
Article:
Nockels, J. orcid.org/0000-0002-4577-6596, Gooding, P. orcid.org/0000-0003-1044-509X and Terras, M. orcid.org/0000-0001-6496-3197 (2024) The implications of handwritten text recognition for accessing the past at scale. Journal of Documentation, 80 (7). pp. 148-167. ISSN 0022-0418

https://doi.org/10.1108/jd-09-2023-0183

Received 18 September 2023
Revised 1 December 2023
Accepted 5 December 2023

The implications of handwritten text recognition for accessing the past at scale

Joseph Nockels
School of Literatures Languages and Cultures, The University of Edinburgh, Edinburgh, UK
Paul Gooding
College of Arts and Humanities, University of Glasgow, Glasgow, UK, and
Melissa Terras
Design Informatics, Edinburgh College of Art, University of Edinburgh, Edinburgh, UK

Abstract

Purpose – This paper focuses on image-to-text manuscript processing through Handwritten Text Recognition (HTR), a Machine Learning (ML) approach enabled by Artificial Intelligence (AI). With HTR now achieving high levels of accuracy, we consider its potential impact on our near-future information environment and knowledge of the past.

Design/methodology/approach – In undertaking a more constructivist analysis, we identified gaps in the current literature through a Grounded Theory Method (GTM). This guided an iterative process of concept mapping through writing sprints in workshop settings. We identified, explored and confirmed themes through group discussion and a further interrogation of relevant literature, until reaching saturation.

Findings – Catalogued as part of our GTM, 120 published texts underpin this paper. We found that HTR facilitates accurate transcription and dataset cleaning, while facilitating access to a variety of historical material. HTR contributes to a virtuous cycle of dataset production and can inform the development of online cataloguing. However, current limitations include dependency on digitisation pipelines, potential archival history omission and entrenchment of bias. We also cite near-future HTR considerations. These include encouraging open access, integrating advanced AI processes and metadata extraction; legal and moral issues surrounding copyright and data ethics; crediting individuals' transcription contributions and HTR's environmental costs.

Originality/value – Our research produces a set of best practice recommendations for researchers, data providers and memory institutions, surrounding HTR use. This forms an initial, though not comprehensive,

We acknowledge the contributions of Andy Stauder, Managing Director of READ, and Guenter Mühlberger, Founder and Director of READ, throughout the research process, as well as community members of the wider READ COOP. We thank Sarah Ames, from the National Library of Scotland, for insights.

Declaration of conflicting interests: Terras serves on the Board of Directors of Transkribus as Research Director. Transkribus is one of Nockels's industrial PhD partners. This research has ethical approval via the READ-COOP, the National Library of Scotland and the University of Edinburgh's processes.

Funding: Nockels's doctoral research is funded by the UK's Arts and Humanities Research Council Grant Number AH/RO12717/1. For the purpose of open access, the authors have applied a creative commons attribution (CC BY) licence to any author accepted manuscript version arising from this submission. Terras is funded by the AHRC Creative Industries Clusters Programme, with support from the Scottish Funding Council and the Edinburgh and South East Scotland City Region Deal, Award Reference AH/S002782/1.

Journal of Documentation
Vol. 80 No. 7, 2024
pp. 148-167
Emerald Publishing Limited
0022-0418
DOI 10.1108/JD-09-2023-0183

blueprint for directing future HTR research. In pursuing this, the narrative that HTR's speed and efficiency will simply transform scholarship in archives is deconstructed.

Keywords Digital libraries, Grounded theory, Information environment, Optical character recognition, Machine learning, Handwritten text recognition

Paper type Article

Introduction

Before Handwritten Text Recognition (HTR), manuscripts were costly to convert to machine-processable text for research and analysis. With HTR now achieving high levels of accuracy, we ask what near-future behaviour, interaction, experience, values and infrastructures may occur when HTR is applied to historical documents? When combined with the mass-digitisation of content held by galleries, libraries, archives and museums (GLAM), how will HTR's application, use, and affordances generate new knowledge of the past, and affect our information environment? This paper's findings emerge from a literature review surveying current understanding of the impact of HTR, to explore emerging issues over the coming decade. We aim to deconstruct the simplistic narrative that the speed, efficiency and scale of HTR will "transform scholarship in the archives" (Muehlberger *et al.*, 2019, p. 955), providing a more nuanced consideration of its application, possibilities and opportunities. In doing so, our recommendations will assist researchers, data and platform providers, memory institutions and data scientists to understand how the results of HTR interact with the wider information environment.

We find that HTR supports the creation of accurate transcriptions from historical manuscripts, and the enhancement of existing datasets. HTR facilitates access to a greater range of materials, including endangered languages, enabling a new focus on personal and private materials (diaries, letters), increasing access to historical voices not usually incorporated into the historical record, and increasing the scale and heterogeneity of available material. The production of general training models leads to a virtuous digitisation circle where similar datasets are easier – and therefore more likely – to be produced. This leads to the requirement for processes that will facilitate the storage and discoverability of HTR generated content, and for memory institutions to rethink search and access to collections. Challenges include HTR's dependency on digitisation, its relation to archival history and omission and the entrenchment of bias in data sources. The paper details several near future issues, including: the potential of HTR for the basis of automated metadata extraction; the integration of advanced Artificial Intelligence (AI) processes (including Large Language Models (LLMs) and generative AI) into HTR systems; legal and moral issues such as copyright, privacy and data ethics which are challenged by the use of HTR; how individual contributions to shared HTR models can be credited; and the environmental costs of HTR infrastructure. We identify the need for greater collaboration between communities including historians, information scientists and data scientists to navigate these issues, and for further skills support to allow non-specialist audiences to make the most of HTR. Data literacy will become increasingly important, as will building frameworks to establish data sharing, data consent and reuse principles, particularly in building open repositories to share models and datasets. Finally, we suggest that an understanding of how HTR is changing the information environment is a crucial aspect of future technological development.

HTR: an overview

Optical Character Recognition (OCR), the electronic translation of generally printed documents into machine-readable text, came into general use during the 1950s for business operations such as mail sorting (Schantz, 1982, p. 7). By the late 1990s, with the

advent of personal computing and improvements made to scanning technology, OCR began to provide cheap and effective text recognition, with large-scale machine processable text resources of variable accuracy being created from mass-digitised *printed* content (Tanner *et al.*, 2009). Advanced OCR approaches now integrate Machine Learning (ML) techniques, improving accuracy rates while still relying on printed character isolation, achieving Character Error Rates (CERs) as low as 0.5% on historical print (Reul *et al.*, 2019, p. 1).

Unlike OCR, developments in HTR focus on the recognition of text which has no consistent font. Recent advances in AI and ML have constructed predictive language models which can decipher handwritten text lines sequentially (Strauss *et al.*, 2017, p. 6). Since the late 2010s, HTR has been viewed as a solved computational problem (Muehlberger *et al.*, 2019). Before this, full text searching at scale of handwritten texts was difficult (Estill and Levy, 2016). The aspiration of recognising handwriting from people of various backgrounds, nationalities, professions and education, with equal competence, accuracy and speed, has largely been realised; challenging notions that such tools should only be "... used to recognise common keywords in historical texts" (Cirone and Spirling, 2021, p. 19). HTR transcriptions hold great potential for quantitative approaches to analysis of the past, including cleaning, marking-up and analysing datasets, and providing novel data on which to train other computational systems.

