Are Users of Digital Archives Ready for the AI Era? Obstacles to the Application of Computational Research Methods and New Opportunities

LISE JAILLANT and KATHERINE ASKE, Loughborough University, United Kingdom

Innovative technologies are improving the accessibility, preservation, and searchability of born-digital and digitised records. In particular, Artificial Intelligence is opening new opportunities for archivists and researchers. However, the experience of scholars (particularly humanities scholars) and other users remain understudied. This article asks *how* and *why* researchers and general users are, or are not, using computational methods. This research is informed by an open-call survey, completed by 22 individuals, and semi-structured interviews with 33 professionals, including archivists, librarians, digital humanists, literary scholars, historians, and computer scientists. Drawing on these results, this article offers an analysis of user experiences of computational research methods applied to digitised and born-digital archives. With a focus on humanities and social science researchers, this article also discusses users who resist this kind of research, perhaps because they lack the skills necessary to engage with these materials at scale, or because they prefer to use more traditional methods, such as close reading and historical analysis. Here, we explore the uses of computational and more "traditional" research methodologies applied to digital records. We also make a series of recommendations to elevate users' computational skills but also to improve the digital infrastructure to make archives more accessible and usable.

CCS Concepts: • **Applied computing** → *Arts and humanities*; *Digital libraries and archives*;

Additional Key Words and Phrases: Digital humanities, archives, born-digital records, digitised records, user experience, computational methods, artificial intelligence

ACM Reference Format:
Lise Jaillant and Katherine Aske. 2024. Are Users of Digital Archives Ready for the AI Era? Obstacles to the Application of Computational Research Methods and New Opportunities. *ACM J. Comput. Cult. Herit.* 16, 4, Article 87 (January 2024), 16 pages. https://doi.org/10.1145/3631125

1 INTRODUCTION

Humanities and social science research in the digital age is facing new challenges. On the one hand, digitised and born-digital materials [1] have opened-up remarkable avenues for research and innovative methodologies (such as data mining, natural language processing, and working with data at scale). On the other hand, technologies are changing more traditional research methods, such as close reading, and altering the relationship between the user and the archive. The digital humanities are often considered an additional strand of the broader humanities disciplines (such as English or History), and yet, the digital age has become central to almost all scholarly research. Digital collections such as web archives have been increasingly recognised as an essential source for studying cultural and social history in recent decades [2]. But unlike the computer sciences, humanities and

Authors' address: L. Jaillant, and K. Aske, Loughborough University, Epinal Way; Loughborough LE11 3TU, United Kingdom.

ACM 1556-4673/2024/01-ART87
https://doi.org/10.1145/3631125

social science researchers tend to have a more individually tailored approach to researching digital materials, often focussing on the historical or literary analysis of individual texts. In this article, we explore user experience and the methodological approaches applied to digitised and born-digital records.

In acknowledgment of continually changing user needs, user-focussed studies have become a significant part of digital humanities and archival practices research––informing critical approaches, methodologies, and the real-world curating of digital collections and their management [3]. The end user is often referred to in discussions regarding digitisation, searchability, and accessibility of digitised and born-digital materials. But the "end-user" is a term that broadly spans a range of skills, disciplines, and knowledge.

This study, as part of the transatlantic AEOLIAN Network (Artificial Intelligence for Cultural Organisations) [4], asks *how* and *why* researchers and general users are, or are not, using computational approaches to analyse digital collections. Computational research methods include a wide range of approaches, including text and data mining, data visualisation, digital data analysis, and using such methods at scale. The applications of **Artificial Intelligence (AI)**, and its subsets, machine learning [5], natural language processing, **Optical Character Recognition (OCR)**, and handwritten text recognition have also widened the scope of computational research methods. We consider the knowledge and skillsets required to conduct these types of data-driven research, and whether this is a roadblock for humanities and social science researchers who have not necessarily been trained or encouraged to use such methods, or else prefer more individually tailored approaches when using digital and digitised resources.

Are some users being left behind simply by a lack of digital skills? Are their own research needs not being met? Where computational research methods are not used, we question whether it is an issue of training to engage with such methods, simply a preference for more "traditional" forms of archival research (such as close-reading or historical analysis), or even if it is tied to career incentives. For example, if computational research is not well understood or valued by hiring committees within the humanities, then the support or incentives for such research will not be available [6].

This article draws on responses to an open-call survey, completed by 22 individuals, and semi-structured interviews with 33 professionals, including archivists, **GLAM (Galleries, Libraries, Archives, and Museums)** sector professionals, digital humanists, literary scholars, historians, and computer scientists [7]. It should be noted that those who participated in the survey and interviews were mostly based in Europe and North America: our findings are therefore located in the Global North, and may not be applicable to other contexts.

The first section looks at the experience of the end users of digitised and born-digital archives. Drawing on responses to our survey, we show that the first problem to solve is the issue of access to these collections. Without access or with limited access, it is difficult to scale up training in computational methods. In other words, new technologies such as AI requires access to data first.

The second section builds on in-depth interviews to shed light on other obstacles within academia. This includes criticism of computational approaches that neglect critical thinking. Other obstacles include the lack of formal training in computational research, and the model of the "solo researcher," traditional in the humanities and qualitative social sciences. Our interviewees pointed out the importance of collaborations with computer science and other disciplines to analyse huge amounts of data in archives. They gave us examples of collaborative work on AI applied to archives, illustrating new models of research beyond traditional solo work.

The third and final section makes recommendations to overcome the obstacles to applying computational methods to digital archives.

- We suggest that computational training should become embedded in all postgraduate programmes in the humanities and qualitative social sciences. More advanced training should also be readily available to scholars who want to elevate their computational skillset.
- We also recognise that it is neither possible nor desirable for everyone to become an expert in AI and other computational methods. We therefore suggest that cross-disciplinary collaborations should be more

valued (in terms of promotion and other career opportunities), to promote positive alternatives to the model of the solo researcher.
- We also need funding agencies to promote the development of infrastructures that will allow greater engagement with digital resources.

These recommendations could help users of digital archives take full advantage of AI and computational tools, without losing sight of risks (including the risks of biased results based on flawed data).

(1) No Use without Access: Understanding Users of Digitised and Born-Digital Archives

In many recent studies into archival and research practices concerning born-digital and digitised resources, the "end-user" is regularly presented as a means to end––to help develop innovative technologies and anticipate user issues in their development [8]. Among these innovative technologies are AI and machine learning, which have a wide range of applications in the archival sector. AI can be used to automatically create data, or to identify sensitive records and thus make possible the release of non-sensitive materials. However, while machine learning tools are being increasingly applied within the cultural heritage sector, AI is still largely in the experimental stages within archival practice [9]. This means there is a timely opportunity to evaluate user requirements within a digital archival context. With the development of AI applications to archives, are users' expectations being met? Are computational research methods being enabled, and are they being tailored to researchers across disciplines?

