



How can we make born-digital and digitised archives more accessible? Identifying obstacles and solutions

Lise Jaillant¹

Accepted: 23 February 2022 / Published online: 24 March 2022
© The Author(s) 2022

Abstract

Access to data is seen as a key priority today. Yet, the vast majority of digital cultural data preserved in archives is inaccessible due to privacy, copyright or technical issues. Emails and other born-digital collections are often uncatalogued, unfindable and unusable. In the case of documents that originated in paper format before being digitised, copyright can be a major obstacle to access. 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 closed to users. We cannot expect archivists or digital humanists to find a magical solution that will instantly make digital records more accessible. Instead, we need to set up collaborations across disciplines that seldom talk to each other. Based on 21 interviews with 26 archivists, librarians and other professionals in cultural institutions, we identify key obstacles to making digitised and born-digital collections more accessible to users. We outline current levels of access to a wide range of collections in various cultural organisations, including no access at all and limited access (for example, when users are required to travel on-site to consult documents). We suggest possible solutions to the problems of access—including the ethical use of Artificial Intelligence to unlock “dark” archives inaccessible to users. Finally, we propose the creation of a global user community who would participate in decisions on access to digital collections.

Keywords Born-digital archives · Digitised archives · Copyright · Data protection · Artificial intelligence · User community

✉ Lise Jaillant
l.jaillant@lboro.ac.uk

¹ School of Social Sciences and Humanities, Loughborough University, Epinal Way, Loughborough LE11 3TU, UK

Introduction

Access to data is seen as a key priority today. Yet, the vast majority of digital cultural data preserved in archives is inaccessible due to privacy, copyright or technical issues. Emails and other born-digital collections are often uncatalogued, unfindable and unusable.¹ In the case of documents that originated in paper format before being digitised, copyright can be a major obstacle to access. As a literary scholar and digital humanist, I have extensive experience of trying (and often failing) to get access to digital collections. I was trained as a “traditional” scholar and spent several years looking at old letters and other paper documents in archival repositories. Getting access to archival emails, PDFs and Word documents is much more complicated. I was lucky to be able to consult limited number of emails in the Ian McEwan collection at the Harry Random Center, and in the Carcanet Press archive in Manchester. But access to born-digital data remains the exception, not the norm.

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 closed to users. We cannot expect archivists or digital humanists to find a magical solution that will instantly make digital records more accessible. Instead, we need to set up collaborations across disciplines that seldom talk to each other.

In particular, collaborations with Computer Scientists can be fruitful. Artificial Intelligence and machine learning² can be used to unlock archives and make them more findable and accessible—for example by differentiating between sensitive and non-sensitive materials, or by automatically adding tags and other metadata to improve findability. But it is crucial to avoid biases in the selection and processing of data, which could discriminate against certain groups and impact the collective memory. This requires engagement with algorithms rather than treating AI as a “black box.” “Explainable AI” is becoming essential so that humans understand how the machine came to particular outcomes and decisions. Networks, such as AURA (Archives in the UK/ Republic of Ireland and AI) and AEOLIAN (UK/ US: AI for Cultural Organisations), bring together Digital Humanists, Computer Scientists, archivists, librarians and other stakeholders.³

For this article, our team of Digital Humanists conducted a series of interviews with professionals in libraries and archives in the UK, Ireland and the US. We investigated the key obstacles to access to born-digital and digitised archives, from the perspective of custodians of collections. The main bottlenecks are well-known (including data protection legislation and copyright), but our interviews revealed

¹ Born-digital records are defined as content produced in digital form, rather than having been digitised from physical form.

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.

³ www.aura-network.net and www.aeolian-network.net.

other obstacles to access—including risk aversion that pushes many institutions to protect their own reputations and interpret legislation in a very restrictive way. Our interviewees were often frustrated by this institutional position, and told us about the need to find a better balance between access and legal requirements. To address these obstacles, interviewees identified possible solutions to make digital archives more accessible to users. However, many bemoaned the fact that scholars and other users seldom seem interested in the issues of archives in the digital age, and often use very traditional research methods in paper collections instead of engaging with more recent sources.

The first section of this article briefly looks at existing research on the issue of access to born-digital and digitised collections. The next section turns to our methods and approach following our series of interviews. It also discusses results, and argues that researchers often lack a voice in debates on access to digital collections. Recommendations are made so that scholars and other users “sit at the table” and contribute to shaping access policies to digital archival collections.

In sum, this paper makes the following contributions:

- Based on our interviews with archivists, librarians and other professionals in cultural institutions, we identify key obstacles to making digitised and born-digital collections more accessible to users.
- We outline current levels of access to a wide range of collections in various cultural organisations, including no access at all and limited access (for example, when users are required to travel on-site to consult documents).
- We suggest possible solutions to the problems of access—including the ethical use of Artificial Intelligence to unlock “dark” archives inaccessible to users.
- We propose the creation of a global user community who would participate in decisions on access to digital collections.

Related work

In the past three decades, there have been numerous scholarly articles and other outputs on digital preservation. Scholarship often originates from archival studies, while other fields (including Digital Humanities) seldom participate in this discussion. This relative silence of users has had consequences on the way in which debates over digital collections has been framed. Indeed, the focus has been mostly on preservation, with access to digital collections more rarely discussed. One of our interviewees noted: “I’ve felt for a very long time the emphasis on preservation, without thinking about how you provide access was... a start-up failure in the way that the whole question was conceived from the beginning” ([anon.] 2021a).

In the mid-1990s, the archival community started devising strategies to preserve digital materials—including archival emails.⁴ The central narrative was then on endangered materials, at risk of disappearing due to neglect and technological

⁴ See, for example, Bearman (1994), Cook (1994), Ross and Gow (1999).

obsolescence. Growing concerns over a digital dark age led in 2002 to the creation of the Digital Preservation Coalition (DPC), established as a partnership between several agencies operating in the UK and Ireland. In the late 2000s, collaborations between archivists and scholars resulted in the creation of open-source digital library tools for content curation such as BitCurator.⁵ However, these examples of collaborations have remained exceptional and have seldom touched on the issues of access to “dark” archives.

