OXFORD UNIVERSITY PRESS | Journal of Victorian Culture

AI and the Historian: Why Digital Literacy Matters Now More Than Ever.

Journal:	*Journal of Victorian Culture*
Manuscript ID	RJVC-2025-0023.R2
Manuscript Type:	Digital Forum
Keywords:	Historians, artificial intelligence, Victorian, digital history, digital archives, digital literacy, digital humanities

Zoë Alker, Lancaster University, z.alker@lancaster.ac.uk

AI and the Historian: Why Digital Literacy Matters Now More Than Ever.

Journal of Victorian Culture's Digital Forum has been leading discussions on digital transformations in Victorian Studies for well over a decade. From 2016-2019, I co-edited a series of issues with Dr Christopher Donaldson. In a 2017 forum, 'Workflow', Christopher and I cautioned readers about historians' increasing dependence on digital sources, in particular, an over-reliance on that great open sesame of digital archives: the keyword search.[1] Conscious that relying on digital collections could lead us to being less critically reflexive or forthright about our work, we encouraged scholars to open the 'black box' of digital archives and develop new forms of source criticism that acknowledged how original documents are manipulated, recoded, and reframed through the digitization and web-publication process. As Tim Hitchcock argued, 'Algorithm-driven discovery and misleading forms of search, poor OCR, and all the selection biases of a new edition of the Western print archive have changed how we research the past, and the underlying character of the object of study (inherited text)'.[2]

Since that time, the introduction of new tools and techniques, notably machine learning and its subsets, including generative artificial intelligence models such as ChatGPT, means we need to renew existing forms of historical data criticism. As novel digital research infrastructures and tools emerge, how will this impact on our understanding of the nineteenth century and even on how we 'do' history? In the era of misinformation – fake news, conspiracy theories, and deepfakes – data literacy is an increasingly urgent issue for scholars, students and citizens alike. Even in relevant subject areas, like English and History, where digital humanities techniques are applicable and powerful, some academics still fail to recognize the importance of digital literacy to staff and students and the contribution that digital humanities can make. Yet the core competencies of the historians' toolkit – source criticism, provenance, historic contextualisation, uncovering bias – are crucial for developing digital literacy, and Digital Victorianists are uniquely placed to contribute.

This article reflects upon on my experiences co-developing a series of databases and web resources over the past decade. Most recently, I have been involved in two digital history projects: *Data Mining Convict Tattoos* (British Academy/ JISC, 2019) which examined the largest number of tattoos ever recorded: 75,688 descriptions of tattoos, on 58,002 convicts in Britain and Australia from 1793 to 1925 and *Skin and Bone: Interdisciplinary Analysis of Accidents, Injury and Interpersonal Violence in London, 1760-1901* (British Academy/ Leverhulme, 2021-22) which merged convict, hospital and osteoarchaeological datasets and documented 87,903 injuries on 50,659 Londoners, revealing the physical impact of the Industrial Revolution on the body. These projects emerged from the Digital Panopticon (Arts

[1] Zoë Alker and Christopher Donaldson, 'Workflow', *Journal of Victorian Culture*, 22.2 (2017), 222–223.
[2] OCR, or optical character recognition, is the automated conversion of images of printed text to machine-readable and -searchable text. Tim Hitchcock, 'Confronting the Digital: Or How Academic History Writing Lost the Plot', *Cultural and Social History*, 10.1 (2013), 9–23.

and Humanities Research Council, 2014) which created a single, searchable database of 250,000 individuals sentenced at the Old Bailey between 1780 and 1868, and assessed the impact of varying modes of punishment on their lives. Key to all these projects was the reuse and repurposing of data from the *Old Bailey Online* project, and associated databases including criminal registers, prison licences, hospital admissions records, civil records, and osteoarchaeological datasets. *Skin and Bone* and *Convict Tattoos* shared common ground in other areas. The projects developed bespoke techniques derived from machine learning, including data mining, natural language processing, automated record linkage, and data visualization. The projects reused and linked together a wide range of historic datasets, enabling innovative computational research. We made the data and programming codes as open and accessible as possible: to academics and to the wider public.[3] In this article, I will reflect upon what we can learn about the opportunities and limitations of computational humanities research for Victorian Studies and the importance of digital humanities skills for developing digital literacies in the era of artificial intelligence.

