accession records in some semblance of order and where the context was discernible by simply looking at the content and, where possible, its place in the file plan or registry system. This is no longer the case. It is possible using various computational techniques to parse the content so that different genres can be separated, but, in the absence of the familiar files of the analogue world and in face of the fact that so much more survives than before, that still leaves open the question of how to select what should be kept. As a rule of thumb, archivists would claim only to keep some 5% of content in the analogue world, selecting only records that related to policy and discarding the bulk of records relating to individual cases, often referred to as 'particular instance papers'. There were exceptions, such as records relating to major contracts. As historical scholarship has changed and with the growth in demand from genealogists there has been pressure to keep more, particularly big collections rich in genealogical material – for example, records of all the military who served in the wars of the 20th century. It could be argued that, rather than trying to disentangle the content, everything should be kept, but that would impose a considerable cost burden downstream. Irrespective of the interests of family historians, the volume kept is likely to increase to an estimated 20% because the way of transacting

business in the digital environment has changed. It is difficult to be precise, because most of the evidence we have is either anecdotal or what can be observed from individual case studies. Processing and curatorial costs will add significantly to downstream costs. We should be able to develop tools that will help us to identify records that relate to policy by tracing e-mail threads, searching for key words or elements and reviewing the length of documents, but we do not yet have those tools (see, for example, Allen, Connelly and Immerman, 2015).

Once content has been selected for preservation, the expectation is that it will become publicly accessible sometime in the future. For records produced by the public administration this is not the same as Freedom of Information, which normally relates to access to records relating to specific topics, but the opening of all records unless there is some legal reason for them to remain closed, such as the disclosure of personal information or records that might contravene international agreements, such as the Geneva Convention. In the United Kingdom public records are opened with these conditions after 20 years; in some jurisdictions the periods are even shorter. Reviewing content for information that must remain closed, usually termed sensitivity review, is of necessity labour intensive. In the analogue world reviewers look through the papers that are to be preserved and either redact information, such as personal information, or remove individual pieces (pages) if redaction is impractical because too much content needs to remain closed. Only rarely in the United Kingdom today are whole files closed. Inevitably, mistakes are made, but for researchers to find them is like looking for needles in haystacks. This will not be the case when digital content is made available online, as ubiquitous search engines will index the content and make it relatively simple to identify information that should not have been released. This presents the archive with a major obstacle in granting access to born-digital content against a background of tightening privacy regimes and hardening public attitudes to inappropriate disclosure. The US Council of Library and Information Resources (CLIR) has warned collecting archives not to take digital content unless it has been reviewed for such sensitive content because, once deposited, it exposes the archives to contingent liability and can be 'discovered' for litigation. This is very different from the paper world, because the risk of discovery is so much greater (Redwine et al., 2013).

It is impossible to completely automate the process of review, as sensitivity

is nearly always context dependent. For example, if you sign off a document with your name and your role in an organization this is unlikely to be sensitive, but if you are referred to by name in a document this could be sensitive. The problem becomes more acute in the digital world when, by piecing together an e-mail thread in what is termed a 'mosaic', it might be possible to identify sensitive content. However, manually reviewing an enormous quantity of digital content in no logical order would clearly be too expensive and impractical. One response, to which some commentators have already drawn attention, is the precautionary closure of records for a long time (Erdos, 2013). 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 is not known, for minors it is closed for the whole period; for those deemed to be over the age of 16, for 84 or 94 years. There are good reasons for such long closure periods. They safeguard the individual, particularly if the material might affect their health and well-being, and they also help to prevent identity theft – used by criminals and, unfortunately, by law-enforcement agents. These are the longest mandatory closure periods and, inevitably, precautionary closure could be expected to be for a similar time. This is unacceptable in an open democracy, where records are the means by which the executive can be called to account.

Ways need to be found to identify content that might contain sensitive information. This can be done using sophisticated information retrieval protocols that employ techniques which are similar to those used by archivists and documentary scholars, known as diplomatics. What information retrieval protocols do is look for names that might be sensitive, such as those of presidents and kings, or combinations of entities that could identify individuals, such as a name and a date of birth or a role, for example police inspector, a place and so on. They might also look for specific words, such as terrorist, informer and so on, or the length of a document, the number of words used and so on. All these attributes can be bundled together as significant properties. These tools do not yet exist, but there are utilities under development such as Project Abacá at Glasgow and Northumbria universities in the United Kingdom and the Mellon-funded Bit-Curator at the universities of Maryland and North Carolina in the United States.[4] These utilities will be able to distinguish sensitive information at only the most simplistic level, such as an insurance number or details of a bank account; all other instances that are flagged will need to be reviewed. They will be able to rank sensitivity, prioritizing

