Archival Science (2022) 22:285–291
https://doi.org/10.1007/s10502-022-09396-1

EDITORIAL

Introduction: challenges and prospects of born-digital and digitized archives in the digital humanities

Lise Jaillant[1] · Katie Aske[2] · Eirini Goudarouli[3] · Natasha Kitcher[2]

Published online: 26 May 2022

The scale and complexity of digital archives, both born-digital and digitized, are posing enormous challenges for both researchers and memory institutions. In the world of archives, these new types of records are fundamentally changing the landscape as well as the role of archivists and archival institutions. The emergence of new generation technologies also brings a variety of complexities and challenges to archival frameworks, requiring new capabilities and approaches on how best to capture, preserve, contextualize and present the increasingly born-digital and digitized records. However, this technological shift also brings new opportunities for research and experimentation (Goudarouli et al. 2019). For example, technological developments have transformed the way researchers can access and explore archival collections. The digitization of archival materials has opened a variety of large-scale digital collections to the world. Additionally, born-digital archives are beginning to reach terabytes, comprising many different types of media, that can be made accessible online.

By enabling the extraction of archival content as data and moving towards the creation of aggregated large-scale datasets, memory institutions are focusing on providing access to their collections in new ways inviting new explorations and interpretations of their materials by researchers from across the world. In addition, today's reading rooms have been transformed to become more user-friendly than they were

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

Katie Aske
K.L.Aske@lboro.ac.uk

Eirini Goudarouli
eirini.goudarouli@nationalarchives.gov.uk

Natasha Kitcher
N.Kitcher@lboro.ac.uk

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

[2] Loughborough University, Loughborough, UK

[3] The National Archives UK, London, UK

in the 1970s and 1980s, as online finding aids and search tools have replaced card catalogues, making it easier to identify materials. Digital cameras allow users to take their own pictures instead of relying on archival repositories for photocopying; and laptops have enabled the exploration of online collections remotely and away from the physical reading rooms. Twentieth-century historians, literary scholars and other humanities researchers are still exploring old papers in archival collections, as the collections that have been digitized only represent a small portion of all archival holdings, but they are also increasingly focusing on exploring new types of born-digital records such as web archives and social media archives.

However, too often, born-digital and digitized materials are inaccessible due to copyright reasons. For instance, copyrighted texts are not available for download from HathiTrust, a not-for-profit collaborative of libraries preserving 17 + million digitized items (including c. 61% *not* in the public domain). Copyright reasons also largely explain why web archives collected by major libraries (including the British Library and the Bibliothèque Nationale de France) cannot be put online. To consult archival webpages that were once publicly available, users often need to travel to the repositories. Other types of archives born in digital forms, such as emails, Word documents, digital pictures and video files, can also be difficult to access due to copyright – but also privacy concerns, and technical issues. Emails have largely replaced letters, and yet, researchers who need to consult archival emails will very rarely be able to access these born-digital records.

This special issue explores the current challenges and prospects of born-digital and digitized archives for the digital humanities, focusing particularly on the topic of access. It brings together experts from archival science and the humanities, with experts and practitioners from cultural heritage institutions. It is a key research output of the AURA (Archives in the UK/ Republic of Ireland & AI) network, funded by the Arts and Humanities Research Council (AHRC) in the UK and the Irish Research Council.[1] The AURA network was designed to unlock cultural assets that are preserved in digital archives, closed to the public, or difficult to access. By bringing together digital humanists, computer scientists, and stakeholders (including policymakers), the network aimed to find solutions to the problem of inaccessible records in digital archives. To explore the challenge 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 inaccessible archives: we cannot expect archivists or digital humanists alone 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. Until recently, the scholarship on digitized and born-digital records originated from the archive sector and focused primarily on preservation. There were few examples of digital humanists who sat at the same table and took part in these discussions. In 2010, Matthew Kirschenbaum, an American professor of Digital Humanities, co-authored a report on 'Digital Forensics and Born-Digital Content in Cultural Heritage Collections,'

[1] www.aura-network.net.

with professionals in libraries and archives (Kirschenbaum et al. 2010). This report then led to a partnership that developed the BitCurator system[2] now widely used by digital archivists. Ten years later, Ryan Cordell published a report in partnership with the Library of Congress on machine learning (ML) in the library sector. The report mentioned that "access to data is the single greatest practical hurdle to more Machine Learning work in libraries" (Cordell 2020, p 50).

The challenge of access is at the center of the new US/UK network AEOLIAN (AI for Cultural Organizations),[3] which complements the AURA network and its research outcomes. This special issue focuses on problems of access to born-digital and digitized archives in the digital humanities (DH), both from the infrastructural and users' perspectives. But accessing archival collections is not enough. The articles presented here also highlight the need to use innovative AI-based methodologies (such as Natural Language Processing or Linked Data) to support research. It is also essential to develop partnerships between digital humanists, computer scientists, and cultural heritage professionals to fully explore new ways to approach digitized and born-digital archives.