Historians take messy, incomplete, and often disparate sources that are full of errors and biases and make critical interpretations and arguments that respond to their research questions. Digital sources and technologies don't fundamentally alter this, but they do add a layer of technical complexity.[4] Developing techniques for dealing with errors is core to digital history work, and we found that iterative processes were key to the critical evaluation of the data. On both the *Convict Tattoos* and *Skin and Bone* projects, our aim was to extract previously hidden data about either tattoos or scars and injuries from the written physical descriptions of approximately 250,000 convicts in the *Digital Panopticon* collection of databases. Extracting and analyzing information from physical descriptions of convicts is not straightforward, because these descriptions, often entered in the same column on a form, contain a wide range of other information in a variety of formats. The challenge in these projects, then, was to extract the relevant information from all the other information in these descriptions.

In *Skin and Bone*, for example, this required distinguishing between language used to describe injuries, wounds and bodily impairments from the wider information contained within the physical descriptions in the criminal records (eye colour, hair colour, complexion, height, weight, and other distinguishing marks such as boils and pockmarks), and we needed to standardize the varied terms used to describe injuries and body parts across the different record collections (criminal, hospital, osteoarchaeological).[5] We couldn't apply machine learning

[3] Zoë Alker and Robert Shoemaker, *Criminal Tattoos: Analysing Criminal Tattoos through Data Mining and Visualisation* (2022) <https://www.dhi.ac.uk/projects/criminal-tattoos/> [accessed 29 July 2025]. *Tattoos in the Digital Panopticon Database, 1793–1925*, The University of Sheffield, Dataset (2022), doi.org/10.15131/shef.data.13398665.v1. Zoë Alker et al., *Skin and Bone: Interdisciplinary Analysis of Accidents, Injury and Interpersonal Violence in London, 1760–1901* (2023) <https://www.dhi.ac.uk/data/skin-and-bone> [accessed 29 July 2025].

[4] Ian Gregory, 'Challenges and Opportunities for Digital History', *Frontiers in Digital Humanities*, 17 (2014) <https://www.frontiersin.org/journals/digital-humanities/articles/10.3389/fdigh.2014.00001/full> [accessed 31 January 2024].

[5] The data comprised eleven datasets that included convict descriptions, hospital admissions, and data collected on skeletal remains. For further information, see Zoë Alker *et al.*, *Skin and Bone: Interdisciplinary Analysis of*

because that would have required training data, and we didn't have that. So, we developed an approach that used bespoke dictionaries, reflecting the data and domain expertise developed and built upon through various projects since the *Old Bailey Online*, and applied rules-based methods of extraction, classification, and analysis. We found that the computing tools required to carry out the projects could not reliably be achieved with any commercial or off-the-shelf packages, so we developed a bespoke process iteratively, combining automated processes of rules-based learning with manual checking through several iterations. The importance of interdisciplinary collaboration and iterative processes here cannot be overstated. Both projects involved multi-disciplinary teams that included historians, web developers, software engineers and other heritage professionals. The iterative process was essential in drawing on both technical expertise and domain-specific knowledge as it allowed for the continuous refinement of methods and interpretations. Developing the ability to collaborate and communicate evolving project needs is a long-term skill—one that short-term funding often hinders. Yet, in digital scholarship, multidisciplinary teams are essential and cannot be substituted by the entirely automated roles that commercial Large Language Models (LLMs) tend to promote.[6]

We developed a computational humanities approach, combining human and machine intelligence, to better understand and interrogate our source material. Big historical data will always be prone to error, inconsistency, and mess, in the same way the original source material will be, and when using multiple record sets, we need to account for the varied nature of recording on a larger, more complex scale. We found that, because they allowed us to take a macroscopic view of the datasets, data visualizations were especially useful in exposing gaps, errors and biases in the sources that wouldn't be picked up by the naked eye. We experimented with different forms of data visualization, such as bar charts, collocations and heat maps to help us identify meaningful patterns in the data, but also to understand how they were complicated by the particular distinguishing features of the sources used (for example, the uneven survival of records across time, the varied nature of recording by the original institutions and the distorting effect of having one very large dataset from the later nineteenth century- the Metropolitan Police Habitual Criminals Register.[7]

