The Challenge of the Digital and the Future Archive: *Through the lens of The National Archives UK*

Authors

1. **Dr Eirini Goudarouli**, Digital and Technology Research Lead; **Corresponding Author:** Eirini.Goudarouli@nationalarchives.gov.uk

The National Archives, Kew, Bessant Drive, TW9 4DU, UK

2. **Dr Anna Sexton**, Head of Research.

The National Archives, Kew, Bessant Drive, TW9 4DU, UK.

3. **John Sheridan**, Digital Director.

The National Archives, Kew, Bessant Drive, TW9 4DU, UK.

Abstract

On the 7th of June 2018, The National Archives UK held its inaugural Digital Lecture, delivered by Professor Luciano Floridi entitled "Semantic Capital: What it is and how to protect it". The lecture was followed by a poster exhibition, showcasing nine cutting-edge digital research projects at The National Archives.[1]

This paper aims at giving a distinct overview of The National Archives' digital research priorities, drawing on examples from the active and recently completed research projects, which were displayed at the exhibition on the 7th of June 2018. The focus of this paper is to discuss the research challenges that we are facing as we seek to become a second-generation digital archive, that is digital by instinct and design. By placing a particular emphasis on the conceptual and epistemological challenges relating to *trust* and *openness*, the paper suggests that research is the key for us as a rapidly evolving digital archive; enabling us not only to inform but also innovate around the forthcoming digital challenges, and helping us to define future directions and lead the shaping of the future archive.

Keywords

Digital, Challenge, Future Archive, Trust, Transparency, Openness

[1] You can listen to the lecture's podcast and see the posters at: https://media.nationalarchives.gov.uk/index.php/digital-lecture-semantic-capital/ Accessed on 12 July 2018.

1. Introduction

The National Archives is the official archive of the UK government, leader of the archive sector in England and a cultural heritage institution. It is also a recognised Independent Research Organisation (IRO) by the Arts and Humanities Research Council, which means that it is eligible to lead, or be equal partners and collaborators in, research proposals submitted to any of the UK's seven research councils for funding. We, at The National Archives, have recently published our research priorities (The National Archives 2018), and as we draw on our in-house multi-disciplinary expertise to deliver around these, one of our strategic goals is to collaborate with academic and research communities across different sectors and disciplines, both nationally and internationally. We aim to create a research environment which will help us address the major questions and challenges that are fundamental to us as a national digital archive.

Archives Inspire, The National Archives' business strategy (2015 – 19), identifies 'digital' as our biggest challenge (The National Archives, 2015). However, we are not alone in this. In a recent short article, Professor Luciano Floridi (2017) discusses the unsustainable fragility of our increasingly complex environments, which is only one facet of the challenge that digital brings into modern societies. In that piece, Floridi presents ways of recognising where the risk of digital fragility resides and suggests solutions by highlighting the crucial role that all members of the digital world, from the top to the bottom (that is, from the huge corporations and organisations to single users), can play in protecting the information society. The article concludes by arguing that, on the bright side, digital fragility has something particularly useful to offer. It pushes our information society to understand its vulnerabilities and work collaboratively on becoming more prepared to take control of (or even prevent) digital failures by being, from the top to the bottom, more open, more accountable and more responsible.

In the world of archives, the intangible record is fundamentally changing the landscape as well as the role of archivists and archival institutions. The emergence of new generation technologies is one factor that is rapidly leading to an epistemological shift in archival science, or, to put it in Thomas Kuhn's words to a *scientific revolution* (Kuhn 1962; Cook 2000), by moving from a relatively settled scientific framework to the urgent need for a profound change to its principles, methods and practices (James, Johnson, Hunter, 2018). More specifically, the digital brings a variety of complexities and challenges to our archival

frameworks, requiring new capabilities and approaches on how best to capture, preserve, contextualise and present the increasingly born-digital record. Technological shifts also bring new opportunities, for example, in unlocking large-scale collections (either physical, digitised or born-digital) for research and experimentation, by enabling the extraction of their content as data and moving towards the creation of aggregated large-scale datasets.

However, as the paradigm is shifting, archival science is not moving from the relatively steady principles and practices of the discipline to the shaping of a new state of "normal" science. Archival science is moving from relative stability to continual change. Therefore, the changing nature of the increasingly complex digital environment makes the challenge for archives a very demanding on; it is a challenge that can never be completely solved (The National Archives, 2017c, 6).

