Introduction [Archives, Access and Artificial Intelligence]

PLEASE CITE THE PUBLISHED VERSION

https://www.transcript-verlag.de/978-3-8376-5584-1/archives-access-and-artificial-intelligence/

PUBLISHER

Bielefeld University Press / transcript Verlag

VERSION

VoR (Version of Record)

REPOSITORY RECORD

Jaillant, Lise. 2022. “Introduction [archives, Access and Artificial Intelligence]”. Loughborough University. https://hdl.handle.net/2134/14394131.v1.

Introduction

Lise Jaillant, Loughborough University, UK

Digital archives are transforming the humanities and the sciences. Digitized collections of newspapers and books have pushed scholars to develop new, data-rich methods. Born-digital records ("items created and managed in digital form"[1]) are now better preserved and managed thanks to the development of open-access and commercial software. Digital humanities have moved from the fringe to the center of academia. Yet, the path from the appraisal of records to their analysis is far from smooth.

Cultural heritage organizations face at least three main challenges. First, the volume of digital archives makes it extremely difficult for archivists to assess records. Applying Artificial Intelligence (AI) and machine learning (ML) to archives is still at an experimental stage, but AI/ ML could become an integral part of archival processes.[2] To manage the sheer bulk and potential sensitivity of records, archivists will also rely on creators to help them make appraisal and selection decisions at the point of deposit.

Second, most born-digital collections are currently closed due to a wide range of reasons (including technical issues, copyright, and data protection). Regardless of whether archives are digital or not, archivists need to balance individual rights and the public interest in the context of the General Data Protection Regulation (GDPR) in Europe. Nobody would reasonably claim that all born-digital data should be unlocked and openly accessible. Yet, it is important to recognize that "dark" archives contain vast amounts of data essential to scholars – including email correspondence, drafts of manuscripts, digital photos and videos. Within current legal frameworks, making born-digital archives more accessible is an urgent priority to fully make sense of our cultural heritage.

---

[1] Ricky Erway, Defining "Born Digital," OCLC Research, November 2010, URL: https://www.oclc.org/content/dam/research/activities/hiddencollections/borndigital.pdf [last accessed: Mar. 29, 2021].

[2] Artificial Intelligence (AI) is a large concept designating the creation of intelligent machines that can simulate human thinking capability and behaviour. Machine Learning (ML) is an application or subset of AI that allows machines to learn from data without being programmed directly. In practice, the terms "AI" and "ML" are often used interchangeably.

Third, data science and AI are becoming essential tools, but very few scholars (particularly in the humanities) have been trained to master these research methods, a skills gap which in turn has an impact on the training we offer to students. This is a central topic of a recent White Paper from the Alan Turing Institute on *The Challenges and Prospects of the Intersection of Humanities and Data Science*. According to Barbara McGillivray and her co-authors, "the challenge here is to find a way to train and upskill humanities researchers in quantitative and computational methods, while at the same time incorporating the basic principles from these methods throughout undergraduate and graduate degrees, so humanities graduates are well equipped to lead projects but also potentially undertake careers in research software engineering and data science for arts and humanities." The authors suggest setting up basic courses in data science and software engineering to "offer the foundational skills to support humanists in having structured and informed conversations with computer scientists and data scientists needed in interdisciplinary projects."[3]

Automation, Access and AI are becoming keywords to decipher our history. We do not suffer from a lack of records, but from too many records – often locked away in dark archives. Access to dark archives is central but needs to be complemented with data-rich methodologies. How can we shed light on born-digital and digitized archives? How can we give greater access to archives currently closed to the public? What is the role of automation and AI? *Archives, Access and AI* addresses these central questions and explores crossovers between various disciplines to improve the discoverability, accessibility and use of born-digital archives and other cultural assets.

1. *Applying AI to Archives*

Archivists have commented on the digital revolution and its impact on archives for the past three decades. But it was not until the mid-2000s that the scholarship on digital preservation started growing. In addition to the preservation of digital materials,[4] commentators have examined the impact of the digital revolution on

---

[3] Barbara McGillivray et al., *The Challenges and Prospects of the Intersection of Humanities and Data Science: A White Paper from The Alan Turing Institute*, London 2020, see 21, doi:10.6084/M9.FIGSHARE.12732164.

[4] From 2005 to 2007, the UK funder Jisc supported the PARADIGM (Personal Archives Accessible in Digital Media) project, undertaken by the Bodleian Library in Oxford and the John Rylands Library in Manchester. The overall aim was to examine the issues in preserving personal digital materials, and to produce best-practice guidelines. In 2007, the Arts and Humanities Research Council (AHRC) funded the two-year Digital Lives project, led by the British Library in partnership with University College London and the University of Bristol.

appraisal.[5] Focusing on the preservation of born-digital and digitized records, or on the selection of these records, is not enough. Access and the production of new knowledge are issues that need to move to the center of the scholarly debate. In particular, Artificial Intelligence can be used by archivists to identify sensitive records, but also by researchers to process large amounts of digital archival data. AI has the potential to transform archives, but it also brings new challenges (including ethical challenges).

The closure of libraries, archives and museums due to the COVID-19 pandemic has highlighted the urgent need to make archives and cultural heritage materials accessible in digital form. Yet too many born-digital and digitized collections remain closed to researchers and other users due to privacy concerns, copyright and other issues. Born-digital archives are rarely accessible to users. For example, the archival emails of the writer Will Self at the British Library are not listed on the Finding Aid describing the collection, and they are not available to users either onsite or offsite. At a time when emails have largely replaced letters, this severely limits the amount of content openly accessible in archival collections. Even when digital data is publicly available (as in the case of web archives), users often need to physically travel to repositories to consult web pages. In the case of digitized collections, copyright can also be a major obstacle to access. For instance, copyright-protected texts are not available for download from HathiTrust, a not-for-profit collaborative of academic and research libraries preserving 17+ million digitized items (including around 61% not in the public domain).

The primary aim of the project was to develop ways to secure the personal archives of individuals in the digital era. In 2008, the Andrew Mellon foundation funded the futureArch project at the Bodleian Library to find solutions to the problem of born-digital but also hybrid archives (composed partly of paper materials). In particular, Bodleian Electronic Archives and Manuscripts (BEAM) worked on digital preservation infrastructure and researcher interfaces for hybrid archives. The 2010s saw the development of guidelines to preserve email archives. See Christopher J. Prom, *Preserving Email - DPC Technology Watch Report*, Digital Preservation Coalition 2011 (rev. ed. 2019).