Exploratory, machine-driven reading of historic 'big' data can reveal statistically meaningful patterns illegible to the human eye, but these techniques need to be combined with close reading, so that the data can be understood both at the scale of the full dataset, and at the individual datum in its fullest evidential context. We need to understand how that information was retrieved, check the outputs for accuracy, and interpret our findings accordingly. Andreas Fickers terms this 'scalable reading': 'Learning to move easily between these two forms of reading will require training a new generation of historians in a new cultural technique of

Accidents, Injury and Interpersonal Violence in London, 1760–1901 (2023) <https://www.dhi.ac.uk/data/skin-and-bone> [accessed 29 July 2025].

[6] Lauren Tilton, 'Relating to Historical Sources', *The American Historical Review*, 128.3 (2023), 1354–1359.

[7] Zoë Alker and Robert Shoemaker, 'Convicts and the Cultural Significance of Tattooing in Nineteenth Century Britain', *Journal of British Studies*, 61.4 (2022), 835–862. For a general guide to data visualizations, see Yale University Library, 'Data Visualization', *Yale University Library Research Guides* (2025) <https://guides.library.yale.edu/datavisualization> [accessed 29 July 2025].

information retrieval and interpretation which I frame as “scalable reading”’.[8] This level of data literacy is portable, as Fickers argues, not just to academic research, but to a wide range of disciplines and professional sectors. As we move further into the AI era, combining human and machine intelligence will continue to be fundamental, but doing so effectively requires us to develop new forms of historical source criticism. This isn’t to suggest that all scholars must become digital humanists or computer scientists, but as digital tools become increasingly commonplace in our research and teaching, knowledge of the underlying data and algorithms must be central to our source criticism. Indeed, ‘digital humanists do not need to understand algorithms *at all.* They do need, however, to understand the transformations that algorithms attempt to bring about’.[9] Artificial intelligence increasingly underpins the tagging, classification, organization, and filtering processes that determine which digital sources historians encounter.[10] Commercial information companies like ProQuest are developing algorithms that use features such as phrase recognition to enhance filtered searching and relevance rankings. Increasingly, for-profit companies like Proquest are building databases of digital sources and applying AI to improve search functionality, while creating data mining tools like ProQuest’s TDM Studio to market as new products.[11] Consequently, we face the risk of surrendering not just historical sources but entire research methodologies to corporate control.[12] But as Lauren Tilton contends, ‘[u]nderstanding the algorithms that built the path through the collection that led to the results is becoming an important feature for understanding our evidence.[13] Future historians must be aware of how history is increasingly filtered and shaped through algorithmic decision-making shaped by AI. Integrating these skills is crucial in ensuring the discipline doesn’t involve a slackening of academic standards. Scholars need to equip current and future humanities students with two essential skills: first, embed computational humanities techniques into undergraduate and postgraduate curricula to demonstrate the potential of technology for humanities scholarship, and second, foster critical, politically aware engagement with web-based tools like ChatGPT, focusing sharply on bias, provenance, and ethics. Digital humanities techniques should not revolve solely around ‘the digital’ but should be used as a tool to make new knowledge that contributes to existing historical research. As Ian Gregory contends, ‘The work that will ultimately prove the relevance and importance of digital resources and methods will not stress the digital, it will stress the applied and contribute to knowledge on particular topics within history that “non-digital” historians will be interested in’.[14]

[8] Andreas Fickers, ‘The Future of History in the Digital Age’, *Radar: The Science & Diplomacy Anticipator* (2023) <https://radar.gesda.global/the-future-of-history-in-the-digital-age> [accessed 20 July 2024].

[9] Benjamin M. Schmidt, ‘Do Digital Humanists Need to Understand Algorithms?’ in eds. Matthew K. Gold and Lauren F. Klein, *Debates in the Digital Humanities* (University of Minnesota Press, 2016), <https://dhdebates.gc.cuny.edu/read/untitled/section/557c453b-4abb-48ce-8c38-a77e24d3f0bd> [Accessed 20 January 2024].