Trust has always been central to archives. However, as the nature of records and archives evolves more quickly and the digital contests long-standing archival practices, *trust* comes to the forefront of the discussion. One of the major questions related to this, is how archives retain the legitimacy they confer on the digital evidence they capture, preserve, contextualise and present. Again, and similar to the case of protecting the information society from digital fragility, *accountability* is key in order for archivists as well as institutions to retain their confidence in the integrity, reliability and authenticity of the archival records that they preserve and present, and in order to remain trusted custodians of the future archive in the eyes of the public (The National Archives 2017c, 6).

In what follows, we will discuss the conceptual and epistemological challenges relating to *trust* and *openness* in rapidly evolving digital archives. As we introduce a synthesis of our digital research projects, we will argue that research plays a central role not only to inform but also to innovate around these challenges, helping to define future directions and lead to the shaping of the future archive.

2. Embracing the evolving nature of the digital archive

The National Archives is one of the few fully functioning digital archives in the world (The National Archives, 2017a, 5). Our digital repository is capable of safely, securely and actively preserving enormous quantities of data, whilst our search and access are provided through *Discovery*, our digital catalogue, in which records are described using long-established catalogue standards. However, we are still a 'first generation' digital archive, digitally

simulating a long-established body of archival practices devised for physical records. If we are to become a second-generation digital archive, that is digital by instinct and design, we must find ways to successfully disrupt and adapt entrenched practice (The National Archives, 2017a, 6). To ensure public trust and openness, and make the most of the new opportunities afforded to us, we must strive to maintain some of the long-held underpinning principles around our practice (accountability and transparency) and the record (integrity, authenticity and reliability); but we must do so through the application of radically different methods to those that enabled the maintenance of these principles in the analogue world. The *why* we seek to maintain a trustworthy and open public record has not changed, and the principles outlined above also maintain their relevance, but the scale and depth of the challenges we are now facing means that every element of *how* we manage records and archives needs to adapt and shift. In this context, we, along with archives across the UK and beyond, must radically reimagine our practice to meet the evolving challenges of preserving digital records and providing access to them, and we must be open to our practice speaking back to, and shifting our theory. To do this, we seek to position our research agenda at the forefront of this shifting ground to create and engage with emerging concepts of the record and recordkeeping (The National Archives, 2018).

Undoubtedly, the use of emergent technologies, such as Snapchat, Google Docs, neural networks, blockchains, hashing algorithms, cryptography and the cloud have profoundly altered the nature of archives, by disrupting how information is created, recorded, captured, encoded, curated, shared, made available and used. These shifts require fundamentally new capabilities and approaches on how best to capture, preserve, contextualise and present increasingly digital public record. Therefore, the archivist's relationship to emergent technologies needs to become multi-layered. We need to understand the digital landscape, and the changing nature of how society creates and shares records in the light of new generation technologies, and be willing to apply these new advanced technologies in our archival response to these changes. This sparks a new era for archives, as today's archivist must become equipped with emergent technologies as their own tools of the trade.

At The National Archives, as we rethink the nature of the record and our archival practices around the record in the light of digital, we combine practical considerations with explorative research into infrastructure, methodology, tools, techniques and user requirements, drawing on innovation across cultural heritage, academia and relevant

industries. Our vision to become a fully functioning second-generation digital archive is only achievable by embedding new generation technologies in our recordkeeping practices to help us manage our rights and responsibilities as we go about capturing, preserving, contextualising and presenting digital records.

In an era of increasing dependency on Artificial Intelligence (AI), ethical implications relating to digital archival preservation, description, representation and use need to be considered thoroughly. We see a future where AI and emergent technologies become part of our everyday recordkeeping practices, and in their use, questions around *trust* and *openness* become fundamental to us. For example, increasing societal dependency on decision making by algorithms is fundamentally shifting the evidential landscape. AI that uses data to create and train algorithms, which are then applied to new data, is becoming increasingly more widespread and multifarious. At the more complex end of the spectrum, where neural networks operate in a dynamically evolving and interactive environment, it becomes challenging to be transparent about the basis on which outcomes are determined. The uncertain and unbounded nature of AI requires us to entirely rethink how we preserve evidence of both the system, and the decisions made. Therefore, AI also requires us to rethink what the 'record' is in this context.