Academic libraries are progressively providing services to allow users to explore digital collections in inventive ways and at scale [10]. Yet Mary Burke et al. have highlighted that there is an ever-present necessity to re-evaluate end-users as their needs continue to change [11]. Not only is there no such thing as a typical "end user," these multiple users also have changing needs that archival institutions need to address. Are we at risk of unduly promoting specific types of computational research, while qualitative approaches and methodologies are left behind [12]? Maemura et al. have noted that traditional approaches to archival research, including close reading and historical analysis, depend on the archival processes that preserve historically significant works through appraisal [13].

The scale of web archives and other born-digital databases confounds these traditional appraisal processes, meaning these digital resources "are replete with information that may not be significant to the research questions being asked" [14]. As Bell et al. point out in their discussion of the UK's Government Web Archive, studies into the use of web archives as primary sources for humanities research remain concerned with problems of working at scale, discussing issues of their complexity and inaccessibility, and the "unsuitability of keyword searching" as a primary form of exploration [15]. Indeed, Bell, who works as Senior Digital Researcher at The National Archives UK told us in an interview: "I've been looking at different ways of applying machine learning to make sense of the web archive. And not necessarily just machine learning. I'm doing some more traditional text analysis or information retrieval techniques as well, … to make it more accessible or easier to understand" [16]. Web archives, which provide vast amounts of important cultural and socially significant data, are, for the most part, inaccessible without advanced computational tools.

Even users who have the skills to do computational research might not be able to access data in the first place. In the case of the UK web archive, 19,000+ websites have permission from their owners to be viewed from anywhere online. However, many other collections are not accessible remotely, and users need to travel to the British Library or other legal deposit libraries to view these websites. Getting access to other kinds of born-digital collections (such as email collections) can also be extremely complicated for several reasons including copyright and data protection [17].

Limitations in usability impact born-digital collections but also digitised materials. Technologies such as OCR are used to make digital texts searchable and useable, but they can misrepresent the original text. This is a particular concern for older, type-set texts, such as those available on large historical databases, including Early

English Books Online; Eighteenth-Century Collections Online; or Nineteenth Century Collections Online. In an article entitled "'Q i-jtb the Raven': Taking Dirty OCR Seriously," Ryan Cordell points out that "we must avoid the myth of surrogacy proffered by page images and instead consider directly the text files they overlay" [18]. Going back to the (physical) reading room is all the more important as databases often include only one copy of any given edition. Addressing the challenges of "dirty OCR" has been central to the collaboration between Gale and Text Creation Partnership, which has led to the development of **Text Encoding Initiative (TEI)** and the improvement of the transcriptions on ECCO [19]. The advancement of complex digital-data infrastructures is not always paralleled by the development of complex computational research methods. The facilities for actually using digital and digitised records remain underdeveloped, and often users wishing to apply more traditional methodologies are limited to keyword searches, Boolean operators (AND, OR, NOT, AND NOT, etc.), and predetermined categories to conduct their research.

While working on a close analysis of a digitised text, a researcher may be able to adapt their approach to counter the potential mistakes within a digital copy, as Anna Kuslits (doctoral researcher in the History of Science at the University of Edinburgh) suggested in our interview. But when searching across databases, inaccuracies in descriptions or transcriptions can mean "those sources would be invisible, and we would not use them or know that they actually exist" [20]. In this way, the apparent ease of digitally accessible records, and the potentially unsuitable ways of using them, are impacting what records are studied, and how [21]. As Richard Dunley and Jo Pugh have suggested, "In reality it can be very difficult [to] know which areas of the ocean are well charted, and which are almost bare. This leads us to the question of how representative are our representations" [22].

Indeed, there remains a gap between "usage" of digital collections and *user* data [23]. How then do we begin to measure what is *not* being used, and why? Do humanities and social science researchers have the practical understanding and access to knowledge required to make full use of digital collections for their research? And, if not, then how can we ensure that the methodological approaches that are being applied, are acknowledged within archival practices? Whether users of digitised and/or born-digital archives are working at scale or conducting close analysis, understanding the potentials and the limitations of such records, and how to address them, is a necessity for all users.

This article focuses not on user participation, but on user experience: what is the user doing or not doing with these materials? The study examines whether technological improvements to digital databases and advancing archival practices might isolate users who do not engage with computational research methods, which are rarely taught within humanities and social science disciplines. We investigate why users resist this kind of research––perhaps because they lack the skills necessary to engage with digital archives computationally, or because they prefer to use more traditional methods, such as close analysis. Qualitative research methods remain central for many researchers, and yet, their needs are often peripheral to technological processes applied to archives. Technological innovation should enable, not direct, the ways research is conducted. In a context of increasing pressure to make data (including data in digital archives) more accessible, we step back to question just how this "accessibility" might impact the future of humanities and social science research. Here, we propose that qualitive data on user experience and research methodologies across many disciplines should inform decisions made on archival processing and digital record-management. What would digital collections such as the UK government's web archive look like if more attention was paid to the needs of users who do not know how to use computational methods? And what kind of training can be offered to help users who wish to develop their computational skills?

Our survey "on the use and non-use of computational research methods on digitised and/or born-digital materials" was open in June 2022 and circulated via listservs mostly used by Digital Humanities scholars and GLAM sector professionals, reaching an audience in Europe, North America, and elsewhere. It should be noted that listservs such as the Humanist discussion group gather diverse groups (including scholars who do not use computational methods but who are interested in digital transformations). The survey remained open for three weeks, and we received a limited number of responses (22). The small sample was completed by interviews, as explained later.

Most survey respondents (59%) were academics at an early stage in their careers (either MA or Ph.D. students or early career researchers within 6 years). The rest of the group was composed of mid-career and senior academics (34%), GLAM professionals (18%), or other civil servants/policy makers (9%). Eighty-nine percent of respondents said they used digitised or born-digital records as part of their profession. In a multiple-choice question, 82% said this research was undertaken for publication or written outputs, with the same number for talks or presentations, academic or otherwise. Fifty percent of the participants said they also conducted personal research, and funded research projects. When asked about the frequency of their research using digitised or born-digital records, 64% said they undertook such research on a daily basis (working week), with 27% saying weekly basis.

The group of respondents was familiar with the opportunities and challenges of digital archives around the world. When asked what research facilities were used, online or in person, there was a good balance of digital resources and physical libraries from across the globe, including Gale's *ECCO*; HathiTrust; Internet Archive; Google Books; and Pacific and Regional Archive for Digital Sources in Endangered Cultures (PARADISEC); as well as national libraries and museums, including The National Archives UK; National Library of Scotland; The British Library; National Library of Brazil; New York Public Library; BNF (French National Library); Huntington; Cleveland Museum of Art; Smithsonian, Victoria & Albert Museum and others. Asking participants what issues they faced when using digitised and/or born-digital records, "limited availability of digital records" was the highest scorer with 86%, with some highlighting specific issues with OCR, copyright, and inaccurate metadata as their biggest concerns. The next biggest issue for 68% of those surveyed was the "discoverability of digital materials," closely followed by the "availability of data to apply computational research methods to" and "online accessibility" (both 64%); and then "search tools" (59%). The "lack of technical knowledge or skills" was an issue for over a quarter of those surveyed (27%), and a "lack of guidance or instructions for users" was an issue for four of those surveyed.