[10] Lauren Tilton, ‘Relating to Historical Sources’, *The American Historical Review*, **128**.3 (2023), 1354–1359, doi.org/10.1093/ahr/rhad365.

[11] Ibid., p. 1355.

[12] Ibid.

[13] Ibid.

[14] Ian Gregory, ‘Challenges and Opportunities for Digital History’, *Frontiers in Digital Humanities*, 17 (2014) <https://www.frontiersin.org/journals/digital-humanities/articles/10.3389/fdigh.2014.00001/full> [Accessed 31 January 2024].

Digital humanists are already examining the opportunities of using LLMs for work that extends beyond text, especially as AI capabilities increasingly support multimodal analysis including images and maps.[15] It remains to be seen whether LLMs will revolutionize digital humanities; certainly, modes of access (open source versus commercial) and the models' capabilities for learning context bring challenges. Looking to the future, I could imagine the possibilities of building bespoke LLMs which could be relevant to sub-disciplines in History. As a crime historian, I could envisage a bespoke LLM relating to crime, law and associated materials in the eighteenth to twentieth centuries. The data could incorporate the wide range of open access datasets available, including smaller databases collected by individuals in their research, and associated metadata would provide accurate provenance. Machine translation and transcription in LLMs, including AI tools such as Transkribus, provide tools for extracting, classifying and translating rare and indigenous languages, making once marginalized histories more widely visible and accessible.[16] Such work both preserves linguistic heritage and makes non-English academic resources available to a broader audience.[17] ChatGPT-4 can power through some steps in the historians' toolkit – improved OCR, data cleaning, complex, multivariate analysis, data visualization, and translation, for example – but commercial LLMs fail spectacularly at *being* the historian. ChatGPT and other commercial generative AI tools use webscraping to feed and train their models, meaning that bias, presentism, and a lack of provenance are all current barriers to using these tools for historical research and teaching. The models are unable to assess authenticity, cross-check information across different records, nor can they be ethical or self-reflexive. But perhaps a bespoke LLM, driven by historical data and informed by critical contextualization and developed by multidisciplinary teams, can help the historians of the future navigate the Victorian and contemporary digital world in new ways, all whilst maintaining those core skills and competencies that are relevant to both historical study and cultural citizenship.

Bio: Zoë Alker is a digital historian of nineteenth-century crime and punishment. Her work focuses on histories of gender, violence and the body, primarily amongst the working classes. With colleagues she has created a series of resources helping to give the public direct access to an extensive range of primary sources evidencing the history of modern Britain, including Digital Panopticon, Convict Tattoos, and Skin and Bone.

[15] Taylor Arnold and Lauren Tilton, 'Explainable Search and Discovery of Visual Cultural Heritage Collections with Multimodal Large Language Models', *CHR 2024: Computational Humanities Research Conference* (Aarhus University, Denmark, 4–6 December 2024) <https://arxiv.org/abs/2411.04663> [accessed 31 March 2025]; Katherine McDonough, Kaspar Beelen, Daniel C. S. Wilson, & Rosie Wood (2024). Reading Maps at a Distance: Texts on Maps as New Historical Data. *Imago Mundi*, Volume 76, Issue 2, pp. 296–307, doi.org/10.1080/03085694.2024.2453336.

[16] Paty Murrieta-Flores, Rodrigo Vega-Sánchez, Alexander Sánchez-Diaz, and Hector Cruz-Ríos, 'Unlocking Colonial Records with Artificial Intelligence: Achieving the Automated Transcription of Large-Scale 16th- and 17th-Century Latin American Historical Collections', *STAR: Science & Technology of Archaeological Research,* 11.1 (2025), doi.org/10.1080/20548923.2025.2484828.

[17] Andrea Cigliano, Francesca Fallucchi, and Marco Gerardi, 'The Impact of Digital Analysis and Large Language Models in Digital Humanity', in *ICYRIME 2024: 9th International Conference of Yearly Reports on Informatics, Mathematics, and Engineering* (Catania, 29 July – 1 August 2024) <https://ceur-ws.org/Vol-3869/p01.pdf> [accessed 31 March 2025].