Following the work of the Email Task Force, the DPC report *Preserving Email* (Prom 2019) points out that preservation is no longer the challenge it once was. The focus has now moved from preservation to appraisal and selection of materials of lasting value. The RATOM project is developing Natural Language Processing for appraisal and processing of email archives.⁶ The UK National Archives is also in the process of using Artificial Intelligence to appraise and select government records for permanent preservation as the historical record.

There is no point preserving archival records that cannot be used, immediately or at a later stage. In a recent article on national archives, Sara Martínez-Cardama et Ana Pacios found that “access” and “use” were listed by 14 and 12 institutions as priorities. In practice, however, major obstacles—including privacy, copyright or technical issues – make digital cultural data very difficult to access. In the case of email archives, for example, privacy concerns often lead to the closure of entire collections.⁷ When the British writer Ian McEwan sold his archive to the Harry Ransom Center in Texas, he included seventeen years of emails, from 1997 to 2014. The finding aid includes a brief mention of McEwan’s email correspondence, which “has not been processed and is not available to researchers at this time.”⁸ Many collections are hidden, and users are not even informed of the existence of digital records. For instance, the archival emails of the writer Will Self at the British Library are not listed in the finding aid describing the collection, and they are not available to users either onsite or offsite. The fact that very few scholars have had access to email archival collections lead to a gap in scholarship on this topic. Privacy and copyright concerns can also lead to difficulties at the publication stage. In a presentation at the 2020 conference on “Archives, Access and AI,” Sophie Baldock explained that she had privileged access to the email collection of the British poet Wendy Cope at the British Library when she was a collaborative doctoral student (Baldock 2020). Yet, her research proved difficult to publish since email correspondence often contain private and sensitive information. Obtaining permissions to publish requires support from the author of the emails, but also third parties mentioned in the emails. It is not surprising that researchers often choose to focus on more accessible collections, with data easier to quote in scholarly outputs.

⁵ See (Kirschenbaum et al. 2010). This report became a cornerstone of the partnership that developed the BitCurator system now widely used by digital archivists.

⁶ <https://ratom.web.unc.edu/about/>. Accessed 17 Nov. 2021

⁷ See Jaillant and Johnston et al. for a discussion on access to email archives.

⁸ <https://norman.hrc.utexas.edu/fasearch/findingAid.cfm?eadid=01073>. Accessed 17 Nov. 2021.

Unlike archival emails, digitised collections are often perceived as easier to access. Yet, collections that were previously accessible can be closed due to privacy and copyright issues. An emphasis on individuals' right to privacy can thus be found in Michelle Moravec's article on feminist research practices and digital archives, which focuses more particularly on the British Library's digitization of the feminist magazine *Spare Rib* in 2013. "Have the individuals whose work appears in these materials consented to this?," asked Moravec (186).⁹ In December 2020, the British Library announced that it would close down the *Spare Rib* resource once the UK leaves the European Union. Indeed, the resource had been made available on the basis of the EU orphan works directive. In the context of Brexit, the changing copyright framework led to the decision to close access to the entire run of *Spare Rib* magazines.¹⁰

Archivists are increasingly paying attention to the issue of access. In 2020, the Born-Digital Archives Working Group (a group of archive professionals mostly based in the US) released "Levels of Born-Digital Access"—a set of benchmarks and practical guidelines supporting access to born-digital materials. "The Access Levels were created by practitioners with practitioners in mind," declared the report. "They are intended to be used and referenced by those who possess a baseline understanding of digital archives tools and concepts" (Peltzman et al. 2020). Created by and for digital archivists, the report outlines several levels of access. At Level 1, researchers need to travel on-site to consult materials on a public access computer. At the other end of the spectrum, records can be accessed remotely and analysed using sophisticated tools. In other words, users can look at archives from their desk at home and apply computational methods to digital records.

Although this work is valuable, more engagement with scholars and other users is urgently needed. Not all users of digital collections have the same needs. Users with advanced level computational skills are a minority among Humanities scholars and non-academic users. Instead of one-way guidelines designed by archivists and targeted to users, it is essential to co-design strategies to make digital archives easier to access.

Methods, approach and results

Inaccessible archives have an obvious impact on researchers, and yet, the literature on digital preservation and access has been written by archivists—not users. In this article, we have addressed this gap in scholarship with a cross-disciplinary perspective. While Digital Humanists have rarely engaged with the issue of closed digital archives, our research approach was to put researchers directly in contact with archivists and other professionals in libraries, archives and museums. After obtaining approval from the Ethics Committee at our institution, our

⁹ For more information on privacy concerns in relation to digitized collections, see LeClere (2018) and Hart et al. (2019).

¹⁰ See <https://journalarchives.jisc.ac.uk/britishlibrary/sparerib>. Accessed 17 Nov. 2021.

team of three Digital Humanities scholars conducted 21 interviews with 26 professionals in archives, libraries and museums. Each interview lasted between thirty minutes and one hour, and the online format allowed us to reach professionals based in the UK, Ireland and the United States. Most interviewees agreed to be named in the resulting article and other research outputs.

Selection of interviewees was done through existing personal contacts. When the people we approached were not available, they often suggested other colleagues, and this snowballing technique led to additional interviews. Recruiting interviewees via existing networks can distort results—especially when all contacts have the same kind of background. We have therefore paid attention to including a wide range of profiles in terms of gender (with 12 women and 14 men), career stage (from early career to retired professionals), and geographical location. We interviewed practitioners at the National Library of Scotland, the National Library of Wales, the National Library of Ireland and Ivy League institutions on the US East Coast (Yale, Harvard). Our interviewees were based at large metropolitan cultural institutions, such as the British Library or The National Archives UK, but also at smaller institutions, such as Seven Stories (a museum that specialises in children’s books in Newcastle-upon-Tyne) or the Irish Traditional Music Archive in Dublin.

Despite our attention to the diversity of our sample, we acknowledge that large cultural institutions are over-represented. This is due in part to the fact that these institutions have access to financial resources and skilled staff that allow them to actively manage their born-digital and digitised collections. For smaller collections, digitisation programmes or strategic engagement with born-digital resources are often out of reach. Rachel Hosker (Archives Manager and Deputy Head of Special Collections, Centre for Research Collections, University of Edinburgh) recognised this divide between relatively privileged and less privileged institutions. “We have skills but a lot of places don’t have developers to work with,” Hosker pointed out (Hosker 2021).