In addressing challenges like these, we initiate cutting-edge, high-quality research that impacts upon collections management, archival science, the archive sector and interdisciplinary discourse around the archive. In our research, we seek to uncover innovative new models and methods, technologies and tools, to manage the risks bound up in recordkeeping in a digital age, negotiating a complex set of legal, ethical and social boundaries.

2.1. Risk, Uncertainty and Trust

Digital transformations have altered the traditional boundaries of the archive as an authoritative institution and created new challenges around safeguarding records and data entrusted to them. In a rapidly shifting recordkeeping landscape, understanding and effectively managing all relating risks lie at the heart of our work.

Building auditability and transparency with Distributed Ledger Technology

One of the fundamental challenges associated with born-digital records is how we can guarantee that the record we archive today is the same one we release to the public in 20 or even 100 years' time. We are experimenting with Distributed Ledger Technology (DLT), such as blockchain, as a tool to maintain trust in digital records through time. DLT uses cryptographic techniques and a distributed storage model to create verifiable proof that an archived object has not been tampered with by producing multiple openly available instances of that proof. This means that if a digital file has been changed in some way it becomes impossible to hide the fact that it is now different to when it was registered in the distributed ledger.

This is an enticing feature of DLT, but digital archiving soon creates a new issue. The original archived object is rarely going to be the one presented to the user of the archive. In the analogue world, we often provide digitised facsimiles of a paper document, and similarly in the digital world we will provide redacted and reformatted versions of the original digital document. A word processor document in an obsolete format may be converted to a PDF for compatibility purposes; individual emails could be extracted from an email archive; high definition video can be reduced to a compressed format to minimise impact on download bandwidth. Each of these procedures, whilst not fundamentally affecting the content of the object, will still create something cryptographically different to the original.

As we focus on ensuring the sustainability of digital archive, we seek to build a system of auditability and transparency into the digital archive with DLT at the centre, paying particular attention to the traceability of the functions used to manipulate original digital objects. The core principle is that each function, used to manipulate a digital file, is itself registered in the distributed ledger, and only verifiable functions may be used. This way a presentation copy of a digital object can be traced back to its original version via the set of all the functions used to create or manipulate it throughout its lifespan. We believe that using this technology could transform the sustainability of digital archives, but also enable archives to share stewardship of the records and, by sharing, guarantee the integrity of the records they hold (Collomosse et al. 2018a; Collomosse et al. 2018b, 7; Green 2018).

Using Bayesian methodology to deal with preservation risks for digital archives

Shifting to a second-generation digital archive requires a change in our approaches to measuring and managing digital preservation risks. In a constantly changing digital

environment, our current qualitative approaches to risk assessment cannot provide us with adequate assurance.

As we seek to trial a statistical approach to quantify our exposure to digital preservation risk, we have identified a Bayesian approach as being potentially well suited to addressing this problem. Digital preservation is a long-term activity, and many risk events and failures occur too infrequently for us to have collected much hard data to inform an empirical approach. The practice of robust measurement to support and evaluate decision making in digital archives is still maturing. These factors can make it difficult to apply quantitative techniques. However, the profession does have a wealth of experience developed over two decades of digital preservation and data is available from industry and from archives pooling their knowledge.

Building a predictive model based on Bayesian approach allows us to combine different sources of information, using hard data where it is available, and supplementing it with our best judgement where it is not. This approach also allows us to operate in the face of uncertainty and imperfect information, offering the potential to refine our views as new data becomes available, and that can accommodate a shifting environment and our changing risk appetite. We are seeking to build a new model of digital preservation risk, which will be grounded in data, but flexible enough to accommodate our changing understanding. It will encompass a wide range of threat factors, looking beyond format diversity to system dependencies, software, technical skills and organisational policies.

Undertaking risk modelling enables us to develop a robust cost model for our preservation functions. This means we can effectively plan our interventions balancing likelihood of risk against the resource implications. Through predictive risk modelling, we will be able to ground our decision-making around preservation actions empirically, whilst being able to articulate the benefits and costs of different preservation options. Drawing on expertise across our conservation and digital teams, we seek to validate our developing model through the application of risk scenarios, and use this work to develop research collaborations to expand the model further.