We asked participants whether or not they applied computational research methods (such as text or data mining, machine learning, topic modelling, and the like) to digitised and/or born-digital archives at scale. Seventy-seven percent said yes, with the majority (59%) suggesting they were self-trained in computational research methodologies; two respondents described themselves as formally trained; and two did not specify. However, when asked if they felt they had the "necessary knowledge and/or skills to apply computational research methods (for data at scale) effectively," only seven participants (32%) said yes. We then asked participants if they preferred more "traditional" research methodologies, which we defined as close reading; historical or literary analyses; or consulting physical archival records. Most participants answered no (64%) or suggested a preference for both computational and traditional research methodologies. Only one participant said they preferred a traditional approach. Of the 20 participants who responded to the question if certain digital archives or resources were difficult to use or inaccessible for applying computational research methodologies, 50% said yes, and they listed issues such as copyright and data protection that limit access to data, issues with downloading data, the lack of availability of open-source tools/software, and broader issues in knowing what can and cannot be used in a research capacity.

More than half of respondents answered yes to the question, "Do you think your research would benefit from computational research methods that could be applied to data at scale?" Following this question, we asked if participants would therefore gain from "tutorials, guidance, or training with specific tools (such as data wrangling, software carpentry, text and data mining, data visualisation, managing digital and digitalised assets, digital data analysis) to learn how to apply computational methods to digitised and/or born-digital archives more effectively." Most said yes (64%), and the remaining participants suggested that they had either used or were using these tools; or that they felt they would like to use such tools if they were more tailored to their discipline, particularly in the humanities.

Several participants left further comments on the inadequate training in humanities disciplines. One told us that "technology should not dictate the direction research takes in the humanities"; and another suggested that

"we might need either more collaboration across disciplines to bridge these gaps," or else find out "to what degree it is necessary to also teach humanities students these kinds of skills in today's age."

Drawing on the survey results, we formed a series of interview questions to discover more about the potential issues of training and accessibility within digitised and born-digital archives. These interviews with researchers and GLAM sector professionals shed light on computational research methodologies that are being used or facilitated within digital archival collections. They also illuminate the obstacles that users face when encountering huge amounts of records without having the tools to analyse these data at scale.

2 LACK OF COMPUTATIONAL TRAINING AND SOLO RESEARCHER MODEL: NAVIGATING OBSTACLES WITHIN ACADEMIA

In June and July 2022, we conducted a series of 24 interviews with 33 participants on the application of computational research methods. The interviewees were made up of researchers who use digitised and/or born-digital records, and archival professionals who work with these types of records. We used our own networks of contacts to select interviewees, paying attention to geographical reach (Europe and North America), gender, and career stage. Getting a comprehensive overview of the global academic and GLAM communities was well beyond the scope of this study, and we focused instead on gathering data from well-informed stakeholders. Each interview lasted between 30 min and 1 h, and was conducted via MS Teams before being transcribed. Here, we discuss the findings from these interviews and consider the key themes that emerged:

- research methods (both "traditional" and "computational")
- accessibility
- issues of bias, representation, and transparency
- skills and training
- ways of doing research (solo researcher versus teamwork)
- using computational methods in the AI age

2.1 Research Methods (Both "Traditional" and "Computational")

Adopting computational and more generally quantitative methodologies remain problematic for many humanities scholars. More than ten years ago, James English wrote "Academic disciplines (and even interdisciplines or hybrids) are relational entities; they must define themselves by what they are not." Taking the example of literary studies, he added that the humanities are not "counting" disciplines [24]. This largely explains why digital humanities––a field that values quantitative and computational approaches––has attracted severe criticisms. In 2016, Daniel Allington, Sarah Brouillette, and David Golumbia published an influential (and controversial) article in the *Los Angeles Review of Books* denouncing the "fetishizing of code and data and the relative neglect of critical discourse within Digital Humanities" [25]. **Digital Humanities (DH)** was presented as a neoliberal tool aimed at destroying traditional modes of humanistic inquiry, which value critical analysis and political engagement. By valuing quantitative approaches, DH scholars were complicit in the neoliberal attack on the humanities. For the authors of the article, "Digital Humanities as social and institutional movement is a reactionary force in literary studies, pushing the discipline toward post-interpretative, non-suspicious, technocratic, conservative, managerial, lab-based practice" [26].

Reflecting this hostility or at least ambivalence toward quantitative approaches, some of our interviewees expressed concerns for implementing computational methodologies into their research. Anna Kuslits told us that "a lot of people who work in the humanities aren't taking on these methods for good reasons, because these methods don't work with the kind of understanding that we have about meaning-making and how complex meaning-making is in human culture." For Kuslits, the computational approach was not necessarily appropriate for the kind of critical inquiries favoured by many humanities researchers. Yet, funding pressures often compel them to attempt to adopt these tools, or at least to try to integrate digital elements in their research. "The shrinking

budget that we have for humanities research is extremely skewed in the direction of digital research," noted Kuslits. Funders expect researchers to be "competitive and efficient" and adopt an "entrepreneurial mindset." Kuslits added, "I find it is very aggressively pushing humanities and qualitative social science research [toward] a position that doesn't really understand what this research is really about" [27].

Kuslits echoed arguments that have shaken the digital humanities in the past decade. Allington, Brouillette, and Golumbia have deplored the funders' obsession with DH to the detriment of other humanistic methodologies and critical frameworks. "It is difficult to get six-figure grants for English scholarship without engaging in computational work," they argued [28]. These funding pressures were presented as a conservative attack against the humanities, a discipline often associated with the Left in the North American context. Responding to this line of criticisms, the discipline of DH has turned in the past few years toward more overtly critical approaches, often influenced by gender and race studies. For example, a recent project funded by the AHRC (Arts and Humanities Research Council) states: "Digital Humanities has a problem. It is built from inherited heteronormative, gendered, and frequently racist brick and mortar" [29]. This project, *Full Stack Feminism in Digital Humanities*, aims to address this central problem by engaging with intersectional feminist theory. In short, this project and others bridge the gap between literary criticism and digital humanities, a discipline often presented as a-theoretical.

While many humanities researchers we interviewed perceived a "gap" in their ability to use computational research methods, others combined the traditional forms of document analysis with computational methods. For example, Graham Jevon (Endangered Archives Programme, British Library) explained that his doctoral research in history was based on more traditional research methods, but since joining the BL, he has had the "chance to get a postgraduate degree in data science." This has allowed him to undertake data processing using Python and machine learning methods for a new crowdsourcing project. "I think the computational side of things [benefits] from that traditional close reading, so that you can have that real understanding of what the documents are saying" [30]. Abby Gondek (Florida International University) told us she "had experience with some qualitative data analysis software like N Vivo" but "had never used machine learning or AI technologies" before collaborating with computer scientists and archivists on a larger project with the **Franklin Delano Roosevelt (FDR)** Presidential Library [31]. Collaborative work between disciplines enabled this research team to combine different methodologies, leading to new findings.