Questions explored in this special issue include: How can we increase use by digital humanities scholars of born-digital and digitized archives? How can we give greater access to collections important to digital humanists that are currently restricted? Collectively, the articles in the special issue problematize the challenges and prospects of digital and born-digital archives. They offer new theoretical interpretations, apply research methodologies to new case studies, and offer innovative perspectives on present and future archival digital collections.

We invited contributions from interdisciplinary voices and received responses from digital humanities scholars, emerging academics, trained librarians, and archivists. While looking at a broad range of issues, from social media use to Python notebooks, Handwritten Text Recognition to the semantic web, these contributions examine similar themes as they seek to make digital archives accessible through a range of new theoretical frameworks and practical tools. The initial contributions to this special issue consider the subject of accessibility and other ethical challenges faced by digital archives. The second section focuses on new computational tools for archives.

Our first article, 'Digital critical archives, copyright, and feminist praxis' by Claire Battershill et al. look at the challenge of creating equal archives. With a special focus on twentieth-century publishing history, Battershill introduces the Modernist Archives Publishing Project (alongside Helena Clarkson, Matthew N. Hannah, Ilya Nokhrin, Elizabeth Willson Gordon, and Nicola Wilson, all from the project team). The article acknowledges that archives of different sizes face different challenges when trying to set up collections. Smaller archives face greater challenges, as their digitization processes depend on what scholarly research gets funding, so unequal practices can trickle-down and impact what is or is not available for users to view. The Modernist Archives Publishing Project has emerged in this context,

[2] https://bitcurator.net.

[3] www.aeolian-network.net.

looking to make the selection process when digitizing objects clear to users. The aim is to "create a digital archive that embodies feminist principles at all levels of practice." The first phase of the project ran from 2012 to 2020 and started by looking at records associated with Hogarth Press, Leonard, and Virginia Woolf's publishing house. The authors note that copyright was a considerable barrier to making this collection available online since much of the material is from the twentieth century. To overcome this challenge, it was possible to gain permission to digitize from the legal owners, but this meant that physically digitizing the archives was just as time-consuming as gaining clearance to do so.

In 'Archives, linked data and the digital humanities: increasing access to digitized and born digital archives via the Semantic Web', Ashleigh Hawkins also considers the huge amount of work that needs to be done behind the scenes to digitize archives, as well as the great value this process can have. The article is an introduction to linked data, looking at the benefits of, and current industry barriers to, archival linked data, and the future directions for DH as data. This includes the possible incorporation of AI, which has already started to take place, for example in the use of the open-source tool ePADD for enabling access to email collections. Hawkins notes that where digital data has been produced, it is not always readily available because of issues such as intellectual property. For this data to be made machine-readable and accessible, archivists need to be involved in the conversation. Hawkins argues that while the infrastructure for archival linked data is emerging, archivists need to participate in the production of these tools so that the metadata, content, and other context are of a high enough standard for the semantic web. Hawkins is one of many in this special issue who calls for an interdisciplinary approach to digital archives.

Furthering the discussion of accessibility and the ethical challenges presented by privacy issues is 'A survey on email visualisation research to address the conflict between privacy and access' by Zoe Bartliff, Yunhyong Kim and Frank Hopfgartner. This paper explores the perpetual cycle between email data access and concerns for privacy and private information held within email archives. Where Bartliff et al. express the need to make emails more accessible data sources, they also address the seemingly impossible balance between access and privacy—a need not met by current visualisations for researching email data. The article proposes a categorisation of email visualisation attributes and a graded scale, as a means to identify the extent to which privacy conscious data management can impact research on email collections.

No matter what the archive is, or how big it is, granting users access in a way that clearly shows how and why some records are available while others are not, is a key challenge for digital archives. Sustainable ways to maintain and create archives need to be identified, and articles in this special issue illuminate several ways this could be done. A cross-disciplinary approach seems almost inevitable given the status of the digital humanities as a tech-focused study of historical concern. Once the issue of access to archives has been overcome, new computational tools can be developed to either help explore these archives or complete the digitization themselves. Two examples are presented in this special issue: Handwritten text recognition technology (Transkribus), and Python Notebooks. Both tools come with their own unique

set of prospects and challenges, showing that while the future of AI in archives may be bright, there is still a need to scrutinize how and why we use digital archives throughout the process.

Joseph Nockels et al. take a close look at Transkribus, a software that has been used to speed up primary source transcription for digitization, in their article 'Understanding the application of handwritten text recognition technology in heritage contexts: a systematic review of Transkribus in published research'. Transkribus was originally funded as an EU Horizon 2020 project which launched as an online tool in 2015 and is now a pay-to-use HTR technology. Nockels conducted a study of all articles (not just peer-reviewed journals) that have used Transkribus between 2015 and 2020. The authors found that Transkribus lends itself to a variety of studies, and this is only getting more eclectic as time goes on (and the article flags a recent branching out into botany). They also note that while the tool originated in academia, its use may go well beyond that, and there is a need to monitor its impact. This can be done with open forums and peer support to assist and continuously remind users of the need for caution when using the tool. As with many digital projects, handwritten text recognition in general and Tranksribus, in particular, have benefitted from Covid-19, which increased a focus on digital projects in cultural heritage institutions. This paper is the first review of Transkribus in published research, although if use of the technology continues to increase at the rate it has in the last five years, it will probably not be the last.