[5] See Ross Harvey/Dave Thompson, Automating the Appraisal of Digital Materials, in: *Library Hi Tech* 28 (2/2010), 313-322, doi:10.1108/07378831011047703; Kate Cumming/Anne Picot, Reinventing Appraisal, in: *Archives and Manuscripts* 42 (2/2014), 133-145, doi:10.1080/01576895.2014.926824; Anne Gilliland, Archival Appraisal: Practicing on Shifting Sands, in: Caroline Brown (ed.), *Archives and Recordkeeping: Theory into Practice*, London, 2014; William Vinh-Doyle, Appraising Email (Using Digital Forensics): Techniques and Challenges, in: *Archives and Manuscripts* 45 (1/2017), 18-30, doi:10.1080/01576895.2016.1270838; Victoria Sloyan, Born-Digital Archives at the Wellcome Library: Appraisal and Sensitivity Review of Two Hard Drives, in: *Archives and Records* 37 (1/2016), 20-36, doi:10.1080/23257962.2016.1144504; André Vellino et al., Assisting the Appraisal of E-Mail Records with Automatic Classification, in: *Records Management Journal* 26 (3/2016), 293-313, doi:10.1108/RMJ-02-2016-0006.

Archives, Access and AI is particularly timely. “Born-digital archives” are among the new research priorities highlighted in the UKRI (UK Research and Innovation) Infrastructure Roadmap Progress Report (2019): “The complexity of ‘born-digital’ archives […] and the challenges of archiving for discovery across many different formats raise significant questions about how to preserve, catalogue and make available these materials discoverable and accessible in a coherent fashion, in perpetuity.” The report adds that “this is a fertile area for the arts, humanities and social sciences to explore natural crossovers with other research domains.”[6] A recent UKRI report on UK’s research and innovation infrastructure reiterates this priority on access to cultural collections using new technologies.[7]

Machine Learning applied to data in libraries and other cultural institutions is also at the center of current debates in the US, in Europe and elsewhere. Ryan Cordell recently wrote a Library of Congress report on ML,[8] which built on previous work in the same field – including Thomas Padilla’s report on Data Science, ML and AI in libraries for the OCLC (Online Computer Library Center). In particular, Padilla gives examples of applications of AI/ ML to enhance descriptions of records at scale: “semantic metadata can be generated from video materials using computer vision; text material description can be enhanced via genre determination or full-text summarization using machine learning; audio material description can be enhanced using speech-to-text transcription; and previously unseen links can be created between research data assets that hold the potential to support unanticipated research questions.”[9] *Archives, Access and AI* builds on this booming interest in new technologies applied to born-digital and digitized collections.

While many digital materials are completely “dark” and inaccessible to users, other records are open in theory, but difficult to find in practice. As Mark Bell, Tom Storrar and Jane Winters show in this edited collection, the UK Government Web Archive (UKGWA) is open to anyone with an internet connection, but discoverability is an issue for several reasons, including the inadequacy of keyword search to find relevant materials. The problem of searching huge amounts of records was also at the center of a 2019 article by Winters and Andrew Prescott. “With the rise of very large born-digital resources such as e-mail archives, Wikileak dumps and web

---

[6] UKRI, Infrastructure Roadmap Progress Report, 2019, see 59.

[7] UKRI, The UK’s Research and Innovation Infrastructure: Opportunities to Grow our Capability, 2020, see 3, https://www.ukri.org/wp-content/uploads/2020/10/UKRI-201020-UKinfrastructure-opportunities-to-grow-our-capacity-FINAL.pdf [last accessed: Mar. 31, 2021].

[8] Ryan Cordell, *Machine Learning + Libraries*, Washington D.C., 2020, https://labs.loc.gov/static/labs/work/reports/Cordell-LOC-ML-report.pdf?loclr=blogsig [last accessed: Mar. 29, 2021].

[9] Thomas Padilla, *Responsible Operations: Data Science, Machine Learning, and AI in Libraries*, Dublin, OH, 2019, see 12, doi:10.25333/xk7z-9g97.

archives, the limitations of Google-type searching are becoming more evident."[10] For Winters and Prescott, it can be useful to re-examine the work of mid-twentieth century precursors of the web such as Vannevar Bush and Ted Nelson, who viewed the recording of links and information as the most effective way of processing very large quantities of information.

In addition to textual records, digital pictures can be very difficult to find without adequate metadata and cataloguing. The issue of discoverability is central for the Frick Collection in New York City, an art museum that houses the collection of industrialist Henry Clay Frick (1849 – 1919). Digital images and documentation for hundreds of thousands of art works have already been made freely available on the institution's online digital archive. An ambitious program of digitization is underway for the Frick Art Reference Library's Photoarchive – a research collection with more than 1.2 million reproductions of works of art in the Western tradition. As Ellen Prokop and her co-authors highlight in this edited volume, the rapid pace of digitization has led to a backlog of images that have been digitized but lack metadata and cataloguing information. In turn, this lack of information makes these digitized pictures unfindable and unusable. To solve this problem, the Frick Collection has worked with computer scientists to apply AI to the Photoarchive, automatically annotating images in the collection with the headings used in the archive's classification system. The team harnessed the power of Convolutional Neural Networks (CNNs), a deep learning technique used for classification and computer vision tasks – including image classification and face recognition.

Handwritten manuscripts also fall under the category of hidden collections, difficult to find, search and analyze at scale. While Optical Character Recognition (OCR) can decipher machine-generated texts to make them fully searchable, the technology does not work well for human-generated texts in digital form. Unlike typed characters, no handwritten characters are identical. Handwritten Text Recognition (HTR) is still at an early stage, but significant progress has been made in the past few years using Convolutional Neural Networks. In this edited volume, Tobias Hodel draws on the example of the European project Transkribus, a comprehensive platform for the digitization, AI-powered recognition, transcription and searching of historical documents. He shows that large amounts of data are used to train algorithms to recognize hand-written characters. As a consequence, "the resulting models are highly biased by the material they are trained on," as Hodel points out.

In recent projects on handwritten text recognition with deep learning, researchers often used the IAM Handwriting Dataset to train their models. This dataset contains 115K+ English words by 600+ authors. Since deep learning model

---

[10] Jane Winters/Andrew Prescott, Negotiating the Born-digital: a Problem of Search, in: *Archives and Manuscripts* 47 (3/2019), 391-403, doi:10.1080/01576895.2019.1640753.

need at least 10^5 - 10^6 training examples in order to perform well, the IAM Handwriting Dataset meets those requirements. Using large sets of data is a prerequisite for HTR, but it can also lead to a "cycle of bias" as Hodel puts it. Understanding the source materials and the methods used are essential to engage critically with HTR and more broadly, with any AI-powered techniques applied to archival collections.