2.2 Accessing and Using Data for Computational Research

Applying computational methods requires accessible and usable data. But getting to the stage where the data is usable requires an entire workflow, relying on the labour of archivists as well as automated processes. The stage of pre-processing digital data is often hidden from view. Dispelling the myth that data comes ready to be used by researchers, archivist Aurélia Rostaing (Archives Nationales, France) said that "data always needs to be worked out. You never get data that you can use right away, that doesn't exist" [32]. Similarly, Professor of Data Science Frank Hopfgartner (Universität Koblenz-Landau) stated that "digitised material" is not "immediately usable . . . you have to do a lot of processing beforehand" [33]. The expectation that digitised and born-digital materials should come ready to use for any number of research methodologies is putting unprecedented demand on archivists. As Jenny Bunn (The National Archives UK) told us, "If my job as an archivist is to maintain the data in a useable form and what is a useable form requires more and more, and more pre-processing, where do we draw the line?" [34].

Even fully processed digital collections are not always discoverable to users, for example, due to issues with metadata. And not all collections have been made discoverable online (either via full digitisation or via an online catalogue). Librarian Treasa Harkin (Irish Traditional Music Archive) told us that not all of their collections have been digitised, and that "somebody who's not finding what they want to find" online, could easily interpret their collections as non-representative [35]. As in the case of paper archives, digital collections are never complete. Data is missing, and digital archives are not the exact reflection of reality. Ryan Cordell (University of Illinois Urbana-Champaign) gave the example of *Chronicling America*, a large-scale database of U.S. newspapers of the

nineteenth century. States had to put forth a plan for their digitisation efforts, and they often chose geographic spread as the driving force behind their programmes. For example, when the state of Pennsylvania decided what they wanted to digitise, they picked one newspaper from each of the ten largest metropolitan areas in Pennsylvania in 1870. "What it means is that they only digitised one newspaper from the city of Philadelphia, which was a publishing powerhouse in the 19th Century," Cordell noted. He added that *Chronicling America* "dramatically under-samples urban centres. It really is skewed toward smaller cities, more rural cities" [36]. This can lead researchers to overrepresent the importance of small cities and to downplay the centrality of Philadelphia in 19th-century print culture. This is all the more problematic when finding information about the digitisation priorities is a complicated process. When Cordell started working on *Chronicling America*, there was no public information about the choice to favour geographic spread above other criteria.

2.3 Issues of Bias, Representation, and Transparency

Greater accessibility and discoverability are often presented as desirable, but they also come with risks that records will be taken out of context, and for cultural and racial biases to be further perpetuated. Often in the digitisation process, the generated item metadata will continue to reflect the outdated structures inherent to most archival institutions in the UK, and beyond [37]. For example, many archives created during the colonial era contain racist or otherwise problematic language. This in turn poses issues of discoverability, as outdated language would not necessarily be included by users in keyword searches. These issues were an important topic for many participants, whether researchers or practitioners within the archival sector, but the responses showed active approaches and possible solutions.

There is a growing perception that mass digitisation is potentially increasing the risk of bias in archives. Archivists are aware that they have an important role to play to mitigate risks. Jenny Bunn (The National Archives UK) told us, "In the archive profession, we are getting much more aware about our own kind of agency" [38]. This growing awareness has led to many positive steps being taken to tackle the issue of transparency head on. For example, historian and archivist Kai Naumann (Landesarchiv Baden-Württemberg) told us that actually, "artificial intelligence might empower us . . . in bringing automated indexing, in bringing automated transcription" to improve accessibility [39].

The issues of bias and transparency can in part be addressed through more critical uses of computational approaches. Tobias Hodel (Assistant Professor in Digital Humanities, Universität Bern, Switzerland) thus suggested that machine learning can be used critically, rather than being treated as a black box [40]. As Anna-Maria Sichani (Congruence Engine project, School of Advanced Study, London) commented, while bias is "embedded in the design and development of the systems and the tools that we are using," and these "are often reinforcing existing historical patterns of discrimination," it is also "becoming more obvious and more visible *because* of the tools that we are using" [41]. Indeed, computational methods also present an opportunity to correct these issues, and one interviewee said that they thought bias was actually being "reduced by digital resources and the availability of information" as research was once limited "to what you had on your library shelf" [42].

To further raise awareness of bias and potential misrepresentation of digital archives, the first big step is transparency––being upfront with the user about what has been done, what is missing, and what is still being addressed. As Sara Thompson (Digital Archivist, University of Edinburgh) told us, "I think the better we can document things the more transparent we can be, the more we have conversations with the communities that we serve and are represented in our archives, the more transparent we can be" [43]. But it is not entirely on the archives to provide this transparency, as there are also steps that researchers can take.

Although humanities methodologies tend to focus on the final output, rather than how they got there, there is an increasing demand for more accuracy when referencing the kind of research done using digital resources [44]. If users expect archives to be more transparent, then they need to show the same transparency in their use of those same records. As Aurélia Rostaing told us,

> Researchers more than ever should stick to this basic research principle, which is documenting what you are doing and enabling anyone to check it. Like quoting the reference in footnotes, but then you need to quote which tool you have been using and which version of it, and which version of the algorithm. And you should be able to enable anyone to replay the programme on the same corpus of data [45].

While Maemura et al. argued that "computational reproducibility of results is not a goal for the humanities," we suggest the opposite: that reproducibility should be a core part of humanities methodologies. Yet, the computational approaches being applied often remain invisible, as this is the part of the process that does not always get published. For example, when literary or historical research is conducted using keyword searches on digital databases, tracing research could be adopted within the humanities: Which keywords were used? Which databases were consulted? What search limitations were applied? How was the data sourced? As humanities and social science researchers are now regularly using digital resources as part of their research, there is an ever growing need to document their specific scholarly practices. This would not only improve the transparency of such methods but also allow archives to better understand users of digitised and born-digital archival collections.

2.4 Skills and Training

Independently of their career stage, most of our interviewees felt compelled to adopt computational methodologies (either driven by an inner conviction that this methodology was right for their research, or by external pressures, such as funding). However, a common concern was the lack of computational training available to researchers in the humanities and qualitative social science. Sara Thompson (University of Edinburgh) noted that "a lot of the tools developed to work with data, to do data science, to do data analysis requires programming skills which just aren't really taught in humanity subjects" [46]. Likewise, Lawrence Evalyn (Visiting Assistant Professor of English at Northeastern University) said: "I taught myself to code in Python using a textbook, just on my own. And everything else… is really self-taught" [47]. Similarly, Tobias Hodel (Universität Bern) told us: "I'm mostly self-taught" [48]. Hannah Ringler (Assistant Teaching Professor of Humanities, Illinois Institute of Technology) shared similar comments about the lack of training available to humanities researchers. Training in computational research "wasn't a big part of my Ph.D.," Ringler said, but she took a class called "coding for humanists," which gave a basic introduction to Python, and to natural language processing [49]. While various tools are proving promising at writing code based on human prompts, learning the basics of coding allows researchers to remain independent of external providers and software.