We are particularly interested in exploring questions on how we can develop statistical models that help us understand and quantify the likely impact of preservation actions and invest resources to greatest effect as well as how these risk models can be relevant to the wider archive sector. Moreover, answering these questions will help us to map and negotiate

the complex legal, social and ethical landscape that sits around the provision of access to our holdings. Articulating risk and developing an understanding of where our 'risk appetite' needs to sit, is an important step in developing a workable infrastructure around the presentation of our digital records (Mulinder et al. 2018, 3).

Tackling Uncertainty in large-scale historical collections with probabilistic techniques

One of our major questions is how to effectively deal with ambiguity and uncertainty in and around our collections, whilst we create fuzzy-linking approaches for historical data that is messy, incomplete and inconsistent.

To answer this question, we have been actively experimenting with probabilistic techniques for linking the people and places who appear in our records. By developing algorithms and statistical models for probabilistic linking of messy historical data, we are seeking to create new routes for navigation across our collections and those of other archives (Bell, Hillyard, Ranade 2018, 5; Ranade 2016; Bell and Ranade 2015). The links we have published through this experimental research have not been a curated assertion of authoritative truth but rather an expression of confidence in the likelihood of an entity match. Instead of trying to cleanse and standardise data, our techniques aim at leveraging uncertainty by quantifying and working around the 'fuzziness' found in our large-scale collections. Our aim is to enable the user to make robust, data-driven access decisions (Bell and Ranade 2017).

The use of probabilities in our record linkage algorithm opens new access routes for heritage collections and offers a practitioner's perspective on the implications for archival description. Our initial work to create links across our collections is also open for extension as now, instead of a catalogue of boxes of paper, connected only by their archival context, we can create a connected web of people, places, events, and ideas. However, there is more work to be done to extend and develop our probabilistic methods in order to manage ambiguous big data and present results back to users as an expression of confidence. As Bell and Ranade explore in a recent magazine article 'we still need to explore methods for making these links accessible to new and old audiences, combining our advances in linking data with modern visualisation and data mining techniques to deliver new insights' (2017, 30).

In a world where archival data is now metadata, another aspect of uncertainty sits with the implication of volume of digital (either digitised or born-digital) collections and their

archival descriptions. We are committed to implement automatic generation of item level metadata around our born-digital records, moving us away from the provision of a manually generated descriptive layer over our holdings. Understanding the implications of this shift, in the broader context of evolving standards for archival description and related schemas and ontologies, will help us explore questions of authority and quality assurance around automatically generated metadata and lead us to explore new ways of helping our users to engage with and interrogate uncertain description (The National Archives, 2017b).

2.2. **Openness, Access and Transparency**

Opening our collections to our users is a vitally important part of The National Archives' remit, and this brings many challenges and opportunities. As custodians of an ever-increasing volume of digital records, both digitised and born-digital, we develop new methods of (i) providing access to data for use and reuse on an unprecedented scale, securely and within the law, and (ii) being transparent on the archivists' decisions and processes involved in the development of new services and innovative models of content delivery.

Enabling big and complex data analysis, by transforming the way we think about access

Traditional concepts of archival use mainly reflect the time when users conducted document-based research in reading or research rooms. Today, the reading room is still a physical space where users go to predominantly explore analogue records and material artefacts but, equally, there are various digital spaces where digital records are produced or reside, such as websites, online catalogues or other digital infrastructures of archives, cultural institutions and other corporations.

Undoubtedly, digital has transformed accessibility to archival collections, bringing new challenges as well as opportunities to the forefront. Rethinking the ways of opening up our records in the light of digital, our research programme has led to identifying new methods of accessing our collections through innovation in our digital catalogues and other digital services, creating new opportunities in collection, analysis, use and reuse of our content.

More specifically, as the archive of the UK Government, and official publisher of UK legislation and The UK Government Web Archive (UKGWA), The National Archives is in a position of being able to contextualise and explore the government record of today against the backdrop of over a thousand years of history interwoven with an evolving governmental

and legal system. Being in this unique position, we are particularly interested in representing UK legislation and UKGWA as data that are open and accessible for research. As we seek to innovate in access, analysis and reuse of our content, we are developing the digital infrastructure to unlock our content and enable the application of data analytic approaches. The fundamental challenge that we face here links to how best to capture, collect, preserve and present the vast size of information relating to UK legislative activity as well as to the government online presence, when the content of both is quickly and constantly changing; only in the Statute Book, 100,000 words are added or changed every month. Let alone, the ongoing changing activity from the government social media and online presence.