The case studies featured in this book will be useful to our intended audience – archivists, digital humanists and social scientists, computer scientists and anyone else interested in the issues faced by archival collections in the twenty-first century. In their 2019 article "More Human than Human? Artificial Intelligence in the Archive," Gregory Rolan and his colleagues note the "barriers to the uptake of AI technology for recordkeeping knowledge work." One explanation is "a lack of compelling case studies": "there are not many real-world examples within the academic or professional literature."[11] Intertwining practical case studies and theoretical insights, the edited collection aims to fill this gap in scholarship. It presents advanced work in Archives and AI, while also "zooming out" and looking at the big picture.

To solve the problem of access to digital archives, cross-disciplinary collaborations are absolutely essential. The big challenges of our time – from global warming to social inequalities – cannot be solved within a single discipline. The same applies to the challenge of "dark" archives. We cannot expect archivists or digital humanities to find a magical solution that will instantly make digital records more accessible. And it is not enough to encourage collaborations between disciplines that are very close (for example, history and literary studies). Instead, we need to take a radical step outside our comfort zone and set up collaborations across disciplines that seldom talk to each other. This is the main goal of the AURA network (Archives in the United Kingdom/ Republic of Ireland and AI), funded by the Arts and Humanities Research Council in the UK and the Irish Research Council in 2020-2021.[12]

Led by a management team with expertise in digital humanities, archives and computer science, AURA organized three workshops to bring people together and offer a forum for discussion and future collaborations: "Open Data versus Privacy" (Workshop 1); "AI and Archives: Current Challenges and Prospects of Born-digital archives" (Workshop 2); "AI and Archives: What comes next?" (Workshop 3). While

---

[11] Gregory Rolan et al., More Human than Human? Artificial Intelligence in the Archive, in: *Archives and Manuscripts* 47 (2/2019), 179-203, see 186, doi:10.1080/01576895.2018.1502088. See also: Basma Makhlouf Shabou et al., Algorithmic Methods to Explore the Automation of the Appraisal of Structured and Unstructured Digital Data, in: *Records Management Journal* 30 (2/2020), 175-200, doi:10.1108/RMJ-09-2019-0049; Tim Hutchinson, Natural Language Processing and Machine Learning as Practical Toolsets for Archival Processing, in: *Records Management Journal* 30 (2/2020), 155-174, doi:10.1108/RMJ-09-2019-0055.

[12] www.aura-network.net[last accessed: Mar. 29, 2021].

AURA focuses mostly on the UK and Ireland, AEOLIAN (Artificial Intelligence for Cultural Organizations) strengthens connections between British and American partners. Funded by the AHRC in the UK and the National Endowment for the Humanities in the US, AEOLIAN runs from 2021 to 2023 and consists in a series of meetings and case studies that bring together a team of experts to develop new approaches to improving access to and use of digital archives.[13]

Other initiatives have created partnerships between experts in various fields, benefiting cultural institutions in general and archival collections in particular. The AHRC-funded Computational Archival Science (CAS) research network and the Advanced Information Collaboratory explore the conjunction of big data methods and technologies with archival practice. In Germany, the Leipzig Computational Humanities group includes experts in the humanities and computer science. In the USA, the HathiTrust Research Centre and Stanford Literary Lab also foster computational research in the humanities. The global spread of these initiatives show that cross-disciplinary collaborations are not enough: we also need to bring the best people together, independently of their nationalities and professional affiliations.

Published by a German press and written by an international team of contributors, the six chapters of this book feature examples of collaborations between researchers in computer science and engineering, archivists and scholars in the digital humanities. Often, these collaborations are made possible thanks to external funding and are hosted in large cultural institutions in world cities. Applying AI to archives would be difficult for a small archival collection in, say, Loughborough (a market town in the North of Britain). The combination of prestigious metropolitan institutions, ground-breaking technology, and advanced expertise can be intimidating. But this does not have to be the case. AI is still a very imperfect technology, with applications to the archival sector at an early, experimental stage.

In a 2019 article entitled "Artificial Intelligence – the Revolution hasn't happened yet," Michael Jordan notes that when the term "AI" was coined in the late 1950s, it referred to the ambition to build hardware and software possessing human-level intelligence. He uses the phrase "human-imitative AI" to refer to this aspiration to create an entity that would resemble humans. AI was meant to focus on "the high-level or cognitive capability of humans to reason and to think," Jordan points out. "Sixty years later, however, high-level reasoning and thought remain elusive. The developments now being called AI arose mostly in the engineering fields associated with low-level pattern recognition and movement control, as well as in the field of statistics, the discipline focused on finding patterns in data and on making well-founded predictions, tests of hypotheses, and decisions."[14]

---

[13] www.aeolian-network.net[last accessed: Mar. 29, 2021].

[14] Michael I. Jordan, Artificial Intelligence – The Revolution Hasn't Happened Yet, in: *Harvard Data Science Review* 1 (1/2019), 1-9. doi:10.1162/99608f92.f06c6e61.

If we look at the specific case of archival collections, it is certainly true that AI has been used for low-level tasks: identifying sensitive information such as credit card numbers in emails, tagging pictures, transcribing handwriting for examples. These tasks could be done by any normal teenager with a minimum of training. The value of AI is not its ability to perform complex high-level tasks that require contextualization, theorization or creativity. Instead, the value of AI comes from its capacity to process huge amounts of data very rapidly – something that no human can do single-handedly.

Michael Jordan argues that success in human-imitative AI has been quite limited. We are very far from having artificially intelligent systems that can compete with humans at the higher levels of intelligence. A focus on human-imitative AI can distract us from a key challenge of our times: making sure that AI works for humans, that it makes our human lives better rather than worse. For Jordan, we are witnessing the creation of a new discipline: *human-centric engineering*. "Whereas civil engineering and chemical engineering built upon physics and chemistry, this new engineering discipline will build on ideas that the preceding century gave substance to, such as information, algorithm, data, uncertainty, computing, inference, and optimization."[15] Since the new discipline will focus on data from and about humans, it will need the perspectives of social scientists and humanists.

Applied to archival collections, *human-centric engineering* will focus on building artifacts and designing processes to make archives more accessible. The new discipline will bring together not only engineers, data scientists and computer scientists, but also archivists and scholars in the humanities and social sciences. Working collaboratively, these interdisciplinary teams will pay close attention to issues of privacy and biases. Unlocking archives should not come at any price, and the hype surrounding Open Data and AI must not distract us from the need to comply with data protection regulations and to address bias associated with black-box algorithms. In short, we need to mitigate the risks of malicious AI in our quest to unlock archival records.