instances of possibly the highest sensitivity. Inevitably there will be errors, as there were in the analogue environment, but they will be easier to detect and the owners of the information will need to be satisfied that the level of risk is acceptable. This will vary from one organization to another. For example, in the United Kingdom, the Foreign and Commonwealth Office and the Ministry of Defence, which handle a great deal of sensitive information, will be much more risk adverse than, say, the Department of Energy and Climate Change. Of course, sensitivity does not just apply to public records, but to all records. Given the risks associated with inappropriate disclosure in the digital environment, information needs to find a place on risks registers. This is new territory for information services and archives, which have often seen themselves as the final arbiters of what should be selected for deposit and, in many jurisdictions, of the terms of access. If risk is overlooked, then the archive or library will be exposed unnecessarily to contingent liabilities. Although this is new territory, these developments are as much a consequence of changing public attitudes to privacy as of the digital environment itself, which has enabled the so-called 'surveillance society'. These issues are explored by Scott David and Barbara Endicott-Popovsky in Chapter 5.

Once content has been reviewed for sensitive information and decisions have been taken regarding what content should not be released immediately, there is a further problem that makes the digital environment very different from the analogue. In the analogue environment content can be listed easily, down to item level (a file or a volume of assorted material) and sometimes an individual object (a letter, a telegram, a memorandum, and so on). This will not be possible for a large blob of 'stuff', and for two reasons: there will be too much of it and the listing would not be particularly useful. It would be possible mechanically to capture details of objects from what is called ambient metadata; crucially, its date and some indication of the author and to whom an object may have been addressed, and possibly the subject. It would also be possible to link objects together using graphs and digital forensic techniques. Such techniques would allow the user to navigate pathways through the content and avoid blind alleys leading nowhere, for example to people who were copied in just for the sake of it. The user's experience would not be the same as for records in the analogue environment, where the reader can follow the order of documents from a register or a file, nor would it be like using commercial search engines that yield results randomly ranked. Just like the software that needs to be developed for appraisal and sensitivity review,

information-retrieval utilities will need to navigate a sequential logical route through a maze of 'stuff', which, to satisfy users, will need to be as precise as possible. One of the challenges of information retrieval even in the analogue world is that objects often relate to one or more entity, usually referred to as a one-to-many relationship. This was resolved in the analogue world by filing copies of a document in several places or by elaborate cross-referencing. In the digital environment navigational tools will need to have the flexibility to chart multiple routes, possibly across several collections. For example, in the United Kingdom all government policy involves expenditure that, if it is large enough, requires the approval of the Treasury, so there will inevitably be interaction between the Treasury and the sponsoring department. This will require, as Norman Gray, Tim Gollins and Emma Bayne discuss, in Chapters 3 and 6, the re-engineering of cataloguing.

Using technologically dependent tools to interrogate a large blob of heterogeneous 'stuff' in order to discover information may seem at first sight to be different from using conventional analogue finding aids; but will it? In the analogue world users are very dependent on what the cataloguer has chosen to catalogue; they do not have the luxury of free-text searching. Before online catalogues were introduced, users were equally dependent on the way in which indexes were constructed. However much archivists and librarians liked to pretend that cataloguing was objective, it inevitably reflected contemporary preoccupations and the individual interests of the cataloguer. This is what Clifford Lynch, director of the Coalition of Networked Information, dubbed as 'the haphazard historical gerrymandering of knowledge into institutional collections belonging to communities' (Lynch, 2001).

Users will need to become familiar with new utilities that are now only in development, in the same way that they have got used to using online catalogues and search engines to find useful stuff. What may make the new utilities different is the precision with which they should be able to locate and visualize relevant content, which is why the risk of inadvertent disclosure is so much higher. The lack of utilities helps to explain why the digital world has as yet made little impact on scholarship in the humanities, as Valerie Johnson and David Thomas argue in Chapter 9. Referencing should be straightforward, as there should be sufficient ambient metadata to make it possible to identify individual objects within the overall aggregation. What of course is very different is ubiquitous access from the desktop that brings with it all sorts of

intellectual property rights and copyright issues, which are addressed by Helen Morgan and Gavan McCarthy in Chapter 8. No longer will it be necessary to visit an archive or a library to access material. This does not mean that archives and libraries as we know them will cease to exist; this will not be the case, the same as online shopping will not extinguish shops. But there will be fewer of them and they will have to reinvent themselves so as to deliver a range of online offerings and services (Hernon and Matthews, 2013).