A variety of HTR platforms exist. Transkribus [1], operated by Recognition and Enrichment of Archival Documents (READ) under a cooperative of 85 institutions and 50 individuals (as of May 2023), spanning 30 countries, is the largest consumer-level HTR system. Transkribus has a reported 127,332 active users, collectively processing 12,946 pages daily. Since 2015, 46.9 million images of historical documents have been uploaded and 21,032 HTR models have been trained using convolutional neural networks (CNN) (READ-COOP, 2023). Transkribus models trained on printed-text, such as the French language model trained by Brando and Becquet, on the Directory of owners and properties in Paris and the Seine (1898–1923), have registered CERs as low as 1%. Models on handwritten text have registered CERs around 5%, such as Hodel's model for 19th century German current writing [2]. Elsewhere, Monk is a HTR tool with a different approach, "... adding labels at the page-description level, adding line transcriptions at the level of line-strip images and adding zone labels for words and characters ...", forming a tailored index for a collection (Schomaker, 2019, p. 225). Since 2009, Monk has been applied to 15th century texts, as well as Chinese and Arabic scripts [3]. As of 2013, Monk's servers hosted a total of 370,000 harvested and human-confirmed word labels and 20,000 documents. eScriptorium [4] is another HTR software using neural networks; although it remains reliant on character isolation, tagging glyphs with set characteristics and using corresponding variant image files for model training. Loghi, [5] an open-source HTR platform that aims to make scanned historic documents digitally readable and searchable, was unveiled in April 2023, from the KNAW (Royal Netherlands Academy of Arts and Sciences) and the Nationaal Archief of the Netherlands.

Efforts have also been made by publishers such as Adam Matthew (AM) digital to utilise HTR to elevate under-represented historical sources. Their HTR, Quartex, uses a similar neural network to Transkribus, making "linguistical assumptions" to predict language patterns [6]. Quartex was used in 2020 by Baylor University Libraries to decipher 40 handwritten pages from the 19th century poet Robert Browning, and seven others, with a 90% accuracy rate for Browning's hand (72% across all scribes) [7]. Major technological providers are now also moving into this space, although their generalised tools are not yet specialised on historical handwriting from particular periods, or allowing bespoke models to be trained and improved. Google has provided the ability to detect and transcribe handwriting within its Cloud Vision Application Programming Interface (API) since 2015 [8] and Microsoft has supported the conversion of images to text via its handwritten input panel in Windows 10 since 2017 (Cottuli, 2022).

HTR can now generate trustworthy machine-readable texts at low cost, especially compared to traditional human-led transcription. Combined with mass-digitisation, HTR can allow the complete transcription of collections of historical documents. It therefore holds the possibility to profoundly change what is available as part of the digital record.

HTR and the digital past

Recently, the library profession has responded to mass digitisation by reimagining the relationship between libraries, collections and the digital, with a focus on how to make collections more amenable to computational methods. Padilla (2017) notes that viewing "collections as data" begins with the need to reframe all digital objects as data. In particular, there has been a move to destabilise established concepts and ways of working, including: breaking down organisation silos which developed to deal with analogue resources (Padilla *et al.*, 2019, p.15); understanding how data structures inform our conceptualisation of memory work (Lincoln, 2017); addressing how material complexities and collection biases are recognised via data provenance (Padilla, 2017; Ames, 2021); and developing new ways of publishing library content for emerging audiences (Candela *et al.*, 2020, pp. 1–2). Simultaneously, we have had to reconsider how access and reuse of data occurs. Many interactions with GLAM collections are informational, in that they are mediated through discovery and search interfaces (Gooding, 2017, p. 173). This has necessitated a reconsideration of how interfaces might allow users more flexibility in engagement with digital sources (Whitelaw, 2015), and an exploration of the limitations and implications of digital materiality for historians and GLAM collections (Conway, 2013; Gooding, 2023). However, despite recognition of the impact and potential of HTR (Cordell, 2020, pp. 25–26), its impact has yet to be addressed to the same extent.

Recent developments in HTR disrupt and extend our relationship to historical handwritten documents, their use to understand human history and society, and how the data contained within them may feed into technological systems, including becoming training for emergent AI systems. This paper therefore provides a timely consideration of how abundant machine-processable text generated from handwritten documents will affect the wider information environment and our relationship with the past. To establish key issues in the development and utilisation of HTR, various aspects need to be taken into account. These aspects include those related to research and collections such as: research into corpora, from the lens of handwritten sources, and the impact HTR will have on access. Issues related to tool and dataset development supporting HTR are also considered. This article also focuses on HTR's unique affordances to research by detailing how historians can develop novel methods and approaches. In establishing how HTR can enable new avenues for research, we demonstrate how training and professional development can be developed specifically for HTR and its resultant datasets. Finally, the ethical and moral aspects of deriving data from handwritten texts via AI are considered.

Methodology

We developed a Grounded Theory Method (GTM) to undertake a thematic analysis of the current impact of HTR in research and the broad information environment. This method was deployed to identify the remaining unanswered questions concerning the implications of HTR in the near future by using extant literature as a robust evidence-base for a systematic horizon scanning exercise. As such, we largely followed Glaser and Strauss (1967), analysing qualitative data from which concepts could emerge, although our GTM dealt with the identification of gaps in literature, following a more constructivist approach. This involved reading general texts to provide context before the coding of categories, providing more

"theoretical sensitivity" in inferring emerging concepts (Glaser and Holton, 2004, p. 11). Following Charmaz (2006, pp. 96–100), the coding process was completed in multiple reiterative stages, selecting and categorising texts, and returning to the literature until thematic saturation was reached. We used the scholarly materials identified and catalogued by Nockels *et al.* (2022) which cite Transkribus in published research as a foundational list of projects to consider.

In addition to this, materials were collected through purposive sampling, beginning with Rosenzweig's germinal article *Scarcity or Abundance? Preserving the Past in a Digital Era* (2003). We charted articles which cited Rosenzweig (2003) in Google Scholar [9], expanding our reading to incorporate those that took debates beyond history, discussing the wider information environment and forming a cycle of sampling, data collection and analysis to identify gaps and omissions. We organised references relevant to our consideration of HTR in Zotero [10]. In line with Hanson (2017), negative case analysis was then undertaken in order to remove irrelevant works from the corpus. This formed a stronger understanding of the context of individual arguments and broader debates concerning HTR's impact, based on 131 published texts [11]. A close reading of relevant texts to identify their concepts via content analysis (Krippendorff, 2004), was followed by an iterative process of concept mapping among the authors in a workshop setting, beginning on the 19th of May, 2022 at the University of Edinburgh's Bayes Centre, an innovative hub for AI and data science research, and writing sprints, while considering wider literature in aligned fields. This process concluded in April 2023. Emerging themes were identified, explored and confirmed via group discussion and further interrogation of the relevant literature. This results in a thematic analysis that explores the recent societal impact of HTR, in turn providing a roadmap for considering the technology's near future (next ten years) implications.

Results

The following sections provide detail on the extant themes identified through our GTM collated from literature relating to HTR's impact on the information environment. Having identified notable gaps in the literature, what follows thereafter is a consideration of the near future implications of HTR on research, dataset creation and documentation; digital infrastructure, ethical and legal frameworks as well as the environment.

Recent impact of HTR

Building accurate datasets. With accurate transcriptions from HTR, historians can begin to confidently use large datasets derived from handwritten texts, instead of relying on in-person visits to archives (Cohen, 2010) and manual transcription.