2. *The Threat of Dark AI*

Using AI to make dark archives accessible is risky: sensitive information in government archives could inadvertently be released and fall into the hands of criminals; private information in email archives could be leaked, leading to distress for individuals and breaches of data protection laws; pictures in digital collections could be mis-labelled, leading to embarrassment and damage to the cultural institution's "brand." Automatic image labelling has a long history of producing embarrassing

---

[15] Jordan, Artificial Intelligence, 3.

results – in 2015, Google Photos labelled a picture of two black people as “gorillas.” Google Maps and Flickr have also suffered from race-related problems. Training datasets that contain biases result in problematic, sometimes appalling outcomes. In 2016, after engaging with users on Twitter, a Microsoft chat box began sharing racist, genocidal and misogynistic messages. And in 2020, the researcher Timnit Gebru said she was fired from Google after co-writing a paper on AI-generated language, which replicates unsavory biases found in online text. To unlock dark archives, we cannot rely on dark AI – defined as AI that is making human lives worse, not better.

Bias in large datasets is at the center of Thomas Padilla’s and Ryan Cordell’s recent reports on ML applied to libraries and archives. Cordell gives the example of the *Chronicling America* newspaper collection, a searchable database of US newspapers with descriptive information and selected digitized pages. *Chronicling America* is produced by the National Digital Newspaper Program (NDNP), a partnership between the National Endowment for the Humanities and the Library of Congress. The NDNP relies on institutions in each state to select and digitize approximately 100,000 newspaper pages representing that state’s history and geographic coverage. The website gives the impression that digitized collections reflect the diversity of each state:

> Participants are expected to digitize primarily from microfilm holdings for reasons of efficiency and cost, encouraging selection of technically suitable film, bibliographic completeness, diversity and “orphaned” newspapers (newspapers that have ceased publication and lack active ownership) in order to decrease the likelihood of duplicative digitization by other organizations.[16]

Yet, far from reflecting diversity, “the data skews to newspapers serving the majority,” Cordell argues.[17] For example, the collection privileges newspapers read by white middle-class audiences in the nineteenth century, rather than black and other minority-run papers. As Benjamin Fagan points out, *Chronicling America* does currently list forty-six black newspapers in its digital archive (c. 2.5% of a total of 1,799), but all were printed in 1865 or later.[18] There are no black newspapers among the digital copies of 215 newspapers published before 1865. NDNP participants prioritized geographic spread, inadvertently deemphasizing racial representation.

---

[16] https://chroniclingamerica.loc.gov/about/[last accessed: Mar. 29, 2021].

[17] Cordell, *Machine Learning + Libraries*, 14.

[18] Benjamin Fagan, Chronicling White America, in: *American Periodicals: A Journal of History & Criticism* 26 (1/2016), 10-13, see 11, https://muse.jhu.edu/article/613375 [last accessed: Mar. 29, 2021].

"Consequently any ML projects based on *Chronicling America* will reflect those same oversights and exclusions," writes Cordell.[19]

Bias in large newspaper datasets is also at the center of the *Oceanic Exchanges* project (2017-2019). Melodee Beals and her co-investigators argue that the national focus of digitized newspapers collections obscures the fact that international news exchange was central to the nineteenth-century press. ML projects based on national newspapers risk missing important international links with other papers. To mitigate these risks, *Oceanic Exchanges* produced the *Atlas of Digitised Newspapers and Metadata*, an open access guide to digitized newspapers around the world. Highlighting the history of digitized newspapers and digitization choices, the atlas examines metadata available in these collections. In particular, it "explores how machine-readable information about an issue, volume, page, and author is stored in the digital file alongside the raw content or text."[20] This project sheds light on information that is often hidden. It invites researchers and other users to see digitized newspaper datasets as human constructions, rather than unproblematic data.

Managing bias is essential for libraries and archival collections. For Thomas Padilla, eliminating bias entirely is not an option: in the hope of cleaning the dataset, elimination risks introducing more bias. Instead, Padilla proposes a bias management strategy to reflect on and integrate bias within the cultural organization. One of the recommendations is to:

> Hold symposia focused on surfacing historic and contemporary approaches to managing bias with an explicit social and technical focus. The symposium should gather contributions from individuals working across library organizations and focus critical attention on the challenges libraries faced in managing bias while adopting technologies like computation, the internet, and currently with data science, machine learning, and AI. Findings and potential next steps should be published openly.[21]

Symposia and other communication strategies would promote self-conscious attention and criticism at every stage of an ML project. But, as Cordell argues, this may not be enough. "To create ML projects that reflect data justice, […] libraries cannot pretend to be neutral or objective in relationship to race, class, gender, sexuality, or culture, but instead must consciously strive to forefront marginalised voices."[22]

The notion of data justice is closely related to the black box problem, explain Catherine D'Ignazio and Lauren Klein in *Data Feminism*. Machine learning algo-

---

[19] Cordell, *Machine Learning + Libraries*, 14.

[20] https://www.digitisednewspapers.net/dhawards/[last accessed: Mar. 29, 2021].

[21] Padilla, *Responsible Operations*, 10.

[22] Cordell, *Machine Learning + Libraries*, 15.

rithms are so complex that they are often described as incomprehensible black boxes: you put data in, and you get something out, but what happens inside the box is a mystery. For D'Ignazio and Klein, data justice aims to "ensure that past inequities are not distilled into black-boxed algorithms." While terms such as *ethics* "locate the source of the problem in individuals or technical systems," *justice* acknowledges "structural power differentials" and works "toward dismantling them."[23] Data justice seeks to address the structural inequalities in the training datasets that lead algorithms to produce less-favorable outcomes for women and ethnic minorities.