During our investigation around how to best provide access to UK legislation and make its content available as a resource for research, we created new ways to enable advanced research across complex large-scale collections, by applying techniques such as Natural Language Processing and structure-aware search. In particular, we developed a new website for researchers to use, new datasets such as the first core reference dataset listing all UK legislation over 800 years, and a suite of advanced tools that make it easy to query legislation content and enable structurally and temporally aware searches (Sheridan and Riley, 2018, 1).

Additionally, as we move our digital services to the cloud, we are also interested in making the UKGWA easily accessible to researchers who undertake large-scale research on our archived government websites, government department videos and social media activity. Combining Optical Character Recognition (OCR) techniques with new advanced technologies for capturing and preserving UK central government information published on the web, we aim at enabling our large-scale digital collection (including digitised documents in PDF format as well as born-digital) available for research and exploration (Storrar, Newing and Feissali 2018, 9; Goudarouli, Winters 2018).

In this context, we seek to enhance and showcase the relevance of our collections to highlight the contribution they can make to contemporary discourse, and to expose the richness of their contents in illuminating past and present debates. We seek to explore the relationship between records and the legal system through research that analyses legislative impacts on our own organisational role, and the role of archives across the wider sector. We also seek to gather evidence on the functioning of the current legislative frameworks that underpin the preservation, access, use and reuse of records and make recommendations for their future direction in support of public policy objectives.

Unlocking large-scale collections with the use of algorithms and AI

The National Archives holds approximately 200km of paper documents in its repositories, including millions of pages of digitised records documenting 1,000 years of human history and culture. The scale of our collection is one of our challenges when it comes to opening our collections, both digitised and digital-born, for research and experimentation.

When the challenge of scale is combined with questions of how to transcribe printed and handwritten historical records, issues around access and use of data need new approaches and interpretation. Recent advances in technology offer us a wealth of new opportunities to unlock the printed as well as handwritten historical content within our records in new ways for our users. In particular, by combining OCR and Handwritten Text Recognition (HTR) with high-performance computational techniques, we have the opportunity to explore how the presentation of, access to, and analysis of our complex large-scale collections could be enhanced through extracting the content of our tangible records as data in order to move us beyond traditional methods of digitisation. In relation to handwritten historical content, one of the challenges of applying HTR techniques is accuracy. This opens up questions around how best to integrate HTR with (potentially crowd-sourced) human intervention to produce transcriptions of content at scale. As we open up the potential to transform our records into data, we aim to enable interdisciplinary research that manipulates that data in a multiplicity of ways to trace patterns across and in records, recombining our data to tell new stories, and finding new ways to navigate and visualise our holdings (Dunley 2018a; Dunley 2018b, 6; Liem et al. 2018; Goudarouli and Alexander 2018; Goudarouli and Perin 2018).

Furthermore, when the focus moves from unlocking our large-scale historical analogue collections to the future challenge of providing access to a large-scale born-digital collection, the challenge of access brings questions around data protection and appropriate levels of openness to the forefront. We are currently transitioning from a 30-year to a 20-year rule for receiving records of enduring value from government departments. This essentially means that Government records selected for permanent preservation are now being transferred to us 20 years after creation. We are very aware that although we are currently receiving a relatively slow trickle of born-digital content reflective of a 1980s and 1990s ‘print to paper’ culture, the trickle of born-digital records will rapidly become a tsunami as we take in Government transfers from the turn of the century onwards. As this change impacts our

reality, the born-digital records will increasingly dominate our attention, and repositories. There are many complex challenges associated with managing and providing access to born-digital large-scale collections. For example, our current processes for reviewing records for sensitive content before release are manual, a status quo which is neither sustainable nor applicable in the digital-born context. To cope with the volume of born-digital records that will come our way, we need to develop and use algorithms to help human reviewers to make important decisions around record selection, retention, closure and openness. We are also looking to develop a presentation system that can enable gradated access to born-digital content, ideally with the capability to automate the suppression of content in the light of legal and related access constraints. We therefore anticipate a future that involves an increasing reliance on algorithms to aid our processes, from preservation through to presentation. It is fundamentally important that we pay attention to ensuring fairness, accountability and transparency throughout all the recordkeeping processes which involve algorithms and AI-assisted decision-making.