HTR offers a step-change in accuracy beyond OCR for printed text. Cordell (2017, p. 194) argues that scholars remain unaware of the inaccuracy of OCR-transcribed text, which can contain complex errors that affect search, processing and interpretation (Conway, 2013, p. 18). Zaagsma (2019, p. 842) stresses that without the correction of OCR datasets, which is necessary for historical scholars to accurately read texts at full-length, societal memory construction could be influenced. HTR has therefore been framed as a way to create more accurate transcripts enabling work on collections at scale (Kaukonen, 2021). In the same way that OCR technology and keyword searches opened up ". . . a new level of everyday cultural discourse" (Nicholson, 2013, p. 67) for typewritten materials, the increased accuracy of transcriptions generated by HTR holds similar potential for keyword searching of handwritten materials at scale, a task which previously relied on an array of expertise or required collaborative infrastructures (Unsworth and Tupman, 2016, p. 231). For example, Amsterdam City Archives crowdsourced 10,000 pages of accurate transcriptions from their

16th-17th century Notary Archives. This ground truth data was used to train an HTR model, then applied over several hundred thousand pages [12]. Scholars have since used this resource to discover "new details of a mixed and growing population in long-gone neighbourhoods, proof of new global connections ..." and "the provenance of numerous European paintings" [13].

Cleaning existing datasets. HTR provides the means to create more accurate transcriptions from existing digital images, particularly those with complex fonts or layouts. In 2018, Vienna City Library applied OCR to their Lehmann address books, a record of all the City's main tenants between 1859 and 1942, but it struggled to recognise Fraktur script. Subsequently, they turned to Transkribus, producing a fully searchable resource of approximately 200,000 pages (Egger, 2021). The National Library of Finland constructed a workflow to reprocess almost two million Finnish newspaper pages from 1771–1918 using Transkribus, replacing insufficient OCR-derived text with more accurate transcriptions (Kaukonen, 2021). Institutions can therefore use HTR to revise existing workflows to enhance the accessibility and usability of the digitised record. This relies on seeing digitised archives as in a constant state of correction, thinking beyond the captured resources toward decision making at content-holding organisations (Cordell, 2017, p. 207). There is opportunity here, too, to mathematically quantify improved rates of accuracy of HTR, validated against previously OCRed datasets, for benchmarking purposes.

Access to a greater range of language materials. The ability to train HTR to recognise a large range of languages can enhance usage of materials that may have been previously neglected. In the case of Eastern European languages, the Kazakh Offline Handwritten Text Dataset (KOHTD) (Toiganbayeva *et al.*, 2022), a dataset of Cyrillic exam papers, and the Handwritten Kazakh and Russian (HKR) database, have recently emerged (Nurseitov *et al.*, 2021). These databases are essential for training and the eventual recognition of texts, allowing these materials to be consulted and analysed for the first time, as exemplified by Abdallah *et al.* (2020, p. 141) producing models with CERs as low as 0.045% and WERs as low as 0.192% using the HKR dataset.

Access to a greater range of personal materials. HTR makes it much easier to transcribe sources from the pre-industrial and pre-mechanical (or at least, pre-print and typewriter) era, increasing their searchability and transformability, and in doing so, changing research questions that can be asked at scale, and themes and periods of focus. In addition, where OCR was limited to printed text (mostly officially prepared for expected audiences), HTR facilitates machine-processable text of private, intimate documents, including personal letters, diaries, institutional correspondence, manuscripts, editions of works, ledgers, accounts and official records such as census materials. This allows different topics, approaches, and questions to be broached. Ólafsson's work (2004, p. 1), analysing the duality of Icelandic book history in the 19th century, shows the difference between the printed and handwritten past (including scribal copies and personal texts). There is a resultant need to audit how handwritten materials vary in scale and content from print, highlighting which gaps exist in the historical record and directing computational methods to extrapolate information from the sources that remain.

Virtuous HTR digitisation circles via general HTR models. With vast datasets of relevant handwritten texts and vocabulary registers now accessible, there are further opportunities for training and re-training of general HTR models, which in turn facilitates the publishing of larger general models across greater time periods, further transcription generation and greater searchability across a wider range of collections. In April 2023, Transkribus released two LLMs: "Dutchess I" combining four previous Dutch models across the 17th-19th centuries, trained by the National Archives of the Netherlands and Amsterdam City Archives, returning a CER of 4.3% [14]; and "German Giant" trained on material spanning the 16th to 21st century in Latin and Kurrent scripts, with a CER of 8.3% [15]. These general algorithms act to increase the pool of training data available for researchers training bespoke

models on controlled corpora [16], and the resulting data created can, in turn, train targeted and specific LLMs. As researchers adopt and improve these models for their own projects, general algorithms can grow more accurate and cover broader time periods.

Access to endangered languages. HTR holds the possibility of preserving endangered language texts, increasing their accessibility and profile. Valy *et al.* (2020) present a novel approach using a Recurrent Neural Network (RNN) to generate machine-readable text from ancient Khmer palm leaf manuscripts. Having accessible databases of training data is also essential for the recognition of nearly lost languages: as Vu *et al.* (2021) highlight for the Old Degraded Vietnamese Handwritten Script Archive (IHR-Nom) database, a collection of 15th century Vietnamese ChuNom handwriting written using Latin characters, and largely displaced after 1920 by modern Vietnamese. Presently, fewer than 100 scholars regularly read the language (Vu *et al.*, 2021, p. 86). Such projects aid transliteration, the process of changing a language from one script or alphabet to another, previously a painstaking task due to the degradation of source material. This approach facilitates a broader culture of inclusion as documents become more accessible to non-specialist audiences and may require a reconsideration of how search is done, especially across global north/global south divides. At-risk materials may benefit from this approach, as we have seen from the development of Ukrainian HTR language models (READ-COOP, 2023) to support the digitisation effort for Ukrainian culture (Saving Ukrainian Cultural Heritage Online (SUCHO), 2023), given the war with Russia and the coordinated destruction of their memory institutions (Marche, 2022). However, questions remain over how indigenous and endangered language communities retain fair control of, and access to, their archival pasts (Owens, 2018, p. 166; Turner, 2020).

Access to multiple voices. Archival material containing multiple hands are often harder to read as each scribe has distinctive handwriting. HTR opens up these collections; for example, Bluche *et al.* (2014, p. 161) used neural networks to recognise Arabic text from a dataset of 455 scribes located throughout Egypt, Iraq, Sudan, Morocco and Algeria. These scribes are not contained within current historical narratives, and so different voices are brought into the digital historical record via HTR, with the potential to reconstitute different corpora to encompass a wider range of perspectives which do not usually end up in the published record, for example personal records relating to woman. With HTR increasing accessibility to collections, details can emerge along the archival grain, revealing what Stoler (2008, pp. 1–3) describes as the nature of imperial rule managed through the written record, and colonial sense, leading to more explicit prejudices in social reality.

Utilising the results of HTR. As the scale of digital archives increases, the use of "machine methods for making sense of this massive database of historical text is no longer a luxury - it is an imperative" (Kelly, 2013, p. 69). HTR allows a greater range of information to become structured, findable, searchable and readable, facilitating both human and algorithmic usage. The technology, therefore, can inform the development of online catalogues and digital records (Moss *et al.*, 2016, p. 120). As the range and scale of the digital corpus grows, users can increasingly adopt the data-led, distant reading approaches outlined by Underwood (2017). HTR therefore helps to stimulate computer use alongside existing discipline-specific practices (Van Lange, 2023, p. 12), supporting a "macroscopic" (Graham *et al.*, 2016) approach allowing researchers to move between individual perspectives and macro histories with greater ease.