In *Algorithms of Oppression*, Safiya Umoja Noble argues that Google search algorithms privilege white people and discriminate against ethnic minorities, particularly women. These algorithms in turn reinforce existing prejudices against women of color with search results presenting black women as "angry" or "sassy." Noble points out that the search operations are invisible: users of Google and other search engines have no access to the algorithms and deep machine learning systems, developed to index masses of information and move some to the first page of results. While the process of searching is hidden, the information users see on their screens becomes a reality and has an impact on decision making. For Noble, Artificial Intelligence will become a major human rights issue in the twenty-first century. "We are only beginning to understand the long-term consequences of these decision-making tools in both masking and deepening social inequality."[24]

Noble, Cordell and others have stressed the role that libraries and other cultural institutions can play to balance the power of tech giants. For Noble, "the public is increasingly reliant on search engines in lieu of libraries, librarians, teachers, researchers, and other knowledge keepers and resources." Yet, it does not make sense "to outsource all of our knowledge needs to commercial search engines" that will return biased results.[25] Librarians and other knowledge keepers can mitigate bias and offer an alternative to commercial interests. Likewise, Cordell argues that "by centering ethics, transparency, diversity, privacy and inclusion, libraries can take a leadership role in one of the central cultural debates of the twenty-first century."[26] Many entrepreneurs in the tech industry still live by Facebook's early motto: "Move fast and break things." But the same motto cannot apply to cultural institutions that value continuity over disruption.

---

[23] Catherine D'Ignazio/Lauren F. Klein, *Data Feminism*, Cambridge, MA, 2020, https://data-feminism.mitpress.mit.edu/ [last accessed: Mar. 29, 2021].

[24] Safiya Umoja Noble, *Algorithms of Oppression: How Search Engines Reinforce Racism*, New York City, 2018, 1.

[25] Noble, *Algorithms of Oppression*, 16.

[26] Cordell, *Machine Learning + Libraries*, 1.

> As institutions of memory and community, libraries cannot be bound to destructive ideologies of technological implementation but must instead model alternative engagements with ML focused on building rather than breaking. Libraries can become ideal sites for cultivating responsible and responsive ML, as that term describes a constellation of technologies that explain data within the contexts of its collection, aggregation, and association.[27]

This debate over the place of libraries in the digital age is not new. In his 1994 article "Electronic Records, Paper Minds," Terry Cook reminded information professionals that their role was to guide users from masses of information onto specific knowledge. At a time of big digital data, this role had become particularly challenging. If librarians and other knowledge keepers failed, they will "be replaced by software packages that can handle facts, and data, and information very efficiently, without any mediation by archivists or anyone else."[28]

In January 1994, shortly before Cook's article appeared, two electrical engineering graduate students at Stanford University created a guide to the World Wide Web. Their website, which was named Yahoo in March 1994, was a directory of other websites, organized in a hierarchy, as opposed to a searchable index of pages. Commercial search engines then evolved to rank results by counting how many times the search terms appeared on the page. The creation of Google in 1998 marked a turning point, with the development algorithms that analyzed links and relationships between websites to determine their importance. By the late 1990s and the massification of internet access, people increasingly relied on commercial search engines to find information, rather than on traditional knowledge keepers.

Tech giants took the advantage over libraries three decades ago and have remained at the forefront of information searching. It is not surprising that the field of AI/ ML applied to cultural institutions is so heavily invested by Google and the like. For the "LIFE Tags" project, Google organized over 4 million images from the LIFE magazine archives into an interactive encyclopedia. Machine learning automatically applied tags to digitized images, which greatly simplified the archiving work since the archive spans approximately 1,800 meters in three different warehouses. LIFE Tags allows users to easily navigate the magazine archives using keywords. For this project, Google drew on a deep neural network used in Google Photo search that has been trained on millions of images. As we have seen, this technology is not neutral: the training dataset can be biased, which in turn can lead to problematic labels (including racist labels).

---

[27] Cordell, *Machine Learning + Libraries*, 2.

[28] Terry Cook, Electronic Records, Paper Minds: The Revolution in Information Management and Archives in the Post-Custodial and Post-Modernist Era, in: *Archives and Manuscripts* 22 (2/1994), 300-328, see 306.

Google also worked with MoMA (Museum of Modern Art) in New York City, to automatically identify art works in archival photos of exhibitions. Since 1929, MoMA has kept thousands of photos of its exhibitions. However, these images were difficult to find and use as they lacked metadata and other information on the works displayed in exhibitions. Google's algorithms automatically identified 27,000 works of art and made MoMA collections more accessible. Again, the issue comes from black-boxed algorithms: the museum has no control over the algorithms used by Google, and over the results produced by these algorithms.

Instead of a black box model controlled by tech giants, a more open model is possible.

Fig 0.1: Implementing an open model for ML projects in libraries and cultural organizations. Courtesy of the author.

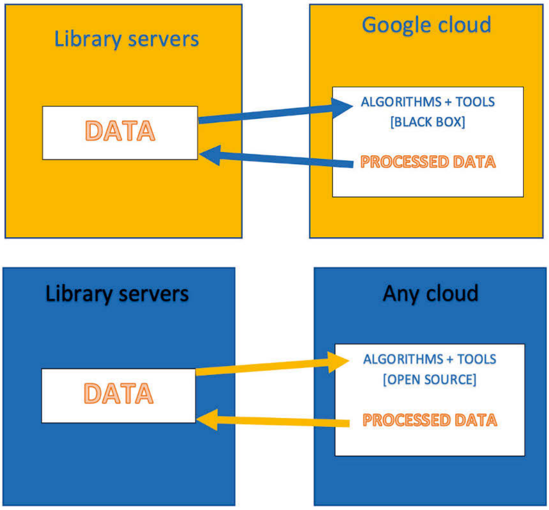

Fig 0.1 gives a simplified account of what goes on when libraries and cultural organizations partner with Google and the like. They send their data to the tech partner and receive processed data in return. What happens in between is largely controlled by the tech company. It is possible to take back control. Instead of relying on tech giants to use their own algorithms (such as Google's Image Content-based Annotation algorithm to generate labels based on image pixels), cultural organizations could work with multi-disciplinary teams to generate their own algorithms. There is no need to re-invent the wheel: many open-source AI software are

easily available, including tools released by Google, Microsoft, Facebook and other tech companies. Teams composed of librarians and archivists, humanities scholars, computer scientists and software engineers would be well-equipped to conduct ML projects currently outsourced to tech giants. There are obvious obstacles, including funding and the availability of expertise, but these concerns should be viewed as challenges rather than showstoppers. Funding bodies are increasingly pushing for cross-disciplinary projects crossing the divide between the sciences and the humanities. "Reuniting the humanities with sciences might protect their future," notes a recent *Economist* article on Digital Humanities.[29]

3. *AI for Good*

What does "AI for good" mean for archival collections? Combined with other technologies, AI has the potential to make digital archives more comprehensive and to guarantee the authenticity of records. Let's start with the issue of comprehensiveness. If you are buying a house, you do not want to request a property title and discover that the record has disappeared from the archive, or never was there in the first place. Yet, anyone who has done archival work knows that archives are rarely complete. Documents are thrown away deliberately or by mistake, either by the record creator or by the archivist. Appraisal, i.e. the selection of records, is central aspect of the archival process. In *Modern Archives: Principles and Techniques* (1956), the American archivist T. R. Schellenberg argued that a record has "primary value" to the creator but may also have "secondary value" (as evidence or information) to historians and future users of the records outside of the originating institution. "Archivists should be empowered to review all records that government agencies propose to destroy."[30] For Schellenberg, only records that have secondary value over the long term should be kept in archival collections.