The format of the interviews was semi-structured, with list of questions sent to interviewees in advance, and follow-up questions asked during the interviews. Although we tailored the questions to each interviewee, our template questionnaire was as follows:

Background and experience

1. Could you tell me more about your background? What did you do before becoming [job title]?
2. Could you describe the kind of collections held at your institution?
3. How many researchers visit your institution each year (approximately)? Do you know if these researchers are mostly academics or members of the public?
4. Could you describe how your institution responded to the challenges brought by the COVID pandemic?

Digitised collections

5. Could you tell us more about current digitisation projects at your institution? How many collections are now available in digital form?
6. Are all the digital collections available remotely? Or are there restrictions to access (need to travel onsite, to log in and provide details, etc.)?
7. If you look at the broad picture, what are the challenges and prospects of digitised archives in the UK and elsewhere?

Born-digital collections

8. As you know, born-digital collections are rarely available to users for many reasons. Could you describe the situation at your institution in terms of access to these archives?
9. How can we make born-digital collections more accessible to users?
10. Do you have any partnerships with other institutions, in the UK and elsewhere, related to born-digital projects and collections?
11. What are the challenges and prospects of born-digital archives in the UK and elsewhere?

Any other business

12. 12. Do you have anything else you would like to tell us?

Interviewees' responses to these questions allowed us to identify common problems that prevent access to digitised and born-digital collections; to ascertain current levels of access to selected collections (for example web archives or email archives) in specific institutions; to probe possible solutions to the problems of access, including the use of advanced technologies, such as AI and ML. Based on these results, we argue that there is a need for more engagement between archivists and users of collections, to unlock closed cultural assets.

Obstacles to access to digitised collections

Digitised collections constitute a minority of all records in cultural institutions, not only because of the cost and technical expertise needed to produce high-quality digital copies and metadata, but also because not all collections can be digitised in the first place.

First, technological obsolescence makes it difficult to find equipment to read records in old formats. For example, players of magnetic tapes in VHS or Betamax format are becoming increasingly rare and may be at risk of disappearing almost completely in three or four years according to Stuart Lewis (Associate Director of Digital, National Library of Scotland). Without players, “even if the media’s in good condition, we wouldn’t be able to digitise them” (Lewis et al. 2021). The NLS recently got a business case approved to digitise 10,000 magnetic tapes, which will

significantly increase the data currently stored due to the size of audio-visual files compared to paper-based digitisation.

Second, digitisation programmes rely on the selection of materials that are deemed high priority – and on the exclusion of records seen as less important or valuable. “There’s a big discussion... around which communities, which records get digitised, who gets represented and not represented in digitisation strategic decisions,” said Anthea Seles (Secretary General, International Council on Archives) (Seles 2021). For example, in the wake of George Floyd’s murder, Dorothy Berry (Digital Collections Programme Manager at Harvard’s Houghton Library) realised that there had been very little digitisation of African American holdings. She then led a project to digitise 2,000 records ([anon.] 2021b).¹¹

Copyright is a third reason why certain collections are not digitised, or not made available when digital copies exist. Treasa Harkin (Librarian, Irish Traditional Music Archive) pointed out that the ITMA has developed over the years a very strong level of trust among the traditional music community: “part of that is built on the fact that you give something to the archive you know it’s safe and you know it’s not going to turn up on the website or online somewhere” (Harkin and O’Connor 2021). The Archive tries to strike a balance between keeping that trust and making more digitised materials available to users. This requires negotiating with copyright holders to obtain permissions, and investigating different levels of access (for example, by providing online access only to registered users with a login and password).

Obtaining permissions becomes extremely difficult when copyright holders or their Estates have disappeared without leaving any traces. Stuart Lewis notes that the National Library of Scotland does not necessarily have permissions to digitise materials from the people that donated or deposited their archives. “Twenty-five years ago, we wouldn’t have asked, ‘Can we digital this, please and put it online?’” (Lewis et al. 2021). The difficulty to obtain copyright permissions retrospectively explains why the NLS does not prioritise the digitisation of manuscript materials, preferring to focus on less problematic records—such as public domain items, which constitute the largest part of the digitised collections available on the NLS’s Data Foundry website.¹²

Copyright issues with materials deposited before the digital age impact other archives, including the Fortunoff video archives at Yale University Library. When this collection of testimonies from Holocaust survivors was created, interviewees gave their permissions to be recorded but not for the interviews to be available anywhere on the internet. This makes the collections less accessible to researchers such as Abby Gondek (Morgenthau Scholar-in-Residence, FDR Library and Roosevelt Institute): “there’s so many survivor testimonies but you can’t listen to them—you can only listen to titbits unless you go either to that archive, or they have, like, satellite sites that they have a relationship with, that you can listen to the recordings at those places” (Gondek 2021).

¹¹ See also <https://news.harvard.edu/gazette/story/2020/07/houghtons-2020-21-digitization-focus-black-american-history/>. Accessed 19 Nov. 2021.

¹² <https://data.nls.uk/data/digitised-collections/>. Accessed 19 Nov. 2021.

Data protection is a fourth major obstacle to increasing access to digitised materials that contain sensitive materials. UK and EU archival institutions need to comply with the General Data Protection Regulation (GDPR) in the European Union and Data Protection Regulation 2018 in Britain. Preventing the release of sensitive materials is a central concern for the Wellcome Collection in London, an institution with confidential records related to health and medicine—including individuals' medical records. The archive also contains other kinds of confidentiality, including business confidentiality in organisational records. As a rule, the Wellcome does not digitise materials that are less than ten years old. For other records, archivists adopt a risk-management approach: they take a proportion of a particular series and assess it first, and if it turns out to be very sensitive, they will then assess more materials of that series. In some cases, it is necessary to do a more detailed item-by-item review to identify records that are culturally sensitive, or have offensive language in it, or are simply unpleasant to look at (such as images of surgery). These records can be digitised and made available online, but users will receive a warning of what they are about to see ([anon.] 2021a).