We therefore need to consider how users navigate the emerging "infinitive archive", described by Turkel *et al.* (2012) as instantly accessible, machine-readable, growing exponentially and constantly being reordered. Enrichment and mark-up of HTR transcriptions can help increase accessibility and structure of collections, including tagging named entities in a TEI-compliant manner [17]. This can result in "radiant textuality", enabling resources with dynamic multi-layers of expression, viewable in different ways using index searches (Thomas, 2004, p. 65). read&search [18], an aligned tool to

Transkribus, automatically assembles components of texts that are displayed in a user-friendly manner using an International Image Interoperability Framework (IIIF) compliant interface [19], thus making collections accessible in a standards-led manner. This allows for something close to a critical edition to be created (Terras *et al.*, forthcoming 2024), holding implications for search capability. The Amsterdam City Archives have used read&search to make hundreds of thousands of handwritten pages from the 17th-18th centuries Amsterdam notarial archives searchable by users [20]. Thematic resources are also available: 16th-17th recessions in Low German cities [21], New Zealand Alpine Heritage [22] and historical mining in Tyrol, Austria [23]. Hanson and Simenstad (2018) show how use of HTR can rethink crowdsourcing interfaces: volunteers on platforms such as Zooniverse [24] normally identify lines of text, reaching a consensus by combining transcription data (Hanson, 2017). HTR models can leverage the online template of crowdsourcing platforms to augment the human transcription process by predicting text [25], redefining user interactions with digital records.

HTR limitations. Most content-holding institutions are grappling with inclusively and sensitively reframing their collections, updating descriptions to reflect changing language (Chew, 2021). This follows several notable failures within the GLAM sector: neglecting that they are keeping "souls in their stacks" (Drake, 2021, p. 8). HTR has the potential to disrupt who is excluded, disposed, disinherited and disembodied by current archival practices. However, the potential for an archive to be discoverable depends on its archival history: "Archives [can be read] as sources of history … but they are also its subjects … with histories and politics of their own" (Yale, 2015, p. 332). Many archives were and are assembled to document and reinforce certain identity narratives (Stoler, 2002). Owens (2018, p. 166) notes that GLAM institutions "… have served and by default, in many cases, continue to serve as infrastructures of colonialism, and oppression. We need to do better." No amount of transcription can return lost sources (Cunningham, 1999), with many subaltern communities sustaining loss and damage of their records, making details irrecoverable.

HTR is dependent on digitisation (Terras, 2022, p. 193), but GLAM's patchwork approach to digitisation strategies, workflows and priorities (Zaagsma, 2019) mean collections teem "with diverse political, legal, and cultural investments and controversies" (Thylstrup, 2019, p. 3), in a way that undermines positivist ideas of archival practices being neutral (Tschan, 2002, p. 176). The provenance of digital resources and the potential motivations in authorising certain parts of heritage into account (Hauswedell *et al.*, 2020, p. 140) becomes increasingly important: those deploying HTR technology on historical collections need to be critical of how they may follow or resist literary, historical and cultural canons. Without this, there is a real risk that HTR outputs only serve to cement hegemonic archival structures. Well-researched and documented collections which have been previously digitised will return stronger models, given they provide extensive training data: "accordingly, the resulting models are highly biased by the material they are trained on" (Hodel, 2022, p. 171). Depending on the training data, HTR can voice colonial artefacts (Stoler, 2008), prioritising the coloniser and not the colonised. For instance, the main Dutch language model was trained on 3 million pages of the 17th-18th century Dutch East India Company Verenigde Oostindische Compagnie (VOC) records [26]. This "set the groundwork for later parts of the digitisation strategy" of the National Archief of the Netherlands [27].

Although HTR improves the readability of handwritten artefacts and produces information retrieval resources (Terras *et al.*, forthcoming 2024), it consolidates an increasing scholarly reliance on databases and keyword searching as the primary way to access archival content. It therefore becomes increasingly harder to understand the material archive, the corpora, upon which our models are based. HTR shifts handwritten texts towards collections as data (Padilla, 2017), in the same way Whitelaw (2015) reimagined search in light of mass-digitisation: we will need to consider whether current search and discovery interfaces are well suited to understanding the complexities of handwritten

sources, and whether current documentation practices truly reflect the interdependencies of HTR generated data sources. In rethinking search and filtering capabilities, set criteria could assess the current levels of source criticism, generosity in discoverability; user content management, exploration and connectivity enabled in interface design (Ehrmann *et al.*, 2019, pp. 4–5). Failure to do so could result in an ". . . increasingly remote and unvisited shadowland into which even quite important texts fall if they cannot yet be explored [or identified] by . . . electronic means" (Leary, 2005, p. 82). In addition, the use of keyword spotting (kws) can result in what Nicholson calls "keyword blinkers" (2013, p. 61) that bypass the wider context of manuscripts and focus solely on textual information, risking the rise of anecdotal studies. Ewing *et al.* suggests that with macroscale digital approaches historians will still need to ". . . accept the 'messiness' of large amounts of data . . ." (2014). Guldi and Armitage (2014, p. 103), argue that digital technologies require a ". . . cautious and judicious curating of possible data, questions, and subjects." Given that "[T]he more automated and efficient our systems of digital discovery become, the harder it can be to look from the flood of data before us" (Putnam, 2016, p. 399), new skills will have to be developed for search and analysis of the results of HTR, requiring the redesign of interfaces to incorporate and navigate archival complexity.

Near future issues for the use of HTR with historical documents

We now turn to topics which were identified via horizon scanning, rather than from our thematic analysis. By signposting current gaps in considering HTR, this section anticipates future developments, possible affordances as well as potential risks in expanding the provision of AI-enabled automated transcription methods.

Encouraging open data practices. The need for HTR to be trained on large quantities of data challenges the siloed nature of many digital GLAM collections, many of which are only accessible behind paywalls, with underlying datasets only available from commercial publishers via permission. This applies to access to datasets for model training, the availability of historical datasets for research, and the publishing of resulting HTR models. GLAM institutions must consider the balance between commercial activities and adoption of FAIR data principles (Wilkinson *et al.*, 2016). Cordell (2020, pp. 44–45) has proposed that libraries develop statements of values for utilising ML technologies. These principles may ensure that the application of HTR within memory institutions proceeds based on core principles of transparency and openness. Recently, HTR United has formed to create a catalogue of training datasets, to support the wider community in accessing materials [28].

Integrating the results of HTR into collection systems and processes. HTR provides opportunities for the description and cataloguing of collections. Automated systems exist which transform HTR results into formats which allow better integration with delivery and analysis systems, supporting findability and reuse, such as the International Image Interoperability Framework (IIIF) (Transkribus and IIIF) and Text Encoding Initiative (TEI) guidelines [29]. Additionally, HTR can support metadata extraction (Skluzacek *et al.*, 2022), with kws facilitating search against tagged named entities such as place names, dates and individuals [30]. This may allow greater archival sensitivity in metadata extraction and creation, ensuring that resources better represent audiences (Havens, 2020). HTR's wider adoption will not automatically improve an institution's metadata, and we should be mindful of "wishing on" technology to solve social separation without enacting new codes of practice (Seely Brown and Duguid, 2000, pp. 15–16). Best practice should be developed for the results of HTR to be used in automated metadata extraction.

Integration with advanced AI processes. HTR – itself a product of ML – has potential for further integration with AI tools and systems. For example, the use of deep learning on epigraphic inscriptions from Classical Athens has shown potential for restoring missing

texts, and attributing original locations (Assael *et al.*, 2022). As of May 2023, Transkribus is investigating how ChatGPT can be integrated into its infrastructure (Stauder, 2023, personal communication), after initial experiments by the Vienna City Library, using a LLM for further HTR output correction (Muehlberger, 2023, p. 13). The integration of HTR tools and datasets into LLM and ML processes provides potential for novel synthesis, comparison and analysis of handwritten data, which will require best practices to avoid disintermediation. In addition, new approaches to information literacy (Pettersson, 2022) are needed to establish the legitimacy of primary sources, including the ability to differentiate between machine-generated information and the human, handwritten record (Donaldson and Conway, 2015). With increased integration, particularly with LLMs, careful documentation is also essential (Bender *et al.*, 2021, p. 618).