Reviewing records manually has become almost impossible in the digital age. Archivists have to deal with huge amounts of records, but also records that are scattered in various places – both within and outside the creator's organizational system. The deployment of cheap cloud technologies, smartphones and other mobile devices have led to a rise in "shadow IT," i.e. IT systems deployed and supported by providers outside the organization's central IT and by definition not aligned to the central IT strategy and direction. For example, civil servants may be tempted to use private emails accounts, WhatsApp, Skype, Facebook, Twitter and other platforms

---

[29] "How Data Analysis Can Enrich the Liberal Arts," in: *Economist*, 19.12.2020, https://www.economist.com/christmas-specials/2020/12/19/how-data-analysis-can-enrich-the-liberal-arts [last accessed: Mar. 29, 2021].

[30] Theodore R. Schellenberg, *Modern Archives: Principles and Techniques*, Chicago 1956, see 32.

to easily share information with their colleagues, instead of using government IT tools. Not only does shadow IT increase the risk of data leaks, but it also makes it impossible to comply with record management obligations.

In the UK, these obligations are outlined in the Freedom of Information Act 2000, the Public Records Act 1958, and the Data Protection Act 2018 following the introduction of the GDPR in Europe. In a nutshell, the Freedom of Information Act provides public access to information held by public authorities. To foster a culture of open government accountable to citizens, disclosure of information should be the default. In other words, information should be kept private only when there is a good reason, and it is permitted by the FOI Act. Moreover, the amended Public Records Act states that records of UK central government selected for permanent preservation shall be transferred not later than twenty years after their creation to The National Archives (rather than the previous thirty years). Finally, the Data Protection Act gives people more control over use of their data and provides them with new rights to move or delete personal data.

For Knowledge and Information Management (KIM) government professionals, shadow IT is a major problem. To respond to Freedom of Information requests, Data Protection subject access and public inquiries, KIM professionals need access to the relevant records. If these records are scattered outside official channels, they often become undiscoverable and inaccessible – making government vulnerable to accusations of secrecy and malpractice that can potentially lead to prosecution.

Even when information remains within government IT systems, it can be extremely difficult to find. This issue was at the center of the *Better Information for Better Government* report that the UK Cabinet Office released in 2017. "While little information has been lost altogether, much of what has accumulated over the past fifteen to twenty years is poorly organised, scattered across different systems and almost impossible to search effectively."[31] At Year 7 following their creation, archival materials are transferred to an internal archive where they stay for thirteen years, before their transfer to The National Archives at Year 20. "We need to know what's here, we need to be able to find it," one KIM professional points out. "When we take it in our archive, we need to be able to index it, to classify it and catalogue it properly" in order to find information easily and respond to possible Freedom of Information requests and other access requests.[32]

Artificial Intelligence has a role to play in bringing scattered records together, making them findable and usable. However, it is a controversial role since AI makes

---

[31] Cabinet Office (UK), *Better Information for Better Government*, London 2017, https://www.gov.uk/government/publications/better-information-for-better-government [last accessed: Mar. 29, 2021].

[32] Conversation with author, Nov. 2, 2020.

an intervention on the records. *Respect des fonds* is a key principle in archival theory that goes back to at least the nineteenth century.[33] According to this principle, archival materials need to be grouped according to their *fonds* or origins. Archivists should maintain records using the creator's organizational system instead of imposing a new order. Established at a time of growing archival records, the principle allowed archivists to save time and avoid any attempts to re-arrange documents created by the same agency, individual, or organization. These attempts would be at best futile, and at worst would tamper with the collection.

In the digital age, the purist and non-interventionist viewpoint of *respect des fonds* has come under attack. Is it better for archivists to passively accept digital information acquired into archives and store it essentially as it comes, without any modification? For some Knowledge and Information management professionals, the principle is no longer adequate at a time when digital information created by organizations is inherently chaotic and unorganized. To easily find information and respond to requests, it is necessary to (re)organise records by grouping, metadata tagging and the like, and in so doing actively interpret them and construct them into thematic archives.

On a small scale, (re)-organizing archival collections can be done manually. One archivist built a digital collection on the 2012 Olympics in London by directly approaching people involved in the preparation and delivery of the events and asking them to send their digital records. The resulting thematic archive does not conform to *respect des fonds* since it was actively created by the archivist rather than passively received. But it served an important purpose: bringing scattered documents together and making them easy to find, search and use.

According to this more interventionist principle, AI's role would be to automatically add metadata, extract names and topics. As Tobias Hodel explains in this edited collection, topic modeling is a statistical method used in machine learning and Natural Language Processing to discover clusters of words or "topics" that occur in a dataset. The approach uses unsupervised machine learning: algorithms identify what words appear together frequently, resulting in the extraction of topics. AI would not only improve the discoverability of records, but it would also make them more accessible.

By automatically identifying sensitive records, AI would allow non-sensitive records to be opened up and made available to researchers and other users. Automatic sensitivity review is still at early stage, but it has the potential to shed light on "dark" archives. This is particularly important for departments that deal with a lot of sensitive and confidential information – such as the UK Cabinet Office as

---

[33] Michel Duchein, Theoretical Principles and Practical Problems of Respect Des Fonds in Archival Science, in: *Archivaria* (16/1983), 64-82.

opposed to, say, the Department for Education (DfE). The risk appetite of the Cabinet Office is very low because many of its records are very secret and sensitive. In contrast, because DfE's policy making is about schools and education and is very public, the risk of leaking sensitive information is much lower.