Training (especially advanced training) exists, but it is seldom mandatory and embedded in curricula. Moreover, the time constraints that apply to all research projects make it difficult to acquire computational skills. These obstacles were discussed by Alice Austin (University of Edinburgh), who has recently finished a Ph.D. project funded by the AHRC. Her thesis examined the archive of the 2014 Scottish Independent Referendum, curated by the **National Library of Scotland (NLS)**. For her research, she looked at what the NLS collected from the web, and she investigated how these born-digital records were integrated with analogue elements of the collection. 'My experiences have been very much using traditional methods in a kind of untraditional setting', Austin said. She reflected on the reason why she did not rely more on computational methodologies. A Ph.D. in the UK is generally completed in three years, and it is difficult to spend an extensive period learning computational methodologies. In other words, spending two years learning Python, and one year writing the dissertation is not feasible. Austin added, "I think there's definitely a bit of a gap between the people who want to do this kind of work and the skills that they need to do it" [50].

Learning new computational skills is not enough: one also needs to regularly update their skills sets to keep up with a rapidly developing field. This is difficult for humanities scholars and qualitative social scientists who might not have easy access to this knowledge, or the time to put it into practice. Being aware of these limits is

important, especially for those who teach computational methods in the humanities. As Tobias Hodel told us, "What's crucial in teaching humanities scholars, or students in humanities, is giving them real life examples that they then can work on for their own thesis or their own credit programmes" [51]. Providing humanities researchers with at least an understanding of what approaches can be taken to use digital data computationally, even if datasets have to be adapted and such skills remain beyond an individual researcher's capacities, was one solution to this issue.

The lack of training opportunities and time constraints were not the only obstacle faced by humanities scholars eager to learn new computational skills. Among humanities scholars, there is frequently a sense of exclusion from quantitative approaches and scholars who use them. Commenting on this perception, David De Roure (Professor of e-Research in the Engineering Science Department at the University of Oxford), told us, "Sometimes I think digital humanities can be a club of people who are self-defined digital humanists. And the people who stand to gain most are actually all the other humanists who could be using digital methods, who may or may not be deterred by the existence of a club in a castle with a wall around it" [52]. Although Digital Humanities often presents itself as a friendly and egalitarian discipline, it can be perceived as a closed group, eager to position itself *against* "traditional" humanities.

2.5 Ways of Doing Research (Solo Researcher Versus Teamwork)

The lack of computational skills is worsened by the model of the solo researcher traditional in the humanities. Lawrence Evalyn (Northeastern University) said: "I think one of the awkward things about the humanities is that many humanities fields are very solo-oriented" and there is an "expectation that they should learn all of the skills that are necessary to complete the project that they have in mind" [53]. Some researchers feel it would be necessary to learn a whole other discipline to conduct computational research. This is in part due to the monograph model, which is often single authored in the humanities. Collaborative work is the exception rather than the norm––a fact that Evalyn found frustrating:

> My dissertation had to be solo work. All of the code in my dissertation had to be solo and computer scientists had written the machine learning that we were using. And so, I had to re-think, "What's a project I can execute fully myself?" So, one of my other answers is, going forward, I hope to be able to forge more collaborations so that I don't have to be the one bringing all of the technical skill to the table [54].

Other interviewees told us that they have moved away from the "solo" model, and that collaborations with colleagues from other disciplines enabled them to acquire new skills. These interviewees had started to do collaborative work after their Ph.D., at a time in their career when solo work mattered less than during their doctoral studies. Katherine McDonough (Alan Turing Institute), who finished her Ph.D. in 2013, told us that "not getting a permanent job right away" and working as a postdoc in research teams helped her develop knowledge and collaborations: "I was encountering very different research goals and digital infrastructures and sources and questions, it really kind of, I would say, helpfully forced me to learn new things." While acknowledging that moving from one postdoctoral position to the other was difficult "personally and professionally," McDonough came to realise "it was quite lucky, because it put me in a position––by the time that I got to the Turing––to have a sort of unique experience of engaging with people across many disciplines" [55].

In this narration, being employed on temporary postdoctoral positions and having to move regularly to other cities and even to other countries is presented as an opportunity to learn new technical skills and ways of working. In their 2016 article, Allington et al. denounced "the rebranding of insecure campus employment as an empowering 'alt-ac' career choice" [56]. In other words, presenting postdoctoral positions as desirable opportunities legitimised precarious employment contracts that have become the norm in neoliberal universities. What bothered Allington et al. was that the model of the sciences (where postdoctoral roles are the norm) had expanded to

the humanities. Their laments on the rarefication of stable academic employment echoed Karl Marx and Friedrich Engels's remarks, that the "constant revolutionizing of production, uninterrupted disturbance of all social conditions, everlasting uncertainty and agitation" are central to the capitalist system [57]. According to this line of interpretation, the discipline of DH served these capitalist interests by disrupting previous academic norms.

Interviewees often presented collaboration as one aspect of the learning process, to be complemented by self-learning or formal learning. For Ryan Cordell (Associate Professor in the Departments of Information Science and English at the University of Illinois Urbana-Champaign), the development of skills through collaborations was accompanied by more formal learning––at the Digital Humanities Summer Institute in Victoria (Canada), for example. Likewise, another interviewee mentioned that they enjoyed learning and applying methods, 'but the ideal is to collaborate with people who are real specialists'. Working with specialists from other fields (such as data scientists and computer scientists) was not always easy. For this interviewee, "communication becomes really key and ensuring there's a common level of esteem, and that everyone's getting out what they want from it in a way that's not too compromised by the collaboration" [58]. In other words, cross-disciplinary collaboration was not self-evident and had to be perfected with experience over time.

Transdisciplinary collaboration was presented as particularly useful when it led to a win-win situation, that benefited humanities scholars as well as computer scientists. Commenting on her experience with the *Living with Machines* project, Katherine McDonough (Alan Turing Institute) said, "A lot of my focus has been on trying to work with computer scientists in a way that develops tools that are meaningful for humanists" [59]. Likewise, Cordell highlighted the benefits of collaborative approaches to digital sources, saying,

> The collaboration between the humanities and computer science allows us to ask, "What can we learn using the data analysis methods? And are there ways in which we can do data analyses that actually help illuminate what's missing?" I think that those moments of bringing in the close reading are really pretty essential if you want this big data work to speak to a wide spectrum of researchers [60].

In this way, the combination of humanities and computational research methods, balancing larger-scale data analysis with close reading, shed new light on digital data.

Training and collaborations with computational disciplines enhance the knowledge and awareness of these approaches, but not always the regular practising of computational methods by researchers in the GLAM sector and in academia. Mark Bell (The National Archives UK) said,

> You can do a training course, you can get the basics, but you have to work at it every day to develop any kind of mastery in it. And people don't have the time. Even if they get some aptitude for it, that's not their day job, but certainly being able to talk to technologists and understand what they're talking about and develop that [conceptual understanding] I think is essential [61].