HTR and legal frameworks. Intellectual Property issues, including copyright, intersect with the digitisation of historical materials. GLAM institutions currently conduct thorough investigations and risk assessments regarding orphaned works, those materials which have no clear copyright holder (Korn, 2009, p. 23), consuming considerable staff time (Secker and Morrison, 2016) and leading some institutions to restrict the use of materials. Generally, enquirers must search for necessary permissions (Korn, 2009, p. 23). HTR work is likely to make deducing the ownership of digitised content more complex. Who will ensure that a work uploaded to an HTR platform is out of copyright? Who owns copyright in the resulting machine-processable text: could these transcriptions be defined as "novel" intellectual products needing protection? How do Intellectual Property Rights of documents intersect with those of machine-learning models which contain their data? Non-western communities often have different frameworks and understanding of ownership and respect regarding the content of historical documents (Romein *et al.*, 2024, p. 18): how is this complicated by ingestion into shared data-models via HTR?

In addition, General Data Protection Regulation (GDPR) issues may arise with the use of HTR. Privacy has always been a concern for historians: there is a responsibility to the dead as well as the living (Lawrence, 2016). However, removal of human moderators may exacerbate situations where "information about the dead can cause social harm to the living" (Lawrence, 2016, p.13). Handwritten materials often contain traceable personal information, and disclosure may have implications not only for authors, but their next of kin and descendants. Anonymisation, data minimisation and timely deletion may be necessary to comply with regulatory contexts, with specific strategies needed for the ethical and legal handling of particular materials. Institutions and individuals should consider privacy when preparing to transcribe mass-digitised private correspondence. Sensitivity review is already a common task for archivists, including with born-digital collections such as email correspondence (Smith, 2021), but a significant increase in the accessibility and volume of handwritten materials may require consideration of when manual review becomes impractical and automated processes are needed (McDonald *et al.*, 2019).

Crediting HTR contributions. HTR's wider adoption conflicts with formal research structures built to credit labour: it remains unclear how to credit individual tasks like model training and the editing and correction of transcriptions. Given that academic credit affects academic career development (Gao *et al.*, 2022), attention must be given to publication of HTR work, and a balance achieved between rights reserved, open-source licensing and named credits (Romein *et al.*, 2024).

Data ethics and bias. Biases and trauma permeate GLAM collections, including: the amplification of historical elisions which reinforce the focus on dominant groups; records of colonial exploitation; and detailed accounts of persecution, atrocities, crime, etc. These are often compounded by explicit and implicit institutional biases which shape and control our access to the past, which then affects and frames our digital collections (Havens *et al.*, forthcoming 2024). New paradigms are therefore needed to mitigate difficult histories embedded in handwritten texts. This may be exacerbated when using sources at scale, or

when this data becomes integrated into wider AI systems. This requires the establishment of a data-ethics led approach to HTR, recognising that work with historical documents should be conducted with a consideration of their societal and ethical contexts.

Environmental costs of HTR. HTR, AI and ML processes are computationally intensive, consuming energy and increasing carbon output. An attempt to quantify the approximate environmental costs of training neural network models for 24 hours found advanced transformer models emit the same carbon emissions as a trans-Atlantic flight (Strubell *et al.*, 2019, p. 4). In 2022, READ unveiled that they were beginning to train HTR models using similar transformer methods (Stauder, 2022).

Digital humanities researchers are increasingly aware of the environmental footprint of their activities (Digital Humanities Climate Coalition, Information, Measurement and Practice IMP Action Group, 2022), while historians recognise the need to debate and research climate change (McNeill, 2016). This follows climate-conscious initiatives prioritising efficient hardware and algorithms, such as SustainNLP (Bender *et al.*, 2021, p. 612), and projects associated with "AI for Good", using computation to solve societal problems (Aula and Bowles, 2023, p. 5). However, methodologies are needed to calculate the carbon footprint of ML research activities (Lacoste *et al.*, 2019), beyond carbon offsetting (Passalacqua, 2021), speaking to existing debates around climate and personal-corporate responsibility (Cuomo, 2011). AI processes can embed biases at the expense of the same marginalised communities already being adversely impacted by the climate crisis (Bender *et al.*, 2021, p. 613). Platform providers are encouraged to be more transparent regarding HTR's environmental footprints in order to allow institutions and individuals to make informed choices about their use.

Recommendations

Given our analysis above, we present several best practice recommendations for collections holders, data providers and data users when creating, stewarding, or utilising HTR generated content. Together, these recommendations form an anticipatory approach that can be tailored through trainings and skills development, clarification over collection management policies and the adoption of already present guidelines. We also signal potential future work to further hone our horizon scanning exercise.

Collaboration in tool and platform development. We recommend that engagement with both the underlying text recognition and development of tools associated with HTR should act as an impetus for collaboration across disciplines, acting as a "trading zone" (Kemman, 2021, pp. 39–58). This should extend to the sharing of outputs including datasets, models, transcriptions, platforms, documentation, research approaches, storage and hosting mechanisms. Terras (2022, p. 183) describes a longstanding community of interdisciplinary researchers developing HTR systems. Collaboration is required to develop best practice in documentation, establish shared vocabularies and workflows, and to ensure that tools and platforms meet the needs of users.

Skills support. Ongoing training and community support is essential for HTR uptake, particularly given advances in related AIs. Professional development activities are needed in: HTR familiarity including limitations and potential errors; the role of traditional palaeography skills; methods to verify accuracy; error detection and best practice in reviewing outputs; and collaboration mechanisms to critique and support technological developments. These require development and additional funding, given GLAM resourcing constraints.

Information documentation and information literacy. There are no standards for cataloguing HTR generated datasets. The GLAM sector need to consistently integrate HTR data sources into content management systems, allowing users to understand datasets, both facilitating and enabling reuse. As HTR datasets proliferate, those using the results of

HTR created by others will have to develop information literacy skills to understand provenance, accuracy, representativeness and relationship to the archival record, to draw upon them to understand the past. Verification and validation of data sets will be necessary before basing analysis upon them, including understanding any biases or limitations present.

Archival inclusion. It is not a question of whether technologies like HTR should be deployed but how, balancing discovery with sensitivity (Odumosu, 2020). Embedding values of inclusion into HTR application is complex and will take conscious planning. We recommend that developers and users of HTR take an activist stance to resist the notion of archival neutrality, considering the interests of a wider range of stakeholders including under-represented and marginalised groups (Romein *et al.*, 2024). In this way, the deployment of HTR (aligned with choices made in digitisation) could form a societal good, naming oppression when seen (D'Iganzio and Klein, 2020) and allowing users to bear greater witness to difficult histories. However, there are concerns regarding the absorption of HTR generated content into LLMs and other AI systems: breaking the link to archival context, and possibly transmitting harmful biases.

Establishing data sharing and data consent principles. HTR uses data created by humans, while outputting data sources which can go on to have other unplanned outcomes. We recommend that individuals, institutions and projects undertaking HTR make their data FAIR where possible (Findable, Accessible, Interoperable, Reusable) [31] but also follow the CARE principles for Indigenous Data Governance: producing data of collective benefit, with authority control, in a responsible manner, holding ethics as a primary concern at all stages of the data life cycle across the data ecosystem [32].

Use of HTR outputs. It is currently not known how best to embed values of transparency, searchability, accessibility and support into HTR technologies, platforms and generated datasets. We recommend that attention is given to: producing open repositories with reusable shareable datasets; how these datasets will be preserved; infrastructure allowing the cross-walking of HTR outputs and content-management systems using more ML approaches; documentation and licensing standards; finding and searching aids; visualisation tools and further publication arenas.