Finally, there is the issue of preservation. In the analogue world records survive for a very long time, quite often in less than ideal conditions, provided that they do not get wet or eaten by rodents. Archivists and librarians store them in strongrooms to which they have the keys. The digital world is different, as the content on our screens is rendered from a binary bit pattern consisting of ones and zeros. Bit patterns are notoriously logically unstable. Every time they are opened, their logical structure changes, and so does some of the ambient metadata, for example the date (Allison et al. 2005). This is a formidable obstacle to preservation, but utilities are being developed, such as the Forensic Tool Kit (FTK) which can capture bit patterns without disturbing the logical structure.[5] The actual process of preservation is relatively simple compared with the other challenges of the transition to the digital environment, but – and it is a big but – it will require much more monitoring and surveillance than equivalent analogue content (Gollins, 2009). No one knows quite how much, but, as with everything to do with the digital environment, however much costs appear to come down, it will be much more expensive and require specialist staff with the necessary technical skills to ensure not just that a digital object is preserved and is what it purports to be, but also that it is held as securely as in the analogue world. As we have learned from WikiLeaks and Edward Snowden, digital repositories without the right safeguards are exposed to the theft and distribution of information on an unimaginable scale. What is the same here is the human factor; what is different is the power of the internet as a distribution channel. Security is Barbara Endicott-Popovsky's subject in Chapter 7.

Because preservation of born-digital content requires intervention, the digitization of analogue content, for all its advantages, is not considered to be an appropriate preservation medium. In other words, analogue content should not be destroyed once it has been digitized, but it can be shrink-wrapped and put into deep storage. There are also other factors: the quality of digital cameras and scanners is improving all the time, as are techniques

for compressing high-resolution images. The digitization of records needs careful thought as it presents issues, often overlooked, that make them different from the analogue form, largely because the principal motive is to make content available on the web. This is not an issue with printed books, as they can simply be 'hung off' a content management system and, if the content has been OCRed, then it can be indexed by search engines. The same is not the case when single objects are digitized, because the only way they can be linked together is through the descriptive metadata. Conventional cataloguing practices are not fit for purpose, as search engines search only the substantive textual element and not the rest. Conventional catalogues also work hierarchically and are often difficult to navigate. Digital consumers navigate content in multiple directions, in much the same way that users of archives and libraries explore collections serendipitously, as David Clark and David Nicholas explain in Chapter 2. It is possible to hang digitized content off an archival catalogue, but then, because only occasionally will it be OCRed, it is just an extension of the catalogue.

The expectation of suppliers (archivists, librarians and their funders) is that digitization will add value to a collection or the objects within it. This is much more difficult than it would seem, largely because the curatorial professions notoriously do not focus on customers in ways that publishers in the print culture must do in order to survive. The first thing that has to be grasped, which should be self-evident, is that consumers can enter a site exposed to the internet at any level, and if they are going to stay they must be able to navigate easily within it. This means being able to move between analogous objects and to discover information that gives context to an object. Most objects have one-to-many relationships and will have multiple contexts and form elements in multiple narratives, which may not be known to those who described them. Well-designed websites that are built around such collections need to be open ended, otherwise they can in no sense be described as a learning resource. They need to embed the concept of co-creation, so that users are empowered to contribute content to the descriptive metadata and can incorporate links into their own collections. This is the subject of Chapter 4 by Ylva Berglund Prytz. All of this demands that a great deal of thought be given to the architecture of a site and how content is going to be selected, described and contextualized, and who the potential customers are, before digitization begins. If sites are to be co-created, then they have of necessity to be dynamic, in other words there must be someone

at the other end of the line to respond to queries and to add new content. Although none of this should be new or different, it is regularly overlooked, largely because it is assumed that analogue practice has nothing to tell the digital.

In practice, from the perspective of archivists and librarians there is more that is the same in the digital environment than is different. Much of the argument that the digital is different came from technologists who have never troubled to learn what it is that archivists, librarians and records managers do. They think that money can be saved by jettisoning what, to their eyes, seemed wasteful practices. Managers, particularly in the public sector where there is pressure to reduce expenditure, are easily persuaded. As a result, practices that have been developed over hundreds of years are being lost and a binary opposition has opened up between the technologists and the curatorial professions. Matters were made worse by those who made an industry out of digital preservation and failed to address substantive issues surrounding content, such as the ingrained problems of appraisal, sensitive content, in particular data protection, and navigation. The chapters in this book are intended to help to breach the binary divide and raise issues that can be resolved only through collaboration.