In an era of both largescale data analytics and AI-assisted recordkeeping science, opening the 'black box' of AI systems will also help us address important practical and ethical issues, which area are linked to the uncertainties and biases in training data and statistics used throughout those processes. The understanding of all the stages of the processes and all levels of dependencies occurred from the use of algorithms and AI in decision-making will consequently help us raise awareness about the ethical implications sitting in this context. This will also help us not only securely opening our collections for complex analysis and research but also improve the delivery of services by ensuring that transparency is at the centre of our practices.

3. Summary: Towards a *Trustful* and *Open* Future Archive

Archives have a meaning and value that goes beyond content; they enrich society and contribute to our cultural heritage. In a rapidly evolving complex digital environment, archive institutions such as The National Archives seek to move towards a sustainable digital future and to remain trusted custodians of our collective memory. Our remit to provide access to an untampered and reliable public record remains constant as our context shifts. Collecting institutions around the world are now focusing on building their capacity in order to maintain the legitimacy they confer on the evidence they capture, preserve, contextualise and present.

In an era of AI-assisted recordkeeping, the need to demonstrate the adaptability, value and sustainability of archives has never been more acute. The article suggests that the challenge of digital brings *trust* into the spotlight, by also raising questions of *openness*, whilst it offers an opportunity to shift the basis of trust, from the authority of the archive as an institution to transparency and accountability of archival practice (The National Archives, 2017c, 6).

The main question is: what needs to be done today to ensure that archives are trusted and open in the digital future, and within a rapidly and constantly evolving complex digital environment? This article suggests that high-quality research, experimentation and collaboration is the answer. Research enables us to unlock the black box of the digital, understand and explore the conceptual and methodological challenges and the ethical implications that the digital brings to our understanding of the record and the archival context and suggest new ways to become more accountable and more transparent for the current and future benefit of government and citizens.

ACKNOWLEDGEMENTS

We would like to thank our colleagues Alex Green, Digital Preservation Services Manager, and Dr Sonia Ranade, Head of Digital Archiving, at the National Archives UK for providing insight and expertise that greatly assisted parts of this paper.

REFERENCES

Bell, M., Ranade, S. (2015). *Traces through time: a case-study of applying statistical methods to refine algorithms for linking biographical data*, In Serge ter Braake, Antske Fokkens, Ronald Sluijter, Thierry Declerck and Eveline Wandl-Vogt (Ed.), Proceedings of the First Conference on Biographical Data in a Digital World 2015, pp. 24 – 32. http://ceur-ws.org/Vol-1399/paper5.pdf Accessed 12/07/2018.

Bell, M., Ranade, S. (2017). *Digital Technologies for a Digital National Archive*, Archives and Record Association, 335, 29 – 30.

Bell, M., Hillyard, M., Ranade, S. (2018). *Traces Through Time*. Poster displayed at the poster exhibition at the first annual Digital Lecture at The National Archives, page 5. https://www.nationalarchives.gov.uk/documents/digital-projects-at-the-national-archives.pdf Accessed 12/07/2018.

Collomosse, J., Bui, T., Brown, A., Sheridan, J., Green, A., Bell, M., Fawcett, J., Higgins, J., Thereaux, O., et al. (2018a), ARCHANGEL: Trusted Archives of Digital Public Documents. Accepted article. The 18th ACM Symposium on Document Engineering. https://arxiv.org/abs/1804.08342v1 Accessed 12/07/2018.

Collomosse, J., Bell, M., Green, A., Mulinder, A., Sheridan, J., et al. (2018b). *Using blockchain to engender trust*. Poster displayed at the poster exhibition at the first annual Digital Lecture at The National Archives, Page 7. https://www.nationalarchives.gov.uk/documents/digital-projects-at-the-national-archives.pdf Accessed 12/07/2018.

Cook, T. (2000). Archival Science and Postmodernism: New Formulations for Old Concepts. *Archival Science: International Journal on Recorded Information*, 1(1), 3– 24.

Dunley, R. (2018a). Machines reading the archive: handwritten text recognition software. Blogpost. https://blog.nationalarchives.gov.uk/blog/machines-reading-the-archive-handwritten-text-recognition-software/ Accessed on 31/07/2018.

Dunley, R. (2018b). *Can machines read handwriting?* Poster displayed at the poster exhibition at the first annual Digital Lecture at The National Archives, page 6. https://www.nationalarchives.gov.uk/documents/digital-projects-at-the-national-archives.pdf Accessed 12/07/2018.