Speculating HTR design. To construct a robust vision of how HTR might impact the near-future of historical research, we recommend the introduction of speculative design methods, which encompass a range of methods such as design fiction, critical design or design for debate (Rüller *et al.*, 2022). This will allow developers and researchers to better frame how HTR could change scholars' engagement with the past, stimulating discussion and providing a "perceptual bridge" around engaging audiences to inform design priorities and aesthetic and functional decisions (Auger, 2013, pp. 11–12). Speculative design has occurred in collaboration with university labs and community groups (Baumann *et al.*, 2017) involving focus groups with users and developers (Rüller *et al.*, 2022). In utilising such methods, users' expectations for platform providers can be clarified to ensure that HTR meets current and emerging stakeholder needs.

Conclusion

In laying out the current impact and future implications of HTR, this article elucidates how abundant digital transcriptions generated from handwritten documents will affect our relationship with the past. Our emerging knowledge of HTR therefore understands the historical digitised record as a constantly forming digital archive with a unique place in the machine-processable present.

This paper has demonstrated that a structured consideration of how recent AI-based tools are changing our access to the past and its sources can be crucial to understanding how to develop best practice in the handling of the results of the process, but also in informing the

future development of tools which represent user needs and inclusive best practice. The method we have used here may also be of benefit when considering the impact of other specific AIs on the information we can access, use and reuse: such high-level consideration of our rapidly changing information environment will allow identification of near future issues for the information and data science community, ensuring that we centre user needs, while minimising harms.

There is an inherent complexity in including information from handwritten historical documents in data-rich environments. This results in a moral responsibility to embed values of transparency, searchability, accessibility and support into HTR technologies, given the changes they are making to the wider information environment. To navigate these complex issues will require collaboration, skills support and the development of best practice principles and protocols, to ensure the future makes the most of our handwritten past.

Notes

1. https://readcoop.eu/transkribus/
2. These models are open access and can be viewed in Transkribus see: https://transkribus.eu.lite.
3. https://www.ai.rug.nl/~lambert/Monk-collections-english.html
4. https://www.escriptorium.uk/escript.html
5. https://github.com/knaw-huc/loghi
6. https://www.amdigital.co.uk/create/am-quartex/htr-and-ocr
7. https://digitalcollections-baylor.quartexcollections.com/abl-collections/the-browning-letters
8. 'https://cloud.google.com/vision/docs/handwriting
9. https://scholar.google.com/
10. https://www.zotero.org/
11. The collated texts gathered in Zotero have been made accessible through Zenodo see: https://zenodo.org/records/10229247
12. These pages are now searchable using Transkribus's Read&Search interface, https://transkribus.eu/r/amsterdam-city-archives/#/
13. https://readcoop.eu/success-stories/amsterdam-notary-archives/
14. https://readcoop.eu/model/the-dutchess-i/
15. https://readcoop.eu/model/the-german-giant-i/
16. https://readcoop.eu/glossary/base-models/
17. The TEI is a consortium which develops and maintains standards for representing texts. 'TEI", https://tei-c.org/. Transkribus allows transcriptions to be exported in TEI-compliant format, although how much editing is necessary to ensure that these products fit set standards is still under research. 'How to enrich transcribed documents with mark-up", Transkribus. https://readcoop.eu/transkribus/howto/how-to-enrich-transcribed-documents-with-mark-up/
18. https://readcoop.eu/readsearch/
19. https://iiif.io/
20. https://transkribus.eu/r/amsterdam-city-archives/#/
21. https://transkribus.eu/r/rezesse-niederdeutscher-staedtetage/#/
22. https://www.nzaj-archive.nz/#/
23. https://transkribus.eu/r/mining-hub/#/

24. https://www.zooniverse.org/
25. https://github.com/danhan52/text_recognition
26. https://readcoop.eu/model/the-dutchess-i/
27. https://readcoop.eu/success-stories/national-archives-of-the-netherlands/
28. https://htr-united.github.io
29. https://tei-c.org/
30. For more information on named entity recognition (NER) see: Appendix 5 – named entitites: current definitions in Nouvel *et al.* (2016), pp. 153–158.
31. https://www.go-fair.org/fair-principles/
32. https://www.gida-global.org/care

References

All URIs accessed 31st July 2023.

Ames, S. (2021), "Transparency, provenance and collections as data: the national library of Scotland's data foundry", *LIBER Quarterly*, Vol. 31, pp. 1-13, doi: 10.18352/lq.10371.

Abdallah, A., Hamada, M. and Nurseitov, D. (2020), "Attention-based fully gated cnn-bgru for Russian handwritten text", *Journal of Imaging*, Vol. 6 No. 9, pp. 141-150, doi: 10.48550/arXiv.2008.05373.

Assael, Y., Sommerschield, T., Shillingford, B., Bordbar, M., Pavlopoulos, J., Chatzipanagiotou, M., Androutsopoulos, I., Prag, J. and de Freitas, N. (2022), "Restoring and attributing ancient texts using deep neural networks", *Nature*, Vol. 603 No. 7900, pp. 280-283, doi: 10.1038/s41586-022-04448-z.

Auger, J. (2013), "Speculative design: crafting the speculation", *Digital Creativity*, Vol. 24 No. 1, pp. 11-35, doi: 10.1080/14626268.2013.767276.

Aula, V. and Bowles, J. (2023), "Stepping back from AI and Data for Good - current trends and ways forward", *Big Data and Society*, Vol. 10 No. 1, pp. 1-12, January-June, doi: 10.1177/20539517231173901.

Baumann, K., Stokes, B., Bar, F. and Caldwell, B. (2017), "Infrastructures of the imagination: community design for speculative urban technologies", *Paper Presented at the 2017 Communities and Technologies Conference*, Troyes, June, 2017, pp. 266-269, doi: 10.1145/3083671.3083700.

Bender, E.M., Gebru, T., McMillan-Major, A. and Shmitchell, S. (2021), "On the dangers of stochastic parrots: can Language Models Be too big?", *ACM Conference on Fairness, Accountability, and Transparency*, Online, 3-10 March, 2021, pp. 610-623, doi: 10.1145/3442188.3445922.

Bluche, T., Louradour, J., Knibbe, M., Moysset, B., Benzeghiba, M.F. and Kermorvant, C. (2014), "The A2iA Arabic handwritten text recognition system at the open HaRT2013", in *2014 11th IAPR International Workshop on Document Analysis Systems*, Tours, 1-10 April, 2014, pp. 161-166, doi: 10.1109/DAS.2014.40.

Candela, G., Dolores Saez, M., Escober Esteban, M. and Marco-Such, M. (2020), "Reusing digital collections from GLAM institutions", *Journal of Information Science*, Vol. 48 No. 2, pp. 251-267, doi: 10.1177/0165551520950246.

Charmaz, K. (2006), *Constructing Grounded Theory: A Practical Guide through Qualitative Analysis*, Sage Publications, Thousand Oaks, CA, New Delhi.

Chew, C. (2021), "Recording. Non-discriminatory library cataloguing practices for sound and moving image", available at: https://scotlands-sounds.nls.uk/index.php/2021/10/08/non-discriminatory-library-cataloguing-practices-for-sound-and-moving-image/

Cirone, A. and Spirling, A. (2021), "Turning history into data: data collection, measurement and inference in HPE", *Journal of Historical Political Economy*, Vol. 1 No. 1, pp. 127-154, doi: 10.1561/115.00000005.

Cohen, D. (2010), *Blog post. Is Google Good for History?*, available at: https://dancohen.org/2010/01/07/is-google-good-for-history/v

Conway, P. (2013), "Preserving imperfection: assessing the incidence of digital imaging error in HathiTrust", *Digital Technology and Culture*, Vol. 42 No. 1, pp. 17-30, doi: 10.1515/pdtc-2013-0003, available at: https://hdl.handle.net/2027.42/99522

Cordell, R. (2017), "'Q i-jtb the raven': taking dirty OCR seriously", *Book History*, Vol. 20 No. 1, pp. 188-225, doi: 10.1353/bh.2017.0006.