Because AI is deployed on huge amounts of data, it would result in vast re-organized archives free of sensitive materials. UK central government departments currently only send a very small proportion of their records to The National Archives (around 5%). Is it advisable to continue this approach? Or would it be better to exploit the potential of digital and AI to make available a larger corpus? The government needs to avoid the release of sensitive material, but it also needs to encourage access and transparency – principles that are at the heart of the Freedom of Information Act 2000 and the National Data Strategy (NDS) policy paper released in September 2020, which aims to "unlock the power of data for the UK." "Data is a non-depletable resource in theory, but its use is limited by barriers to its access – such as when data is hoarded, when access rights are unclear or when organisations do not make good use of the data they already have," declares the NDS.[34]

Although Artificial intelligence can be used to re-organize the archive and add metadata, the technology is fraught with risks. When record creators work with commercial partners to apply AI to their archives, they rarely invite external archivists to the table. Yet, selected records will eventually end up in the external archive. Does the archivist have a role to play in defining the algorithms and code needed for decision making? asked Anthea Seles of the International Council on Archives (ICA) in 2019.[35] She outlined four main challenges and issues. First, archivists will be responsible for the conservation of these algorithms in archives used by historians and other users. Second, archivists are not currently considered stakeholders in discussions related to the development and implementation of AI technologies. Third, archivists currently do not have the capacity and skills to play their role as advisors on good records management to ensure the longevity and sustainability of these new archival documents. Fourth, archivists will need to understand not only how to advise on the conservation of AI algorithms, but also how to deal with important ethical issues – including the black box issue. Even with all the necessary elements are kept, it is often difficult to understand how an algorithm came to a decision. When AI is applied to archives, the risk is to bias the historical document and consequently history as well as our collective memory.

Archivists rightly ask for a seat at the table and an opportunity to shape the discussions on AI applied to archives. "Automation is no longer a choice, it is a necessity but that does not mean that the archivist (the human) is not relevant in the

---

[34] https://www.gov.uk/government/publications/uk-national-data-strategy/national-data-strategy[last accessed: Mar. 29, 2021].

[35] Anthea Seles, Norwegian Triennial Archival Conference, April 8, 2019.

process," Seles points out. Advocating for algorithmic transparency and accountability are becoming key roles for archivists. In 2017, the Association for Computing Machinery issued a statement outlining seven principles to make public policy more transparent and accountable. First, the principle of *awareness* implies that all stakeholders should be aware of the bias problem and potential harm associated with automation. Second, a right to *access and redress* should allow individuals and groups that are adversely affected by algorithmically informed decisions to question these decisions. Third, the principle of *accountability* would hold institutions responsible for decisions made by the algorithms that they use. Fourth, producing *explanations* should be required from institutions that use algorithmic decision-making, in order to understand how specific decisions are made. Fifth, *data provenance* should be documented, including a description of the way in which the training data was collected. Sixth, the principle of *auditability* should encourage institutions to record models, algorithms, data, and decisions so that they can be audited. Seventh, the principles of *validation and testing* should encourage institutions to validate and document their models, but also to routinely perform tests to make sure that models do not generate harm.

Like many humanities scholars, archivists often fear that they lack the skills and knowledge to productively participate in these debates. But there is no need to be a computer scientist or software engineer to have a productive discussion on AI, transparency and accountability. The private sector offers a valuable model to bring together professionals with various specialisms and expertise. When an IT consulting company is contracted to implement a new system or deliver a solution to a data problem, the first thing they do is to bring together a committee with representatives of the client's internal IT and operational services. The committee defines standards to apply to the data, and starts with a pilot (for example, a small amount of data to process). Following an agile process, consultants then work to deliver the project in close discussion with their client's stakeholders from various teams. In the case of government archives, it should be possible to bring to the same table AI specialists (either consultants or internal experts), government record creators and experts from different services including KIM professionals, and external archivists. A central objective would be to improve the organization and comprehensiveness of the digital archive, while also pushing for algorithmic transparency and accountability.

In addition to comprehensiveness, authenticity is central to archival collections. One of the most important roles of archives in our societies is to preserve authentic documents, before making them available to users. If you are buying a house, you will need to access authentic records about previous ownership. And if a historian consults government records of the nineteenth century, they need to trust that the records have not been tampered with. Guaranteeing the integrity of digital records is a key objective for The National Archives UK and other institutions. As technology evolves and software used to read certain formats becomes obsolete, digital records often need to be ported from one format to another. Although these records are easy to copy and modify, their content must remain unaltered while stored in the archive.

The ARCHANGEL project (2017-2019) addressed the challenges around trust, integrity and authenticity of born-digital archival materials by exploring the possibilities offered by blockchain and machine learning.[36] Blockchain is the technology that underpins Bitcoin and other cryptocurrencies, but it has the potential for application to other sectors. With blockchains, data can be added to the chain, but it cannot be overwritten, amended or deleted. A blockchain is therefore a growing list of records, called blocks – with each block containing a cryptographic hash[37] of the previous block, a timestamp and other information. Moreover, the technology is distributed, i.e. no central organization has sole possession or control over of the data. Finally, it is transparent, with all entries in the chain visible to all trusted members who have a copy. Combined with machine learning, blockchain offers a digital fingerprint for archival materials, making it possible to verify their authenticity.

ARCHANGEL prototyped the creation of hashes using machine learning methods, particularly for image and video records. ML can identify the causes of glitches and noise in these records – which could either be caused by transcoding and format-shifting, or by any undesirable process, such as corruption of the files in storage or tampering. Machine learning complements the ARCHANGEL blockchain, which enables archival collections to upload metadata that uniquely identified specific records. In the case of sensitive records, metadata can itself be confidential and sensitive, making it inappropriate to add it to the blockchain. One solution is to add an archival reference and the record's checksum instead (a unique computer-generated string that changes if the file is altered). That data is then sealed into a block that cannot be changed or deleted without detection. Finally, a copy of the data is shared with all trusted members of the network. The ARCHANGEL example shows that AI can work *for* rather than *against* archival collections.

4. *Structure*

Archives, Access and AI is organized in two parts, with three chapters in each section. The first part on "Selection, Appraisal, Discoverability and Access" starts with the example of AI applied to the Photoarchive at the Frick Art Reference Library

---

[36] http://www.archangel.ac.uk/ [last accessed: Mar 29 2021]

[37] A cryptographic hash is an algorithm that takes an arbitrary block of data and returns a fixed-size bit string.

in New York. The project led to the automatic creation of metadata that improved the discoverability of the archive. This collection has been made more accessible and usable thanks to a cross-disciplinary team with expertise in computer science, art history and other fields. Chapter 2 on web archives also moves away from single disciplines to solve the problems associated with large-scale digital collections. Bringing together archive professionals and a digital humanist, the project aims to make born-digital archives more accessible by prototyping a researcher dashboard for the UK government web archive. Chapter 3 focuses on design thinking, a human-centered method to solving business and social problems. It argues that design thinking is a productive way to solve the problems of *access* and *use* of archival collections in the digital age. Researchers should work closely with archivists to shape access policies that will facilitate the use of AI and other innovative methodologies.