Like Bell, other interviewees pointed out the importance of meaningful discussions between disciplines to achieve common goals. As Alice Austin (University of Edinburgh) told us, crossing the boundaries between the humanities and computer science is becoming essential: "As curators and archivists we're going to have to be able to have conversations that can address both sides of that coin" [62]. Discussing the need for collaborative approaches to using "real-world digital assets, digital resources, digital repositories" at scale, Jason R. Baron (Professor of the Practice in the University of Maryland's College of Information Studies) noted that it is "incumbent on archivists and records managers, for active records that are in huge numbers, and for others––journalists, historians, any number of communities––to be aware of the possibility of using machine learning techniques. And to push for those techniques to be used by archival institutions and by others" [63]. An awareness of computational research methods, even without undertaking any specific training, would allow humanities researchers and GLAM professionals to understand what potential answers these approaches could bring to their own research questions.

2.6 Using Computational Methods in the AI Era

While many scholars in the humanities and social sciences do not engage with computational methods, these approaches are increasingly needed to make sense of huge datasets. Jason R. Baron gave the example of the U.S. President records: "The White House has transferred petabytes to NARA [National Archives and Records Administration] at the end of each administration since Clinton." The rapid development of vast quantities of digital records was not accompanied by a rise in the number of archivists. This is impacting the ability of archival institutions to perform their public duties (including responding to Freedom of Information queries). "The Clinton Library has 13 archivists or records people, and they can't possibly handle a Freedom of Information Act queue of 10 million documents in any reasonable time––while also doing a reference function in trying to open other documents that are not part of the queue" [64].

Vast digital collections cannot be searched manually, or even using basic computational approaches. When government archivists try to address queries, they often type keywords to search vast quantities of records. This method can lead to tens of thousands of results, without any specific order. In contrast, machine learning methods are generally more effective, since they produce a rank-ordered list of records related to the search. AI can also identify specific elements of the record, making possible the release of sections rather than the entire document. In the case of US government records, it is possible to separate the factual part of the document from the recommendations or opinions. This is particularly useful for Freedom of Information requests: The factual section can be shared, and the recommendations can be withheld at the discretion of the agency. For Baron, it is essential for archives to adapt their thinking and methods in an age of big data and AI.

AI can also identify patterns in a mass of data, for example in digital reproductions of art works or in textual records. Todd Dobbs (Doctoral Researcher in art authentication at the University of North Carolina at Charlotte) used machine learning algorithms to determine an artist's probability to have created any given painting. His research seeks to maximise accuracy by taking into account a large number of artists and art reproductions. The difficulty to access copyrighted materials has led Dobbs to focus mainly on public domain images available via open access sources such as WikiArt. Likewise, Tobias Hodel (Universität Bern) is applying AI methods to public domain records––including nineteenth-century archives created by the City of Basle to document the evolution of buildings since the Middle Ages. Hodel uses text recognition to analyse these records and create the economic history of Basle based on the evolution of these buildings.

Our interviewees frequently combined computational methods, including AI-driven tools, with human-centred approaches. Crowdsourcing platforms such as Zooniverse make it easy to create a project, upload data and enlist the help of volunteers. Graham Jevon (British Library) thus relied on crowdsourcing for his project on slavery in Barbados. Drawing on Barbadian newspapers from the late 18th and early 19th Centuries, the project aimed to identify specific sections such as "wanted" advertisements (enslavers looking for people to enslave) as well as runaway or captured notices when enslaved people had escaped or had been arrested. The first phase of the crowdsourcing asked volunteers to look at the full-page image, draw a rectangle around a particular notice, and classify it in one of four categories. "That worked really well," said Jevon. "Within about two or three weeks, people had drawn more than 90,000 rectangles, and once it had been processed, we ended up with a dataset of 25,000 adverts from about 12,000 newspaper pages" [65]. The second phase of the crowdsourcing consisted in asking volunteers to analyse the advertisements and input structured data (such as names, places and prices and dates). This proved challenging for contributors, in part, because the advertisements did not always follow a set format. "There were differences in the way that they were written, the style or presented and so on, and people just weren't always sure," Jevon told us [66]. The second phase therefore did not produce the level of accurate information that the project required.

3 RECOMMENDATIONS: TRAINING AND ACCESS IN THE AI ERA

In Britain, several institutions directly facilitate data-led research, such as The National Archives; National Library of Scotland; British Library and the Alan Turing Institute. GLAM Labs are also putting together tutorials

and instructions within the sector to help provide computational use of digital collections, such as the NLS's Data Foundry [67]. Furthermore, the **Digital Research Infrastructure for the Arts and Humanities (DARIAH)** brings together individual state-of-the-art digital arts and humanities activities and scales their results to a European level. DARIAH also facilitates training and education in digital methods [68]. Other initiatives include Distant Reading (Horizon 2020), which aims to establish and share best practices and "develop innovative computational methods of text analysis adapted to Europe's multilingual literary traditions" [69]. In addition, the **Computational Literary Studies Infrastructure (CLS Infra)** is building a shared and sustainable infrastructure for literary studies in the digital age, and providing a training school for digital methodologies [70]. For those wishing to enhance their skills, there are several resources that have tailored computational approaches to the humanities and social sciences.

Yet this kind of training lacks the formality and incentives associated with postgraduate programmes. As one survey participant argued, "We should establish computational methods at least in every field of the humanities and the social sciences. These methods have to become a fixed part of every curriculum." It is not the same thing to attend the Digital Humanities Summer Institute for one week or to follow a short lesson on the Programming Historian website [71], and to take a rigorous semester-long course. Making computational skills a core aspect of Master's and Ph.D. programmes would provide a formal framework and opportunities for deliberate practice and feedback [72]. It is also essential to offer advanced training to scholars who already have a good knowledge of computational approaches. As Ryan Cordell puts it, "If you're a researcher who's been using these methods for a while, there's not a clear place you would go for further development" [73].

Broadening postgraduate training is important, but it is neither possible nor desirable for everyone to become an expert in AI and other computational methods. Cross-disciplinary collaborations should be more valued (in terms of promotion and other career opportunities), to endorse positive alternatives to the model of the solo researcher. Too often, the single-authored monograph and article continue to take precedence over research outputs authored by multiple contributors. Scholars in the humanities often wait until their mid-career, once they have secured a permanent or tenured position, to engage with collaborative work that will be less valued than individual research. Yet, working with colleagues in other fields (including computer science) is beneficial, not only because it unlocks other expertise and skillsets but also because it offers the possibility of addressing complex challenges from a wide range of perspectives. Like global warming or social inequalities, solving the issue of access to digital archives requires input from multiple disciplines.

Universities and GLAM sector organisations would benefit from becoming more open to cross-disciplinary collaborations. Jenny Bunn (The National Archives UK) suggested "restructuring the workplace so that the day job is transdisciplinary…I think there is a tendency, particularly in a large organisation…to value specialism. And sometimes specialism can move into siloes" [74]. An example of this over-emphasis on specialism is the way the REF (Research Excellence Framework) is organised in the UK. Digital Humanists often submit their work in units of assessment that corresponds to their original expertise (for example, English or History), even if their collaborative work bears little resemblance with the standards of single disciplines. The problem of over-specialism is also acute in STEM. Arturo Casadevall (a distinguished microbiologist and immunologist) has compared the current system to medieval guilds. "The guild system in Europe arose in the Middle Ages as artisans and merchants sought to maintain and protect specialized skills and trades," wrote Casadevall and Ferric C. Fang. "Although such guilds often produced highly trained and specialized individuals who perfected their trade through prolonged apprenticeships, they also encouraged conservatism and stifled innovation" [75]. Specialism has a price and encouraging transdisciplinary work can unlock creative solutions to multi-faceted problems.