Cordell, R. (2020), "Machine learning + libraries", Library of Congress, available at: https://labs.loc.gov/work/experiments/newspaper-navigator/

Cottuli, M. (2022), "New handwriting experiences come to Windows 10 Insider build 16215 for PC", OnMSFT, 28 December, 2022, available at: https://www.onmsft.com/news/new-handwriting-experiences-come-to-windows-10-insider-build-16215-for-pc/

Cunningham, A. (1999), "Waiting for the ghost train: strategies for managing electronic personal records before it is too late", *Archival Issues*, Vol. 1 No. 1, pp. 1-10.

Cuomo, C.J. (2011), "Climate change, vulnerability, and responsibility", *Hypatia*, Vol. 26 No. 4, pp. 690-714, doi: 10.1111/j.1527-2001.2011.01220.x, available at: https://www.jstor.org/stable/41328876

D'Iganzio, C. and Klein, L. (2020), *Data Feminism*, MIT Press, Cambridge.

Digital Humanities Climate Coalition, Information, Measurement and Practice (IMP) Action Group, Baker, J., Ohge, C., Otty, L. and Walton, J.L. (Eds) (2022), *A Researcher Guide to Writing a Climate Justice Oriented Data Management Plan*, Report, Online, April, doi: 10.5281/zenodo.6451499.

Donaldson, D.R. and Conway, P. (2015), "User conceptions of trustworthiness for digital archival documents", *Journal of the Association for Information Science and Technology*, Vol. 66 No. 12, pp. 2427-2444, doi: 10.1002/asi.23330.

Drake, J. (2021), "Blood at the root", *Journal of Contemporary Archival Studies*, Vol. 8 No. 6, pp. 1-26.

Egger, A. (2021), "Transkribus projects at the Vienna city library", in *READ-COOP Success Stories*, available at: https://readcoop.eu/success-stories/vienna

Ehrmann, M., Bunout, E. and Düring, M. (2019), "Historical newspaper user interfaces: a review", *Proceedings of the 85th International Federation of Library Associations and Institutions (IFLA) General Conference and Assembly*, Athens, August 2019, pp. 1-24, available at: https://zenodo.org/records/3404155

Estill, L. and Levy, M. (2016), "Chapter 12: evaluating digital remediations of women's manuscripts", *Digital Studies/Le champ numérique, Beyond Accessibility: Textual Studies in the Twenty-First Century*, Vol. 6 No. 6, doi: 10.16995/dscn.12.

Ewing, E.T., Gad, S., Hausman, B.L., Kerr, K., Pencek, B. and Ramakrishnan, N. (2014), "Blog post. Mining coverage of the flu: big data's insights into an epidemic", Perspectives on History (AHA), available at: https://www.historians.org/publications-and-directories/perspectives-on-history/january-2014/mining-coverage-of-the-flu-big-datas-insights-into-an-epidemic

Gao, J., Nyhann, J., Duke-Williams, O. and Mahony, S. (2022), "Gender influences in Digital Humanities co-authorship networks", *Journal of Documentation*, Vol. 78 No. 7, pp. 327-350, doi: 10.1108/jd-11-2021-0221.

Glaser, B. and Holton, J. (2004), "Remodelling grounded theory. Forum qualitative sozialforschung/forum", *Qualitative Social Research*, Vol. 5 No. 2, pp. 1-22, doi: 10.17169/fqs-5.2.607.

Glaser, B. and Strauss, A. (1967), *The Discovery of Grounded Theory Strategies for Qualitative Research*, Sociology Press, Mill Valley, CA.

Gooding, P. (2017), *Historic Newspapers in the Digital Age: 'Search All about it'*, Routledge, Abingdon.

Gooding, P. (2023), "Informational abundance and material absence in the digitised early modern press: the case for contextual digitisation", in Brownlees, N. (Ed.), *The Edinburgh History of the British and Irish Press*, Edinburgh University Press, Edinburgh, Beginnings and Consolidation 1640-1800, Vol. 1, pp. 586-598.

Graham, S., Milligan, I. and Weingart, S. (2016), *Exploring Big Historical Data, the Historian's Macroscope*, Imperial College Press, London.

Guldi, J. and Armitage, A. (2014), *The History Manifesto*, Cambridge University Press, Cambridge.

Hanson, A. (2017), "Negative case analysis", in Allen, M. (Ed.), *The International Encyclopaedia of Communication Research Methods*, Wiley & Son, New York, pp. 1-3.

Hanson, D. and Simenstad, A. (2018), "Combining human and machine transcriptions on the zooniverse platform", in *2018 EMNLP Workshop W-Nut: 4th Workshop on Noisy User-Generated Text*, Brussels, pp. 215-216, available at: https://aclanthology.org/W18-6129.pdf

Hauswedell, T., Nyhan, J., Beals, M.H., Terras, M. and Bell, E. (2020), "Of global reach yet of situated contexts: an examination of the implicit and explicit selection criteria that shape digital archives of historical newspapers", *Archival Science*, Vol. 20 No. 9, pp. 139-165, doi: 10.1007/s10502-020-09332-1.

Havens, L. (2020), "Blog post. Exploring collections as data with jupyter notebooks", National Library of Scotland Data Foundry, available at:, https://data.nls.uk/project/exploring-collections-as-data-with-juypter-notebooks

Havens, L., Alex, B. and Terras, M. (forthcoming 2024), "Confronting gender biases in heritage catalogues: a natural language processing approach to revisiting descriptive metadata", in Ashton, J. (Ed.), *The Routledge Handbook on Heritage and Gender*, Routledge, London.

Hodel, T. (2022), "Supervised and unsupervised: approaches to machine learning for textual entities", in Jaillant, L. (Ed.), *Archives, Access and Artificial Intelligence*, Bielefeld University Press, Bielefeld, pp. 157-178.

Kaukonen, M. (2021), "Improved text recognition for Finnish historical newspapers with transkribus", READ-COOP Success Stories, available at: https://readcoop.eu/success-stories/improved-text-recognition-for-finnish-historical-newspapers-with-transkribus/

Kelly, T.M. (2013), *Teaching History in the Digital Age*, University of Michigan Press, Ann Arbor, doi: 10.3998/dh.12146032.0001.001.

Kemman, M. (2021), *Trading Zones of Digital History*, De Gruyter, Berlin/Boston.

Korn, N. (2009), *In from the Cold: an Assessment of the Scope of 'Orphan Works' and its Impact on the Delivery of Services to the Public*, Naomi Korn Associates, London.

Krippendorff, K. (2004), *Content Analysis: an Introduction to its Methodology*, Sage, London.

Krull, F. and Muehlberger, G. and Terras, M. (2019), "Transkribus and IIIF: beneficial possibilities between image sharing and handwritten text recognition frameworks", *IIIF Conference*, Göttingen, 24-28 June 2019, available at: https://www.pure.ed.ac.uk/ws/portalfiles/portal/215672397/Transkribus_and_IIIF_beneficial_possibilities_between_image_sharing_and_Handwritten_Text_Recognitio.pdf

Lacoste, A., Luccioni, A., Schmidt, V. and Dandres, T. (2019), "Quantifying the carbon emissions of machine learning", *arXiv. Preprint.*, doi: 10.48550/arXiv.1910.09700.

Lawrence, S.C. (2016), *Privacy and the Past: Research, Law, Archives, Ethics*, Rutgers University Press, New Brunswick, NJ.