With very large archives, sensitivity review can be a Sisyphean task. It is always possible that collections that were deemed non-sensitive turn out to contain problematic materials. The National Library of Scotland has a series of safety checks that restricts digitisation to collections that have been screened for data protection and copyright issues and fulfil additional conditions. “To go for digitisation, one of the criteria is it needs to be in final order so that it's arranged, described and foliated so there's a referencing system [and] somebody can refer back to a particular folio within a particular context,” notes Stephen Rigden (Digital Archivist at the NLS) (Lewis et al. 2021).

Following the “More Product, Less Process” movement (Greene and Meissner 2005), some would argue that these conditions are excessive, and that it is always better to digitise and make records available even when collections are not fully arranged and may still contain problematic information. No collection is ever perfectly organised, and no one can guarantee that all sensitive materials have been removed. As Dennis Meissner and Mark Greene claimed in a 2010 article, “our greater ethical vulnerability may come from withholding access. Any vulnerabilities resulting from exposing materials that are ‘sensitive,’ but not contractually proscribed, pale in comparison” (204).

Born-digital collections: from creation to access

Unlike collections that originated in paper before being digitised, born-digital records are rarely acquired single-handedly but instead come to archival institutions as part of hybrid collections. There is still no market for born-digital records. As one interviewee told us: “if somebody offers me originals and then somebody else comes back and offers me photocopies of those originals, I would not purchase the photocopies, even though the content is the same” ([anon.] 2021b). Like photocopies, born-digital records give access to content and can easily be duplicated. The

financial value of a collection still relies on the aura of the original record and its connection to the creator.

One of the reasons why Special Collections value original materials is that they can easily be displayed in exhibitions. Engaging with children and their parents using original records is central to the strategy of Seven Stories museum. Collection Manager Krie McKie said:

We hold Enid Blyton's archive here, or part of it, and we have a lot of her typescripts for Famous Five books, Noddy, etc. And it's really easy to present one of those to a group and say this is a typescript by Enid Blyton, and... there's an authenticity to it... But with born-digital material, I don't know how you would generate that same level of excitement or interest in a Word file (McKie 28 May 2021).

Instead of displaying actual born-digital files, some institutions use artefacts such as original computers. For example, Harvard's Houghton Library is planning to display one of the writer Jamaica Kincaid's laptops in a future exhibition ([anon.] 2021b). The machine that the writer touched during the working process is imbued with an aura that the born-digital files seem to be lacking.

Collections that are acquired or donated today were often created before the digital revolution. With the arrival of a new generation of creators, it is likely that born-digital records will form the largest part of collections in the future. Some institutions have anticipated this moment, and are actively approaching creators who have spent their entire careers using computers rather than typewriters and other tools.

In 2019, the National Library of Ireland thus announced that the bestselling author Marian Keyes (born in 1963) had donated digital materials related to the publication of her novel *The Mystery of Mercy Close*. This includes book cover samples, drafts and preproofs; videos of the author discussing characters in the book, as well as her writing process; and promotional material, social media interactions, and a timeline of the novel.¹³ NLI staff also took photos of Keyes's workspace – in the spirit of *enhanced collection*, a concept developed by the British Library (Keating and Finegan 2021). Enhanced collection seeks to create new content surrounding acquisitions of contemporary archives through various methods—including creating interactive photographs of writers' workrooms, recording interviews with archival creators or documenting video-conversational tours of creators' environments. The initiative was developed to enable researchers to explore additional context, thus re-creating the privileged position of the curator who encounters a new collection for the first time.

Keyes has a strong web presence, and the NLI was already collecting her website and Twitter feed as part of its programme of web archiving. With this born-digital collection, the NLI sends a strong message that records from commercial women writers are worth collecting. It positions itself as a forward-looking institution that is not constrained by models of collecting that have traditionally focused on male

¹³ <https://www.nli.ie/en/list/latest-news.aspx?article=2a51d71e-584f-4779-a3f0-0f3d023cc179>. Accessed 19 Nov. 2021.

record creators, on original paper records, and on strict hierarchies between high and popular culture. “Our director has been a huge champion of diversity and inclusion,” said Della Keating (Assistant Keeper, Digital Collecting, NLI), “and that very much fed into our thought process as well in terms of the role of the library in reflecting society on societal change” (Keating and Finegan 2021). The Marian Keyes archive is part of a series of born-digital pilot projects—which also includes the Yes Equality Visual Digital Archive (on the 2015 marriage equality referendum in Ireland).¹⁴

Transport for London is another example of an institution that has played an active role in approaching creators of records, rather than waiting passively for the collection to be deposited in the archival repository. TFL archivist Tamara Thornhill was instrumental in building a digital collection on the 2012 London Olympics. She worked with tech colleagues to build a suitable platform to host these records. The next step was to approach the teams that had been involved with the Games transport delivery, and to ask them to transfer their records on this online repository. Thornhill’s team also conducted oral history interviews around the staff’s experience of delivering the Olympics. Some senior staff were still in media mode when the interviews were conducted, with a message that gave the impression that everything had been perfectly planned and nothing went wrong. Follow-up interviews were organised two years later, yielding a more nuanced message. In addition to leading the creation of oral history interviews, Thornhill also did “lots of appeals on internal social media, and internal intranet to try and weed out any extra Games born-digital records that are out there, just to make sure, really, that we have got as complete a set as possible. And I feel very confident that we do have everything that, certainly that we need, anyway, and that we should have” (Thornhill 2021).

This active approach is controversial in the archival community. “Archivists tend to look at that slightly askance—are you distorting history there? Or are you interpreting the record?,” said one interviewee who works in knowledge and information management for the UK Central Government ([anon.] 2020). Indeed, the nineteenth-century concept of *respect des fonds* implies a passive role of the archivist, who receives collections from the creators without changing the records and their original organisation. However, born-digital collections often lack any kind of organisation—they can be replicated easily, and sit on multiple devices and platforms without being actively managed. Adopting an active approach to creating born-digital collections can therefore be seen as a pragmatic response to the needs of future researchers. Indeed, users interested in the history of Brexit or the London Olympics will find it easier if records are all in one place and properly indexed, instead of being spread out across several different places.

But as we have seen, most born-digital collections are not open to researchers. This includes the Marian Keyes archive at the NLI, which is currently processed, and the TFL London Olympics archive, with access restricted to a few staff members. What are the key obstacles to access? Like digitised archives, born-digital collections often contain sensitive materials. In practice, many institutions prefer to close access entirely rather than potentially release problematic records. In the

¹⁴ <https://www.nli.ie/en/born-digital-collection.aspx>. Accessed 20 Nov. 2021.