Technology can make digitized sources more accessible, and it can also lead to new knowledge–which is the topic of Leontien Talboom and Mark Bell's 'Keeping it under lock and keywords: exploring new ways to open up the web archives with Notebooks'. This article looks at how Python Notebooks can be used to take users beyond the keyword search, with Notebooks serving as supporting tools that aid search work while giving users a deeper look into their archive sources. The article looks at two Notebooks used to search the UK Government Web Archive (UKGWA). Notebook 1 focuses on available metadata, providing a view of records not available through the UKGWA, while Notebook 2 takes the researcher to the next step–crawling through UKGWA on behalf of the user and extracting relevant content from pages to present an overview of what is available. Notebooks can be used to showcase datasets for researchers, but the authors admit they may not be sustainable long-term as their use depends on the availability of cloud platforms. The authors encourage institutions to think about the ethical issues around computational methods, and again emphasize the need to keep users involved in conversations about the tools used to select and study historical data. Nonetheless, Notebooks are (for now) a good tool to use when accessing archives.

The special issue closes with an article from Lise Jaillant, 'How can we make born-digital and digitised archives more accessible? Identifying obstacles and solutions'. As a literary scholar and digital humanist, Jaillant addresses the key issues faced by researchers when attempting to access "dark" digital collections. Based on a series of interviews with archival and library professionals in the UK, Ireland, and the US, the article explores the common obstacles that often hold back developments for improving the accessibility of digital archives. The article outlines current levels of access to digital collections: some of them are completely closed to users,

while others are accessible on a limited basis (for example, when digital files are available on-site but not remotely). Jaillant suggests possible solutions to the problems of access–including the ethical use of Artificial Intelligence to unlock "dark" archives inaccessible to users. She proposes the creation of a global user community who would participate in decisions on access to digital collections. The articles in this special issue all share a similar focus on the need for interdisciplinary research in archives, the need to include users in conversations with practitioners and scholars when creating new research tools, and the need to stay vigilant regarding possible ethical downfalls in digitization. The landscape of digital archives is changing rapidly, as these articles show in both their scope and their focus on recent and new technologies. What emerges from this special issue is the need for future-proof solutions to the issues of born-digital and digitized archives, and to keep these discussions going well into the future.

Funding This study is part of the AURA project (Archives in the UK/ Republic of Ireland and AI), which received funding from the UK Arts and Humanities Research Council (reference AH/V002341/1) and the Irish Research Council (reference IRC/V002341/1).

References

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 13 Aug. 2021.

Goudarouli E, Sexton A, Sheridan J (2019) The challenge of the digital and the future archive: through the lens of the national archives UK. Phil Tech 32:173–183. https://doi.org/10.1007/s13347-018-0333-3

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 13 Aug. 2021

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.

Katie Aske is a Research Assistant at Loughborough University working on the AURA and AEOLIAN networks. Aske is an award-winning scholar of eighteenth-century literature and the digital humanities. After completing her PhD at Loughborough University in 2015, Aske undertook a Postdoctoral Research Fellowship at Université de Bretagne Occidentale in 2016, working on the digital humanities project 'DIGITENS: Digital Encyclopaedia of Eighteenth-Century British Sociability'. She has published widely on eighteenth-century literature and medicine and has also worked on several major research projects at

Northumbria University, including the AHRC-funded Sterne Digital Library.

Eirini Goudarouli is Head of Digital Research Programmes at The National Archives. Her current research interests focus on digital research in cultural heritage. She is particularly interested in bringing together methods and theories from a range of disciplines that could essentially contribute to the rethinking of digital, archival and collection-based research.

Eirini has extensive experience working on interdisciplinary research projects across the Cultural Heritage and Higher Education sectors. In 2015 she received a doctorate in History of Science from the University of Athens, and she has been a visiting scholar at the University of Cambridge and the University of Helsinki.

Before joining The National Archives, she was a researcher at the University of Warwick and an Associate Research Fellow at Birkbeck, London. In previous years, she spent more than five years working with the University collections belonging to the Historical Archive and the Lab for the Electronic Processing of historical archives at the University of Athens.

Eirini is also a Research Fellow at the Research Centre for the Humanities, Greece, a member of the Digital Committee at the Royal Historical Society, a member of 'Humanities and Data Science' special interest group at the Alan Turing Institute, and a board member of the Advanced Information Collaboratory, an international network with partners from leading academic and cultural institutions spanning five continents.

Natasha Kitcher previously studied History at Royal Holloway, where she completed her BA in History and an MA in Public History. As part of the course she learned a great deal about engaging wider audiences in the past and wrote a play (Mum is MAD!) that was performed at Stanley Halls in 2019.

Now focused on her doctoral research, Natasha was recently the Programme Editor and an Online Tutor for the University of London Worldwide. She is currently the Rapporteur for an AHRC funded project with the Science Museum considering the culture and display of space exploration for future space galleries.