Funding agencies should also promote the development of infrastructures that will allow greater engagement with digital resources. New ways to deliver data are needed. Kirsten Carter (archivist at the FDR Presidential Library) told us that "datafying our existing digital collections" is a key priority to make them more accessible to researchers using computational methods [76]. Abby Gondek, who has done collaborative work with FDR archivists, gave the example of a set of letters in the FDR collection:

> We basically datafied the letters. We created a spreadsheet which included who it was from, who it was to, the content of the letter, the themes that were mentioned, the key words, the geographical location, and then some themes from the content of, for example, the gender of the sender. So, we had these different categories, then I was able to visualise those using a software called Tableau Public [77].

As this example illustrates, archives can be turned into data to be processed using software that does not require advanced technical skills. The problem is that not many tools are currently available. In a talk organised by the Society of American Archivists on 10 December 2022, the historian Ian Milligan talked about the "serious problem" faced by web archives: On the one hand, researchers will need to use these records; on the other hand, "researchers can't analyse them as the tools and supports aren't there" [78]. Likewise, John Sheridan (Digital Director, The National Archives UK) told us that too often developers are not interested in developing tools for the archive sector. "No one is building anything specifically for us" [79]. Instead, the archive sector needs to adapt tools developed for other sectors and primary uses. More funding to develop a digital infrastructure and tools tailored to archival collections would benefit the GLAM sector and researchers alike.

4 CONCLUSION

This article has outlined the many obstacles that users of digital archives face when trying to use computational methods. Too often, archives are difficult to access, or even locked altogether. When data is shared, the format used is not always user-friendly. For example, the Library of Congress, and other organizations involved in web archiving, are preserving web content in the **WARC (Web ARChive)** file format, a format that is rarely used among historians––even those with advanced computational skills. [80] Facing these difficulties to access and use digital archives, it can be tempting to focus instead on more traditional qualitative work. Spending years learning new computational methods, or honing collaborations with specialists in other fields, is not always rewarded by academic institutions. Moving to more transdisplinary work environments is needed if we are to address complex challenges that cannot be solved within single disciplines. We recommend broadening postgraduate training in the humanities and social sciences to include approaches used in data science and computer science. Analysing data at scale should be available to all postgraduates in the social sciences, humanities, and the arts. Not everyone wants to become an AI specialist, and other approaches (such as close reading) are equally valuable. But for those who want to engage with huge amounts of data, it is essential to offer both the infrastructure and the tools needed to make sense of these data. Funding agencies therefore have an important role to play to nudge the development of the digital infrastructure that will democratise computational work in the humanities and social sciences.

REFERENCES

[1] Examples of born-digital records include emails, PDFs, Word documents, audio and video digital files. These born-digital records differ from "digitised" materials that originated in paper form and were remediated thanks to processes such as scanning and photographing.

[2] Emily Maemura, Christoph Becker, and Ian Milligan. 2016. Understanding computational web archives research methods using research objects. In *Proceedings of the IEEE International Conference on Big Data (Big Data'16)*. 3250–3259.

[3] Mary Burke, Oksana. L. Zavalina, Shobhana L. Chelliah, and Mark E. Phillips. 2022. User needs in language archives: Findings from interviews with language archive managers, depositors, and end-users. *Lang. Document. Conserv.* 16 (2022), 1–24; Paul Gooding. 2022. Towards Critically Addressable Data for Digital Library User Studies. *Archives, Access and Artificial Intelligence*, Lise Jaillant (Ed.). Verlag, Bielefeld, 109–130; Matthew S. Weber. 2018. Methods and Approaches to Using Web Archives. *Comput. Commun. Res. Commun. Methods Measures* 12, 2-3 (2018), 200–215. https://doi.org/10.1080/19312458.2018.1447657; Maemura, Becker, and Milligan. 2016.

[4] See the project website. https://www.aeolian-network.net

[5] Artificial intelligence (AI) refers to the use of computational processes to learn, make decision and solve problems. Machine learning (ML), which is often referenced in discussions about AI, is an application of it. ML is the process by which a computer system is able to continue learning and improving on its own based on previous processes it has undertaken. ML is considered a subset of AI.