United States, where data protection legislation is much less restrictive than in Europe, there have been cases where the privacy of living individuals named in born-digital archives has been compromised. One of our interviewees said that the case of the Susan Sontag archive at UCLA led him to change his mind on the issue of privacy in archives:

for the early part of my career, I fundamentally didn't believe it was archive's responsibility to maintain privacy and confidentiality. It's like the record is the record, and often because my perception of how that was used was that sensibility was used to suppress really important information. That sensibility was often used by the powerful to prevent information that they didn't want to be released ([anon.] 2020).

The turning point was the publication of a 2014 article in the *Los Angeles Review of Books*, written by two doctoral students at UCLA who got access to Sontag's archival emails and hard drives (Schmidt and Ardam 2014). The article revealed Sontag's relationship with the photographer Annie Leibovitz, which Sontag had denied during her lifetime (she died in 2004). Readers were also informed that Sontag's hard drives contain sensitive documents regarding Leibovitz's surrogate pregnancy. There was nothing illegal in the article, but our interviewee saw it as a deeply unethical invasion of privacy. As a time when corporations gather massive amounts of data from individuals' phones and computers, the interviewee felt that libraries and archives should resist this trend and actively preserve privacy (an argument echoed by Ryan Cordell in his recent report *Machine Learning + Libraries* 2020).

Confidential and sensitive records also concern paper collections, and archival repositories have long managed access by setting a date in the future when records will be made available. Even business archives—that are not legally obliged to make their collections publicly accessible—often allow researchers to consult selected records after a set period. The question of opening up the archive, or part of the archive, is often hotly debated within the company. Let's take the example of the BT Archives in London, which holds records from the communications company dating back to 1846. If records were created before the privatisation of British Telecom in 1984, they are public. Users have the right to view them under the Public Records Act and Freedom of Information Act. However, “there are a few things which still sit in a space of having a government security marking and therefore not yet available,” said James Elder, who worked as Archives Manager at BT for 32 years (Elder and Hay 2021). Records created after 1984 (including born-digital records) are mostly closed. As David Hay, Heritage and Archives Professional at BT, puts it:

It will continue to prove very difficult to open up even the early post-privatisation records for some period, precisely because of the problem with risk that you mentioned is that the lawyers and senior managers will be considering BT is a huge business, incredibly highly regulated. So the regulator and the government have an interest in what we do. We have huge tax issues, we're always in dispute with HMRC. There are always huge numbers of industrial injury claims. Why do we want to put any stuff which could expose BT to any

possible risk at all? It's much easier to keep it closed if legally we're entitled to. (28 May 2021)

Even within this very restrictive access policy, there are plans to make some materials available within a still-to-be-defined timeframe. “You can imagine a situation where the existence of records might go forward on a rolling 30, 40, 50-year basis or whatever,” said Elder. “And then they would be open to researchers on application, and then that would be on the basis of research based on a defined research objective and a sharing of the results before publication with an ability to have some kind of oversight of that, I suspect” (28 May 2021). Due to fears of censorship, many researchers would object to sharing their research outputs with the archival repository before publication. But this is perhaps the price to pay to have access to records that would otherwise be inaccessible.

The long timeframe between the creation of records and their opening up could in theory encourage archivists to make the collections as user-friendly as possible. After all, it is always easier to have plenty of time to process and arrange records, and perhaps design strategies to make non-sensitive parts of the collections available sooner. However, knowing that collections will not be open for another 20 years or more can lead to a wait-and-see attitude. As one of our interviewees told us, “why would you spend a lot of time working out how to provide access to it now, when actually, we cannot legally provide some of that access?” ([anon.] 2021a).

Data protection is not the only obstacle that prevents users from accessing born-digital collections. The scale of these collections can also be a major concern—especially for smaller institutions with limited human and financial resources to store and actively manage enormous amounts of data. Scale often makes born-digital collections less discoverable, as archival repositories struggle to catalogue them (Hosker 2021). It can also make it harder to preserve multiple copies of the archive—an important principle to ensure the sustainability of a collection. There is currently only one copy of the UK Web Archive held by the British Library and shared with the other legal deposit libraries, although the National Library of Scotland will soon hold a second copy (Lewis et al. 2021). Making entire collections accessible is often not an option, even though smaller parts are sometimes released to researchers. For example, the Irish Traditional Music Archive holds terabytes of materials (including videos of music performance). It requires a huge amount of work to put some of these records online, and to make sure that the copies are of sufficient quality (Harkin and O'Connor 2021).

The differences between preservation and access workflows also create pain points in the user experience. Euan Cochrane (Digital Preservation Manager, Yale University Library) explains that born-digital records are acquired and processed in bulk. At the preservation stage, archivists who check the content can take a disc image of the entire media and store it. In contrast, the access workflow is much more tortuous and time-consuming. “If someone requests access, [archivists] have to get the files out of the disc image, figure out what they are, review them, maybe even catalogue them a bit because they're not catalogued enough” (Cochrane 2021). In short, the access process is much more difficult to automate than the preservation process.

Designing an appropriate interface for users to access born-digital materials is a major challenge for many archival institutions, since they must ensure that users will not be able to change records or delete them (Hosker 2021). As with other aspects of born-digital curation, controlling the way in which users access digital files requires advanced technological skills. Callum McKean (Lead Curator, Contemporary Archives and Manuscripts, British Library) said: “You don’t just need curators and technologists, and that’s useful a lot of the time as well, but I think you need people who can kind of move between those worlds and translate between those worlds to do it effectively” (McKean 2021). McKean himself did an Applied Data Science course, to learn coding techniques to deal with digital collections. This sort of training is still unusual for cultural heritage professionals. For McKean, developing those skills among staff or hiring for those skills should be a key priority for the sector.

Others, such as Mark Bell (Senior Digital Researcher, The National Archives UK), think that coding skills are not essential for all practitioners who deal with digital collections. Even staff with an interest in Artificial Intelligence applied to archives do not necessarily need to know programming languages. However, they do need “skills in interpreting data outputs, whether it’s graphs or probabilities or... numerical or visual outputs” (Bell 2021). To summarise, there is no consensus on the form that technological training should take, but professionals, such as McKean and Bell, agree that the sector has a long way to go to upgrade its skills base.