References

Allen, D., Connelly, M. and Immerman, R. (2015) Crisis in the Archives, Is It the End of History as We Know It?, https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/370930/RECORDS_REVIEW-Sir_Allex_Allan.pdf.

Allison, A., Currall, J., Moss, M. and Stuart, S. (2005) Digital Identity Matters, *JASIST*, **56** (4), 364–72.

Beale, P., Almond, A. and Archer, M. S. (eds) (2011) *The Corsini Letters*, Amberley.

Brabazon, T. (2007) *The University of Google: education in the (post) information age*, Ashgate.

Brabazon, T. (2014) Review of *Mind Change: how digital technologies are leaving their mark on our brains*, by Susan Greenfield, *Times Higher Education Supplement*, 28 August.

Clarke, P. and Zutshi, P. N. R. (eds) (2014) *Supplications from England and Wales in the Registers of the Apostolic Penitentiary, 1410–1503*, Boydell & Brewer.

Coppersmith, J. (2015) *Faxed: the rise and fall of the fax machine*, Johns Hopkins

University Press.
Currall, J., Johnson, C., Johnston, P., Moss, M. and Richmond, L. (2002) *No Going Back? The final report of the Effective Records Management Project*, https://dspace.gla.ac.uk/handle/1905/19.
Erdos, D. (2013) Mustn't Ask, Mustn't Tell: could new EU data laws ban historical and legal research? UK Constitutional Law Blog, http://ukconstitutionallaw.org.
Gillen, J. (1999) Reflexivity, Consciousness and Linguistic Relativity: an attempted link, http://www.did.stu.mmu.ac.uk/cme/Chreods/Issue_13/JGillen.html.
Gollins, T. (2009) Parsimonious Preservation: preventing pointless processes! (The small simple steps that take digital preservation a long way forward), http://www.nationalarchives.gov.uk/documents/information-management/parsimonious-preservation.pdf.
Greenfield, S. (2014) *Mind Change: how digital technologies are leaving their mark on our brains*, Rider Books, Random House.
Hernon, P. and Matthews, J. R. (2013) *Reflecting on the Future of Academic and Public Libraries*, Facet Publishing.
Kittler, F. (1999) *Gramophone, Film, Typewriter*, Stanford University Press.
Lynch, C. A. (2001) Colliding with the Real World: heresies and unexplored questions about audience, economics, and control of digital libraries. In Bishop, A., Butterfield, B. and Van House, N. (eds), *Digital Library Use: social practice in design and evaluation*, MIT Press.
McKemmish, S. (1997) Yesterday, Today and Tomorrow: a continuum of responsibility. In *Proceedings of the Records Management Association of Australia 14th National Convention, 15–17 September 1997*, Records Management Association of Australia.
McLuhan, M. (1964) *Understanding Media: the extensions of man*, McGraw-Hill.
Moss, M. (2013) Where Have All the Files Gone? Lost in action points every one? *Journal of Contemporary History*, **7** (4), 860–75.
Moss, M. (2015) Understanding Core Business Records. In Turton, A. (ed.), *International Handbook of Business Archives Handbook*, Ashgate.
Pacioli, L. (1494) *Summa de Arithmetica, Geometria, Proportioni et Proportionalita* (Venice).
Prensky, M. (2001) Digital Natives, Digital Immigrants, *On the Horizon*, MCB University Press, **9** (5), 1–6.
Redwine, G., Barnard, M., Donovan, K., Farr, E., Forstrom, M., Hansen, W., John, J., Kuhl, N., Shaw, S. and Thomas, S. (2013) *Born Digital: guidance for donors,*

dealers, and archival repositories, CLIR Publication 159.

The National Archives (2010) *Migrating Information between EDRMS*, TNA, Kew, http://nationalarchives.gov.uk/documents/information-management/edrms.pdf.

Verbeek, P.-P. (2010) *What Things Do: philosophical reflections on technology, agency, and design*, Penn State University Press.

Notes

1 www.fugger.de/en/business/organisational-chart.html.
2 www.telegraph.co.uk/technology/news/8753784/The-300m-cable-that-will-save-traders-milliseconds.html.
3 www.queenmaryhistoricalresearch.org/roundhouse/default.aspx.
4 https://projectabaca.wordpress.com/publications/ and www.bitcurator.net/.
5 http://accessdata.com/solutions/digital-forensics/forensic-toolkit-ftk.