[6] As early as 2012, the Modern Language Association (MLA) produced "Guidelines for Evaluating Work in Digital Humanities and Digital Media" (*Journal of Digital Humanities*, 1 (2012), 4). Retrieved from https://journalofdigitalhumanities.org/1-4/guidelines-for-evaluating-work-in-digital-humanities-and-digital-media-from-the-mla/
[7] The survey on the use (or non-use) of computational research methods had 22 responses, and we interviewed 33 participants who use born-digital and/or digitised records in a research capacity.
[8] Paul Gooding. 2016. Exploring the information behaviour of users of welsh newspapers online through web log analysis. *J. Document.* 72, 2 (2016), 232–246; Mark Bell, Tom Storrar, and Jane Winters. 2022. Web Archives and the Problem of Access: Prototyping a Researcher Dashboard for the UK Government Web Archive. *Archives, Access and Artificial Intelligence*, Lise Jaillant (Ed.), Verlag, Bielefeld, 61–82.
[9] Giovanni Colavizza, Tobias Blanke, Charles Jeurgens, and Julia Noordegraaf. 2022. Archives and AI: An overview of current debates and future perspectives. *J. Comput. Cult. Herit.* 15, 1 (2022), 1–15.
[10] Beth Plale, Robert McDonald, Yiming Sun, Inna Kouper, Ryan Cobine, J. Stephen Downie, Beth Sandore Namachchivaya, and John Unsworth. 2013. Hathitrust research center: Computational access for digital humanities and beyond. In *Proceedings of the 13th ACM/IEEE-CS joint Conference on Digital Libraries*, ACM (2013), 395–396; Giovanni Colavizza, Tobias Blanke, Charles Jeurgens, and Julia Noordegraaf. 2022.
[11] Burke, Zavalina, Chelliah, and Phillips. 2022.
[12] E. T. Meyer, R. Schroeder, and J. Cowls. 2016. The net as a knowledge machine: How the internet became embedded in research. *New Media Society* 18, 7 (2016), 1159–1189. https://doi.org/10.1177/1461444816643793
[13] Maemura, Becker, and Milligan. 2016.
[14] Maemura, Becker, and Milligan. 2016.
[15] Bell, Storrar, and Winters. 2022.
[16] Mark Bell, interview, 29 June 2022, online.
[17] Lise Jaillant. 2019. After the digital revolution: Working with emails and born-digital records in literary and publishers' archives. *Arch. Manuscr.* 47, 3 (2019), 285–304. https://doi.org/10.1080/01576895.2019.1640555; Lise Jaillant. 2022. How Can We Make Born-Digital and Digitised Archives More Accessible? Identifying Obstacles and Solutions. *Arch. Sci.* 22 (2022), 417–436. https://doi.org/10.1007/s10502-022-09390-7. Lise Jaillant and Arran Rees. 2022. Applying AI to Digital Archives: Trust, Collaboration and Shared Professional Ethics. *Digital Scholarship in the Humanities.* Oxford University Press, Oxford, UK, 1–15. https://doi.org/10.1093/llc/fqac073
[18] Ryan Cordell. 2017. "'Q i-jtb the raven': Taking dirty OCR seriously." *Book Hist.* 20 (2017), 188–225. https://doi.org/10.1353/bh.2017.0006
[19] Text Creation Partnership. 2022. Retrieved from https://textcreationpartnership.org/tcp-texts/ecco-tcp-eighteenth-century-collections-online [accessed 11 November 2022].
[20] Anna Kuslits, interview, 21 June 2022, online.
[21] Stephen H. Gregg. 2020. Old Books and Digital Publishing: Eighteenth-Century Collections Online. Cambridge University Press, Cambridge, UK; Paul Fyfe. 2016. An Archaeology of Victorian Newspapers. *Victor. Period. Rev.* 49 (2016), 546–577. Retrieved from https://www.jstor.org/stable/26166577
[22] Richard Dunley and Jo Pugh. 2021. Do archive catalogues make history?: Exploring interactions between historians and archives. *Twentieth Cent. Brit. Hist.* 32, 4 (2021), 581–607 (583). https://doi.org/10.1093/tcbh/hwab021. On the issue of "algorithm-driven discovery and misleading forms of search," see also Tim Hitchcock. 2013. Confronting the Digital, Cultural and Social History. *Cult. Soc. Hist.* 10, 1 (2013), 9–23. https://doi.org/10.2752/147800413X13515292098070
[23] For more on user studies in relation to digital collections, see Gooding. 2022. Towards critically addressable data for digital library user studies.
[24] James F. English. 2010. Everywhere and nowhere: The sociology of literature after "the sociology of literature." *New Lit. Hist.* 41, 2 (2010), v–xxiii.
[25] Daniel Allington, Sarah Brouillette, and David Golumbia. 2016. Neoliberal tools (and archives): A political history of digital humanities. *Los Angeles Review of Books.* Retrieved from https://lareviewofbooks.org/article/neoliberal-tools-archives-political-history-digital-humanities/ [accessed 16 December 2022].
[26] Allington, Brouillette, and Golumbia. 2016.
[27] Anna Kuslits, interview, 21 June 2022, online.
[28] Allington, Brouillette, and Golumbia. 2016.
[29] Retrieved from https://gtr.ukri.org/projects?ref=AH%2FW001667%2F1 [accessed 24 November 2022].
[30] Graham Jevon, interview, 27 June 2022, online.
[31] Abby Gondek, interview, 11 July 2022, online.
[32] Aurélia Rostaing, interview, 27 June 2022, online.
[33] Frank Hopfgartner, interview, 29 June 2022, online.
[34] Jenny Bunn, interview, 29 June 2022, online.
[35] Treasa Harkin, interview, July 2022, online.
[36] Ryan Cordell, interview, 30 June 2022, online.

[37] Retrieved from and https://archivesforblacklives.files.wordpress.com/2019/10/ardr_final.pdf [both accessed 24 November 2022].
[38] Jenny Bunn, interview, 29 June 2022, online.
[39] Kai Naumann, interview, 4 July 2022, online.
[40] Tobias Hodel, interview, 28 June 2022, online.
[41] Anna-Maria Sichani, interview, 11 July 2022, online.
[42] Anon., interview, 5 July 2022, online.
[43] Sara Thomson, interview, 30 June 2022, online.
[44] Dunley and Pugh. 2021.
[45] Aurélia Rostaing, interview, 27 June 2022, online.
[46] Sara Thomson, interview, 30 June 2022, online.
[47] Lawrence Evalyn, interview, 30 June 2022, online.
[48] Tobias Hodel, interview, 28 June 2022, online.
[49] Hannah Ringler, interview, 30 June 2022, online.
[50] Alice Austen, interview, 30 June 2022, online.
[51] Tobias Hodel, interview, 28 June 2022, online.
[52] David De Roure, interview, 27 June 2022, online.
[53] Lawrence Evalyn, interview, 30 June 2022, online.
[54] Lawrence Evalyn, interview, 30 June 2022, online.
[55] Katherine McDonough, interview, 29 June 2022, online.
[56] Allington, Brouillette, and Golumbia. 2016.
[57] Karl Marx and Friedrich Engels. 1978. *The Marx-Engles Reader*, Robert C. Tucker (Ed.) Norton Edition, London, 476.
[58] Anon., interview, 5 July 2022, online.
[59] Katherine McDonough, interview, 29 June 2022, online.
[60] Ryan Cordell, interview, 29 June 2022, online.
[61] Mark Bell, interview, 29 June 2022, online.
[62] Alice Austen, interview, 30 June 2022, online.
[63] Jason R. Baron, interview, 23 June 2022, online.
[64] Jason R. Baron, interview, 23 June 2022, online.
[65] Graham Jevon, interview, 27 June 2022, online.
[66] Graham Jevon, interview, 27 June 2022, online.
[67] Retrieved from https://glamlabs.io/computational-access-to-digital-collections/ [accessed 24 November 2022]
[68] Retrieved from https://www.dariah.eu/activities/training-and-education/ [accessed 24 November 2022]
[69] Retrieved from https://www.distant-reading.net/ [accessed 24 November 2022]
[70] Retrieved from https://clsinfra.io/events/training-school/ [accessed 24 November 2022]
[71] Retrieved from https://programminghistorian.org [accessed 24 November 2022]
[72] While this article focuses on humanities and social science researchers, it is also important to note that computer scientists and other STEM researchers would benefit from more humanities education (for example, to design more responsible and explainable AI systems).
[73] Ryan Cordell, interview, 29 June 2022, online.
[74] Jenny Bunn, interview, 29 June 2022, online.
[75] Arturo Casadevall and Ferric C. Fang. 2014. Specialized Science. *Infection and Immunity* 82, 4 (2014), 1355–1360 (1355). https://doi.org/10.1128/IAI.01530-13
[76] Kirsten Carter, interview, 11 July 2022, online.
[77] Abby Gondek, interview, 11 July 2022, online.
[78] Ian Milligan, Society of American Archivists, 2022, online.
[79] John Sheridan, Interview for the Unlocking our Digital Past project, 29 June 2021, online.
[80] Nick Ruest, Jimmy Lin, Ian Milligan, and Samantha Fritz. 2020. The Archives Unleashed Project: Technology, Process, and Community to Improve Scholarly Access to Web Archives. *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries (JCDL'20).* ACM, New York, NY, 157–166. https://doi.org/10.1145/3383583.3398513

Received 19 December 2022; revised 30 June 2023; accepted 15 July 2023