Current levels of access to digital collections

Data protection, scale, and skill gaps: these central obstacles explain why so many born-digital collections are completely inaccessible to researchers. Sally McInnes (Head of Unique Collections and Collections Care, National Library of Wales) notes that the email archives at the NLW are “sitting on the network somewhere”: “we haven’t started to address them, but we will” (McInnes 2021). These collections are completely hidden and undiscoverable by researchers. The same happens with very sensitive email collections at Yale University Library. Cochrane gives the example of a plastic surgeon who deposited his records just before his death. “That included his inbox, his email inbox, which had hundreds of thousands of emails,... a lot of which had patient data associated with them. So, you can imagine—what do you do with that” (Cochrane 2021).

Even when born-digital and digitised collections are accessible, users often need to go on-site to view documents—which limits access to those who have time, funding and the physical availability to travel. For example, Seven Stories has a born-digital collection relating to the illustrator Chris Fulton, which includes Photoshop files, PDFs, JPEGs and other file formats. Researchers need to travel to Newcastle to consult these digital illustrations, in part because the archival institution does not have the capacity to make them available online (McKie 2021). At Transport for London Archive, materials can only be accessed on-site. Not all materials are discoverable, however. As Tamara Thornhill notes, “in our digital systems, we only have the option of open or closed, and we don’t have the option of just pushing through the metadata” (Thornhill 2021). In practice, this means that archivists

do not have the middle-ground option of sharing only the metadata, leaving many records undiscoverable.

In the case of web archives in Britain, the Non-Print Legal Deposit Regulations 2013 also require users to travel onsite to consult materials, which were once publicly available on the internet. The largest part of the collection can only be viewed at the six legal deposit libraries located in nine places: the British Library (London St Pancras and Boston Spa), National Library of Wales (Aberystwyth and Cardiff), National Library of Scotland (Edinburgh and Glasgow), Bodleian Library (Oxford), Cambridge University Library, and the Library of Trinity College, Dublin. Additionally, 19,000+ websites—a small proportion of the 5 to 10 million different websites that are probably in the UK sphere—have permission from their owners to be viewed from anywhere via the UK Web Archive website ([anon.] 2021a).¹⁵

The situation is very different in the United States, where the Internet Archive regularly harvests websites and makes them publicly available online. The law on fair use in the US is much more liberal than the non-print legal deposit regulations in the UK. While the British legal framework is highly restrictive on access, it also provides legal deposit libraries with a clear mandate to collect materials that may otherwise disappear. One interviewee said: “the fact that we have a legal remit and obligation to collect these things is... a good thing because the British Library and other libraries like it tend to think very long term, and I think there’s a reasonable chance that we’ll be here in 50 years’ time, 100 years’ time” ([anon.] 2021a). Others are more critical of this situation, which requires users to access UK web archives onsite using a Library PC. William Kilbride (Executive Director of the Digital Preservation Coalition) notes: “Very few people even know it’s there, or how they would go about using it” (Kilbride 2021). For Kilbride, there is no preservation without access, since it is essential to understand properly the needs and expectations of the user community.

In the case of digitised collections, contracts with commercial companies often leave cultural institutions unable to widely share digital copies of the original materials they hold.¹⁶ Typically, companies will create digital copies that are then placed behind a paywall. Libraries and archives get free digital copies that they can make accessible on-site, but not on their websites and other online platforms. “Essentially what we’re doing is we’re selling our collections to someone else that we can’t then put online,” said Stuart Lewis. The National Library of Scotland is planning to move to another model: “we’re actually trying to say to them, *We will do the digitisation and you pay us to do the digitisation*, because quite often they’d outsource it to a third party and we were never quite happy with their quality of digitisation. So we’re basically trying to take on that commercial role ourselves instead” (Lewis et al. 2021).

The pandemic particularly highlighted this issue of access to digitised collections. As one of our interviewees pointed out, there has not been open access to a

¹⁵ See also <https://www.bl.uk/collection-guides/uk-web-archive>. Accessed 20 Nov, 2021.

¹⁶ For more on commercial digitisation companies (focusing particularly on the case of digital newspaper archives), see Hauswedell et al. (2020), Milne (2002), Smith (2006), Terras (2011), and Gooding (2014).

large proportion of our digitised cultural heritage, because of the arrangements that were made to digitise it in the first place. Although libraries and archives receive a revenue stream from these commercial contracts in the form of royalties, “they’re not seeing the full economic value of the material” that they provided for digitisation projects ([anon.] 2021a). Users are also limited in the kind of research they can do with these digitised materials. Metadata are not openly accessible, which limits the potential of researchers to analyse the substantial quantity of digitised materials that has been created in the past two decades or so.

When cultural organisations were closed during the pandemic, users (and their institutions, in the case of academic researchers) had no other option than to pay for digital copies provided behind paywalls. Paul Seaward, the Director of the History of Parliament Trust, points out that some of the key collections held at the British Library and The National Archives have been digitised by ProQuest. “It’s only accessible through highly expensive... commercial online publication, so that is an issue” (Seaward 2021).

Possible solutions to the problems of access

The problem of access to digital collections will not be easy to solve. Take the example of the non-print legal deposit regulations. Several interviewees pointed out that this legislative framework is not adequate at a time when Open Access has become a major driving force. It makes little sense to restrict access to web archives with the potential to bring valuable research findings, which researchers will then be required to publish in open access journals. Even if current legislation becomes less restrictive, many cultural institutions may well remain risk adverse—which will frustrate many researchers who are denied access to important records. The risk of releasing problematic materials is seen as particularly high for cultural institutions eager to preserve century-long reputations. “We want to be an organisation to be trusted. We think we are already—we have high levels of trust, and we want to maintain those,” declared one interviewee ([anon.] 2021a).