Floridi, L. (2017). The Unsustainable Fragility of the Digital, and What to Do About It. *Technology and Philosophy*, 30(3), 259 – 261.

Goudarouli, E., Alexander, A., (2018) Beyond words: exploring historical document collections at scale. Blogpost.

https://blog.nationalarchives.gov.uk/blog/beyond-words-exploring-historical-document-collections-scale/
Accessed on 30/07/2018.

Goudarouli, E., Perin, C., (2018) Exploring new ways of visualising The National Archives' First World War diaries. Blogpost.
https://blog.nationalarchives.gov.uk/blog/data-visualisation-first-world-war-diaries/
Accessed on 12/07/2018.

Goudarouli, E., Winters, J., (2018) Opening up web archives to research and innovation. Blogpost.
https://blog.nationalarchives.gov.uk/blog/opening-web-archives-research-innovation/
Accessed on 21/09/2018.

Green, A., (2018), Trustworthy technology: the future of digital archives. Blogpost.
https://blog.nationalarchives.gov.uk/blog/trustworthy-technology-future-digital-archives/
Accessed on 31/07/2018.

James, J., Johnson, V., Hunter I. (2018), Archives Matter. *Insights*, 31, 26,
http://doi.org/10.1629/uksg.417
Accessed on 30/07/2018.

Kuhn, T. S. (1962). *The Structure of Scientific Revolution*. Chicago: University of Chicago Press.

Liem, J., Goudarouli, E., Hirschorn, S., Wood, J., Perin, C., et a. (2018). Conveying Uncertainty in Archived War Diaries with GeoBlobs. Accepted and Awarded the *IEEE InfoVis* Best Poster 2018 award. IEEE VIS 2018.
http://vu2044.serv01.menkisys.de/geoblob/geoblob_poster_abstract_vis2018.pdf
And
http://vu2044.serv01.menkisys.de/geoblob/geoblob_poster_vis2018_A0.pdf
Accessed 20/09/2018.

Mulinder, A., Clipsham, D., Underdown, D., Waddington, S., Ranade, S., et al. (2018). *A Bayesian model of preservation risk*. Poster displayed at the poster exhibition at the first annual Digital Lecture at The National Archives. Page 3. https://www.nationalarchives.gov.uk/documents/digital-projects-at-the-national-archives.pdf Accessed 12/07/2018.

Ranade, S., (2016). Traces through Time: A probabilistic approach to connected archival data. Paper presented. IEEE Big Data 2016: 1st CAS workshop. https://ieeexplore.ieee.org/document/7840983/
Copy available at:
http://dcicblog.umd.edu/cas/wp-content/uploads/sites/13/2016/05/2.pdf Accessed 20/09/2018.

The National Archives. (2015). Archives Inspire. http://nationalarchives.gov.uk/about/our-role/plans-policies-performance-and-projects/our-plans/archives-inspire/ Accessed 12/07/2018.

The National Archives. (2017a). Digital Strategy. https://www.nationalarchives.gov.uk/about/our-role/plans-policies-performance-and-projects/our-plans/digital-strategy/ Accessed 12/07/2018.

The National Archives (2017b). Digital Cataloguing Practices. http://www.nationalarchives.gov.uk/documents/digital-cataloguing-practices-march-2017.pdf Accessed 12/07/2018.

The National Archives. (2017c). Archives Unlocked. http://www.nationalarchives.gov.uk/documents/archives/Archives-Unlocked-Brochure.pdf Accessed 12/07/2018.

The National Archives (2018). Research Priorities. http://www.nationalarchives.gov.uk/about/our-research-and-academic-collaboration/our-research-and-people/our-research-priorities/ Accessed 12/07/2018.

Sheridan, J., Riley, J. (2018). *Big Data for Law*. Poster displayed at the poster exhibition at the first annual Digital Lecture at The National Archives. Page 1. https://www.nationalarchives.gov.uk/documents/digital-projects-at-the-national-archives.pdf Accessed 12/07/2018.

Storrar, T., Newing, C., Feissali, K. (2018). *Webrecorder: archiving the un-archiveable*. Poster displayed at the poster exhibition at the first annual Digital Lecture at The National Archives. Page 9. https://www.nationalarchives.gov.uk/documents/digital-projects-at-the-national-archives.pdf Accessed 12/07/2018.