The second part on "Using the Archives: AI and New Knowledge" starts with a chapter on digital library user studies. Investigating the impact of e-legal deposit on UK academic deposit libraries in Chapter 4, Paul Gooding argues that transparent workflows and data documentation should be central to user studies. Currently, library user studies often rely on tools such as Google Analytics and suffer from a black box problem. The quality of library patron data is also an issue, with potentially biased data leading to problematic results. In Chapter 5, Martin Paul Eve and his co-authors examine another large-scale dataset: academic peer review reports in the sciences. The confidentiality of these reports and the difficulties of access make research extremely complicated. Neural networks can be used to machine-read these archives and make them more accessible, but the process is fraught with difficulties. In Chapter 6, Tobias Hodel focuses on Handwritten Text Recognition (HTR). Supervised deep learning approaches have led to astonishing results in deciphering handwriting, but also to new problems – including in terms of transparency and accountability. Hodel also presents an overview of unsupervised machine learning, including topic modeling used on huge amounts of textual data. Continuing the discussion on HTR in Chapter 7, Melissa Terras shows that this technology is now transforming access to our written past. Drawing on a survey of users of HTR on the Transkribus platform, Terras highlights issues raised when inviting machine learning into historical archives. Transcriptions generated by HTR will require new approaches to both history and public engagement, Terras argues, before providing recommendations on how to best support the community applying HTR to cultural heritage materials. Finally, an afterword by Richard Marciano offers further thoughts on the intersection of technology and archives to produce new ways of preserving and making accessible our collective past.

Bibliography

Cabinet Office (UK), *Better Information for Better Government*, London 2017, https://www.gov.uk/government/publications/better-information-for-better-government [last accessed: Mar. 29, 2021].

Cook, Terry, Electronic Records, Paper Minds: The Revolution in Information Management and Archives in the Post-Custodial and Post-Modernist Era, in: *Archives and Manuscripts* 22 (2/1994), 300-328.

Cordell, Ryan, *Machine Learning + Libraries*, Washington D.C., 2020, https://labs.loc.gov/static/labs/work/reports/Cordell-LOC-ML-report.pdf?loclr=blogsig [last accessed: Mar. 29, 2021].

Cumming, Kate/Picot, Anne, Reinventing Appraisal, in: *Archives and Manuscripts* 42 (2/2014), 133-145, doi:10.1080/01576895.2014.926824.

D'Ignazio, Catherine/Klein, Lauren F., *Data Feminism*, Cambridge, MA, 2020, https://data-feminism.mitpress.mit.edu/ [last accessed: Mar. 29, 2021].

Duchein, Michel, Theoretical Principles and Practical Problems of Respect Des Fonds in Archival Science, in: *Archivaria* (16/1983), 64-82.

Erway, Ricky, Defining "Born Digital," OCLC Research, November 2010, URL: https://www.oclc.org/content/dam/research/activities/hiddencollections/borndigital.pdf [last accessed: Mar. 29, 2021].

Fagan, Benjamin, Chronicling White America, in: *American Periodicals: A Journal of History & Criticism* 26 (1/2016), 10-13, https://muse.jhu.edu/article/613375 [last accessed: Mar. 29, 2021].

Gilliland, Anne, Archival Appraisal: Practicing on Shifting Sands, in: Caroline Brown (ed.), *Archives and Recordkeeping: Theory into Practice*, London, 2014.

Harvey, Ross/Thompson, Dave, Automating the Appraisal of Digital Materials, in: *Library Hi Tech* 28 (2/2010), 313-322, doi:10.1108/07378831011047703.

Hutchinson, Tim, Natural Language Processing and Machine Learning as Practical Toolsets for Archival Processing, in: *Records Management Journal* 30 (2/2020), 155-174, doi:10.1108/RMJ-09-2019-0055.

Jordan, Michael I., Artificial Intelligence – The Revolution Hasn't Happened Yet, in: *Harvard Data Science Review* 1 (1/2019), 1-9. doi:10.1162/99608f92.f06c6e61.

McGillivray, Barbara, et al., *The Challenges and Prospects of the Intersection of Humanities and Data Science: A White Paper from The Alan Turing Institute*, London 2020, doi:10.6084/M9.FIGSHARE.12732164.

Noble, Safiya Umoja, *Algorithms of Oppression: How Search Engines Reinforce Racism*, New York City, 2018.

Padilla, Thomas, *Responsible Operations: Data Science, Machine Learning, and AI in Libraries*, Dublin, OH, 2019, doi:10.25333/xk7z-9g97.

Prom, Christopher J., *Preserving Email – DPC Technology Watch Report*, Digital Preservation Coalition 2011 (rev. ed. 2019).

ROLAN, Gregory, et al., More Human than Human? Artificial Intelligence in the Archive, in: *Archives and Manuscripts* 47 (2/2019), 179-203, doi:10.1080/01576895.2018.1502088.

SCHELLENBERG, Theodore R., *Modern Archives: Principles and Techniques*, Chicago 1956.

SHABOU, Basma Makhlouf, et al., Algorithmic Methods to Explore the Automation of the Appraisal of Structured and Unstructured Digital Data, in: *Records Management Journal* 30 (2/2020), 175-200, doi:10.1108/RMJ-09-2019-0049.

SLOYAN, Victoria, Born-Digital Archives at the Wellcome Library: Appraisal and Sensitivity Review of Two Hard Drives, in: *Archives and Records* 37 (1/2016), 20-36, doi:10.1080/23257962.2016.1144504.

UKRI, The UK's Research and Innovation Infrastructure: Opportunities to Grow our Capability, 2020, https://www.ukri.org/wp-content/uploads/2020/10/UKRI-201020-UKinfrastructure-opportunities-to-grow-our-capacity-FINAL.pdf [last accessed: Mar. 31, 2021].

VELLINO, André, et al., Assisting the Appraisal of E-Mail Records with Automatic Classification, in: *Records Management Journal* 26 (3/2016), 293-313, doi:10.1108/RMJ-02-2016-0006.

VINH-DOYLE, William, Appraising Email (Using Digital Forensics): Techniques and Challenges, in: *Archives and Manuscripts* 45 (1/2017), 18-30, doi:10.1080/01576895.2016.1270838.

WINTERS, Jane/PRESCOTT, Andrew, Negotiating the Born-digital: a Problem of Search, in: *Archives and Manuscripts* 47 (3/2019), 391-403, doi:10.1080/01576895.2019.1640753.