Making all digital collections open access on the internet remains a utopia, and some institutions are taking the more pragmatic approach of building virtual reading rooms for registered users. Yale University Library is currently working on an online environment where users could view restricted content using a web browser on their computer. They would have the option to use their own software, but not to download the materials (Cochrane 2021). At The National Archives UK, Mark Bell is also interested in secure virtualised environments which could offer different levels of access. This is based on his experience of a data study group on the web archive organised with the Alan Turing Institute. Since Bell and his colleagues were working on non-sensitive materials, they had the lowest level of security on their virtual environment. This allowed them to add new data sets, to copy and paste, and the like. These functionalities were gradually lost at higher levels of security. On the highest level, it was not possible to take a screenshot, to put in or to take out data. The environment was very locked down, but it still allowed users to work with the data. For Bell, this model that gives access to a computer, with heavy restrictions on

what you can do with it, is “a way to go” to solve the problem of totally inaccessible data (Bell 2021).

A growing number of institutions are exploring the use of Artificial Intelligence and Machine Learning to make their collections more discoverable, and accessible to researchers. Automatically creating metadata is a task that could be undertaken by AI systems. The Irish Traditional Music Archive is thus seeking to use AI/ML to name tunes. For example, a reel-to-reel audio tape recorded in the 1960s could be digitised as an hour and a half of music, containing 50–60 tunes. In the case of collections containing hundreds of reel-to-reels, AI could enable the creation of a dataset of tunes, which would then be made available to researchers (Harkin and O’Connor 2021).

Web archives offer another example of collections with very minimal metadata. This might include the title of the webpage, but nothing more. AI/ML could categorise websites in broad categories, such as politics or culture. This would in turn help researchers who could focus on the categories they are interested in ([anon.] 2021a). In addition, Artificial Intelligence can be used to identify controversial metadata in historical collections, including paintings that have been photographed and then digitised. Antoine Isaac (Research and Development Manager, Europeana) points out that the objective would not be to hide unpleasant historical events: “if we’ve got a painting representing some slave person, we’re not going to be able to hide the fact that slavery has existed” (Isaac 2021). But growing awareness of these controversial terms could lead to new ways of describing collections—at a time when some museums are starting to change the title of their artwork, for instance to focus on the individuality of the enslaved person.

Another use of AI/ML in archival collections is to predict file types. Santhilata Venkata (Digital Preservation Specialist, The National Archives UK) notes that her institution often receives corrupted data records which are text files. Artificial Intelligence can automatically examine the contents of the file, and then make prediction on file types—which is useful in the context of digital preservation and access (Venkata 2021).

Improving metadata and predicting file types are of course very important, but perhaps the most important application of AI/ML to archives is its ability to identify sensitive materials. Sensitivity review is already a feature of ePADD, a free and open-source software developed by Stanford University’s Special Collections & University Archives that supports the appraisal, processing, preservation, discovery, and delivery of historical email archives. But not everyone is convinced that ePADD’s Natural Language Processing feature is enough to separate problematic from non-problematic materials. Callum McKean, who has done curation work on the archive of the British poet Wendy Cope, said: “You have things being flagged up by ePADD like her shopping list comes under drugs because she buys mushrooms with the shopping and... the lexicon classes the word, mushroom, as a drug, so you get lots of false positives” (McKean 2021). On the opposite, ePADD also returns false negatives because Cope provides very personal information about other writers in a coded way that the machine is unable to decrypt.

With the development of Artificial Intelligence and Machine Learning in libraries and archives, it is important that users make their voices heard and shape the ways in

which collections are accessed and used. As we have seen, professionals in libraries and archives can use AI/ML for metadata extraction, or to help with the process of cataloguing and characterisation. Differentiation between sensitive and non-sensitive records is another important application of this new technology. But Artificial Intelligence and Machine Learning can also be used by researchers, to analyse huge amounts of records (for example in web archives). These technological advances require access to data. Without access, new forms of knowledge will not be possible.

While AI and ML have the potential to assist with tasks, such as sensitivity checks and metadata tagging, technological advances alone will not solve the issue of access. The legislative constraints around copyright and data protection tend to paralyse access regimes for digital holdings. It is possible that these laws will evolve to encourage open access to records that are not particularly sensitive or confidential. In Britain, the National Data Strategy stresses the need for data to be “appropriately accessible, mobile and re-usable.”¹⁷ This strategy sets out how best to unlock the power of data for the UK. The British government’s focus on Artificial Intelligence and data science as key priorities to make data more accessible sets the context for the UK AI Council’s AI Roadmap¹⁸ and UKRI AI Roadmap,¹⁹ both released in early 2021. In particular, the UKRI AI Roadmap stresses the problem of “poor access to data” and the need for more research “across disciplines,” including the Humanities and the Sciences.

Two decades ago, the growing awareness of a digital “dark age” led to the creation of the Digital Preservation Coalition in the UK and Ireland. It has since grown into a global organisation, with more than 100 institutional members. Moving the focus from preservation to access, the global user community we propose would aim to co-design practical solutions to the problem of “dark” archives, in partnership with archivists, librarians and other professionals. This community could address the major limitations to access based on legislative constraints—for example by lobbying to relax copyright laws and increase exceptions in legislation for research purposes. The community could also campaign for additional resources for digitisation and copyright clearance. As Mark Bell of The National Archives UK puts it, “it’s going to be really important for us to work as a community on all these challenges.... Whether that’s ideas or the technology itself, or platforms, or just sharing models or data, I think there needs to be a lot of sharing” (Bell 2021).

Funding The funding has been received from arts and humanities research council with Grant no. AH/V002341/1.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative

¹⁷ <https://www.gov.uk/government/publications/uk-national-data-strategy>. Accessed 18 Nov. 2021.

¹⁸ <https://www.gov.uk/government/publications/ai-roadmap>. Accessed 18 Nov. 2021.

¹⁹ <https://www.ukri.org/about-us/what-we-do/strategies-and-reviews/ai-review-transforming-our-world-with-ai>. Accessed 18 Nov. 2021.

Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- [Anon.] (2020) Interview with author on 2 Nov 2020
- [Anon.] (2021a) Two interviews with author on 27 May 2021
- [Anon.] (2021b) Interview with author on 28 May 2021
- Baldock S (2020) Wendy Cope's email archive at the British Library, presentation at the conference archives, access and AI: working with born-digital and digitised archival collections. <https://www.poetrysurvival.com/programme-aracai-conference/>. Accessed 17 Nov. 2021
- Bearman D (1994) Managing electronic mail. *Arch Manusc* 22: 28–50. <https://publications.archivists.org.au/index.php/asa/article/view/8385/8379>. Accessed 12 July 2021
- Bell M (2021) Interview with author on 27 May 2021
- Cochrane E (2021) Interview with author on 26 May 2021
- Cook T (1994) Electronic records, paper minds: the revolution in information management and archives in the post-custodial and post-modernist era. *Arch Manusc* 22: 300–28. <https://publications.archivists.org.au/index.php/asa/article/view/8433/8427>. Accessed 12 July 2021
- Cordell R (2020) Machine learning + libraries. Library of Congress, Washington, D.C. <https://labs.loc.gov/static/labs/work/reports/Cordell-LOC-ML-report.pdf?loclr=blogsig>. Accessed 14 July 2021
- Elder J, Hay D (2021) Interview with author on 28 May 2021
- Gondek A (2021) Interview with author on 20 May 2021
- Gooding PM (2014) Search all about it : a mixed methods study into the impact of large-scale newspaper digitisation. University College London (University of London), Ph.D.
- Greene M, Meissner D (2005) More product, less process: revamping traditional archival processing. *Am Arch* 68:208–263. <https://doi.org/10.17723/aarc.68.2.c741823776k65863>
- Harkin T, O'Connor L (2021) Interview with author on 20 May 2021
- Hart TR, de Vries D, Mooney C (2019) Australian law implications on digital preservation. In: Presentation at the IPRES conference.
- Hauswedell T, Nyhan J, Beals MH, Terras M, Bell E (2020) Of global reach yet of situated contexts: an examination of the implicit and explicit selection criteria that shape digital archives of historical newspapers. *Arch Sci* 20:139–165. <https://doi.org/10.1007/s10502-020-09332-1>
- Hosker R (2021) Interview with author on 19 May 2021
- Isaac A (2021) Interview with author on 28 May 2021
- Jaillant L (2019) After the digital revolution: working with emails and born-digital records in literary and publishers' archives. *Arch Manusc* 47:285–304. <https://doi.org/10.1080/01576895.2019.1640555>
- Johnston JA, Wallace DA, Punzalan RL (2019) Messages sent, and received? Changing perspectives and policies on US federal email as record and the limits of archival accountability. *Arch Sci* 19:309–329. <https://doi.org/10.1007/s10502-019-09318-8>
- Keating D, Finegan J (2021) Interview with author on 28 May 2021
- Kilbride W (2021) Interview with author on 19 May 2021
- Kirschenbaum M, Ovenden R, Redwine G (2010) Digital forensics and born-digital content in cultural heritage collections. *CLIR*. <https://www.clir.org/pubs/reports/pub149/>. Accessed 12 July 2021
- LeClere E (2018) Breaking rules for good? How archivists manage privacy in large-scale digitisation projects. *Arch Manusc* 46:289–308. <https://doi.org/10.1080/01576895.2018.1547653>
- Lewis S, Hamilton G, Rigden S (2021) Interview with author on 19 May 2021
- Martínez-Cardama S, Pacios AR (2022) National archives' priorities: an international overview. *Arch Sci* 22:1–42. <https://doi.org/10.1007/s10502-021-09367-y>
- McInnes S (2021) Interview with author on 26 May 2021
- McKean C (2021) Interview with author on 28 May 2021
- McKie K (2021) Interview with author on 28 May 2021

- Meissner D, Greene M (2010) More application while less appreciation: the adopters and antagonists of MPLP. *J Arch Organ* 8:174–226. <https://doi.org/10.1080/15332748.2010.554069>
- Milne R (2002) The research support libraries programme, the distributed national collection, and the importance of cross-sectoral collaboration. In: McKitterick D (ed) *Do we want to keep our newspapers?* pp 81–85
- Moravec M (2017) Feminist research practices and digital archives. *Aust Fem Stud* 32:186–201. <https://doi.org/10.1080/08164649.2017.1357006>
- Peltzman S, Dietz B, Butler D, Walker P, Farrell J, Arroyo-Ramirez E, Macquarie C, Bolding K, Helms AM, Venlet J (2020) Levels of born-digital access. Digital Library Federation (DLF) part of the Council on Library and Information Resources (CLIR), Alexandria, VA. <https://osf.io/r5f78/>. Accessed 12 July 2021
- Prom C (2019). *Preserving Email*, 2nd Edition. Digital Preservation Coalition. <https://doi.org/10.7201/twr19-01>
- Ross S, Gow A (1999) *Digital archaeology: rescuing neglected and damaged data resources: a JISC/NPO study with the Electronic Libraries (eLib) Programme on the preservation of electronic materials*. Library Information Technology Centre, London
- Schmidt J, Ardam J (2014) On Excess: Susan Sontag’s Born-Digital Archive. *Los Angeles Review of Books*. <https://lareviewofbooks.org/article/excess-susan-sontags-born-digital-archive/>. Accessed 14 July 2021
- Seward P (2021) Interview with author on 28 May 2021
- Seles A (2021) Interview with author on 28 May 2021
- Smith N (2006) Digitising documents for public access. In: MacDonald L (ed) *Digital heritage: applying digital imaging to cultural heritage*. Butterworth-Heinemann, Oxford, pp 3–32
- Terras M (2011) The rise of digitization: an overview. In: Rikowski R (ed) *Digitisation perspectives*. Sense Publishers, Rotterdam, pp 3–20
- Thornhill T (2021) Interview with author on 27 May 2021
- Venkata S (2021) Interview with author on 21 May 2021

Publisher’s Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Lise Jaillant is Senior Lecturer (Associate Professor) in Digital Humanities at Loughborough University. She has a background in publishing history and digital humanities. In the past five years, she has gained expertise on born-digital archives and the issues of preservation/ access to these archives. Since 2020, she has been UK PI for three externally funded projects on Archives and Artificial Intelligence: (1) “EyCon (Visual AI and Early Conflict Photography)” (2) “AEOLIAN: Artificial Intelligence for Cultural Organisations” (3) “AURA (Archives in the UK/ Republic of Ireland & AI): Bringing together Digital Humanists, Computer Scientists & stakeholders to unlock cultural assets.” These international projects aim to make digitised and born-digital archives more accessible to researchers, and to use innovative research methods such as AI to analyse archival data.