Leary, P. (2005), "Googling the victorians", *Journal of Victorian Culture*, Vol. 10 No. 1, pp. 72-86, doi: 10.3366/jvc.2005.10.1.72, available at: https://victorianresearch.org/Googling_the_Victorians.htm

Lincoln, M. (2017), "Ways of forgetting: the librarian, the historian, and the machine", in *National Forum Position Statements. Always Already Computational: Library Collections as Data National Forum*, available at: https://collectionsasdata.github.io/aac_positionstatements.pdf

Marche, S. (2022), "'Our Mission is Crucial': meet the warrior librarians of Ukraine", The Guardian, available at: https://www.theguardian.com/books/2022/dec/04/our-mission-is-crucial-meet-the-warrior-librarians-of-ukraine

McDonald, G., Macdonald, C. and Ounis, I. (2019), "The FACTS of technology-assisted sensitivity review", in *Workshop on Fairness, Accountability, Confidentiality, Transparency, and Safety in Information Retrieval (FACTS-IR (SIGIR'19 Workshop)*, 25 July 2019, Paris, available at: https://arxiv.org/pdf/1907.02956.pdf

McNeill, J.R. (2016), "Historians, superhistory, and climate change", in Jarrick, A., Myrdal, J. and Wallenberg Bondesson, M. (Eds), *Methods in World History, A Critical Approach*, Nordic Academic Press, Lund, pp. 19-43.

Moss, M., Thomas, D. and Gollins, T. (2016), "The reconfiguration of the archive as data to Be mined", *The Journal of Association of Canadian Archivists*, Vol. 1, pp. 1-34.

Muehlberger, G. (2023), "Transkribus for archives or how artificial intelligence is revolutionizing access to historical documents", *Deep-L. Pre-print.*, pp. 1-20.

Muehlberger, G., Seaward, L., Terras, M., Ares Oliveira, S., Bosch, V., Bryan, M., Colutto, S., Déjean, H., Diem, M., Fiel, S., Gatos, B., Greinoecker, A., Grüning, T., Hackl, G., Haukkovaara, V., Heyer, G., Hirvonen, L., Hodel, T., Jokinen, M., Kahle, P., Kallio, M., Kaplan, F., Kleber, F., Labahn, R., Lang, E.M., Laube, S., Leifert, G., Louloudis, G., McNicholl, R., Meunier, J.-L., Michael, J., Mühlbauer, E., Philipp, N., Pratikakis, I., Puigcerver Pérez, J., Putz, H., Retsinas, G., Romero, V., Sablatnig, R., Sánchez, J.A., Schofield, P., Sfikas, G., Sieber, C., Stamatopoulos, N., Strauß, T., Terbul, T., Toselli, A.H., Ulreich, B., Villegas, M., Vidal, E., Walcher, J., Weidemann, M., Wurster, H. and Zagoris, K. (2019), "Transforming scholarship in the archives through handwriting text recognition, Transkribus as a case study", *Journal of Documentation*, Vol. 75 No. 50, pp. 965-967, available at: https://www.emerald.com/insight/content/doi/10.1108/JD-07-2018-0114/full/html

Nicholson, B. (2013), "The digital turn: exploring the methodological possibilities of digital newspaper archives", *Media History*, Vol. 19, pp. 59-73, doi: 10.1080/13688804.2012.752963.

Nockels, J., Gooding, P., Ames, S. and Terras, M. (2022), "Understanding the application of handwritten text recognition technology in heritage contexts: a systematic review of Transkribus in published research", *Archival Science*, Vol. 22, pp. 1-26, doi: 10.1007/s10502-022-09397-0.

Nouvel, D., Ehrmann, M. and Rosset, S. (2016), *Named Entities for Computational Linguistics*, Wiley, New York.

Nurseitov, D., Bostanbekov, K., Kurmankhojayev, D., Alimova, A., Abdallah, A. and Tolegenov, R. (2021), "Handwritten Kazakh and Russian (hkr) database for text recognition", *Multimedia Tools and Applications*, Vol. 80 Nos 21-23, pp. 1-23, doi: 10.1007/s11042-021-11399-6.

Odumosu, T. (2020), "The crying child: on colonial archives, digitization, and ethics of care in the cultural commons", *Current Anthropology*, Vol. 61 No. 22, pp. 289-302, doi: 10.1086/710062.

Ólafsson, D. (2004), "Sagas in handwritten and printed books in 19th century Iceland", *Sagas and Societies*, Vol. 11 No. 1, pp. 1-11.

Owens, T. (2018), *The Theory and Craft of Digital Preservation*, John Hopkins University Press, Baltimore.

Padilla, T. (2017), *On a Collections as Data Imperative*, UC, Santa Barbara, available at: https://escholarship.org/uc/item/9881c8sv

Padilla, T., Allen, L., Frost, H., Potvin, S., Russey Roke, E. and Varner, S. (2019), "Final report - always already computational: collections as data", available at: https://zenodo.org/record/3152935#.X6WOf-LPzIU

Passalacqua, A. (2021), "The carbon footprint of a scientific community: a survey of the historians of mobility and their normalized yet abundant reliance on air travel", *The Journal of Transport History*, Vol. 42 No. 1, pp. 121-141, doi: 10.1177/0022526620985073.

Pettersson, K. (2022), "Teaching information literacy in the humanities: engaging students with primary sources and cultural heritage material", *Nordic Journal of Information Literacy in Higher Education*, Vol. 13 No. 1, pp. 56-62, doi: 10.15845/noril.v13i1.3782.

Putnam, L. (2016), "The transnational and text-searchable: digitized sources and the shadows they cast", *The American History Review*, Vol. 121 No. 2, pp. 377-402, doi: 10.1093/ahr/121.2.377.

READ-COOP (2023), "Transkribus daily report", May 12, 2023, Internal project documentation.

Reul, C., Christ, D., Hartelt, A., Balbach, N., Wehner, M., Springmann, U., Wick, C., Grundig, C., Buttner, A. and Puppe, F. (2019), "OCR4all - an open-source tool providing a (semi-) automatic OCR workflow for historical printings", *Applied Sciences*, Vol. 9 No. 22, pp. 4853-4883, doi: 10.48550/arXiv.1909.04032.

Romein, C.A., Hodel, T., Gordijn, F., Zundert, J.J.V., Chagué, A., Lange, M.V., Jensen, H.S., Stauder, A., Purcell, J., Terras, M., Heuvel, P., van den, Keijzer, C., Rabus, A., Sitaram, C., Bhatia, A., Depuydt, K., Afolabi-Adeolu, M.A., Anikina, A., Bastianello, E., Benzinger, L.V., Bosse, A., Brown, D., Charlton, A., Dannevig, A.N., Gelder, K.V., Go, S.C.P.J., Goh, M.J.C., Gstrein, S., Hasan, S., Heide, S.V.D., Hindermann, M., Huff, D., Huysman, I., Idris, A., Keijzer, L., Kemper, S., Koenders, S., Kuijpers, E., Rønsig Larsen, L., Lepa, S., Link, T.O., Nispen, A., van, Nockels, J., Noort, L.M.V., Oosterhuis, J.J., Popken, V., Estrella Puertollano, M., Puusaag, J.J., Sheta, A., Stoop, L., Strutzenbladh, E., Sijs, N.V.D., Spek, J.P.V.D., Trouw, B.B., Van Synghel, G., Vučković, V., Wilbrink, H., Weiss, S., Wrisley, D.J. and Zweistra, R. (2024), "Exploring data provenance in handwritten text recognition infrastructure: sharing and reusing ground truth data, referencing models, and acknowledging contributions. Starting the conversation on how we could get it done." *Journal of Data Mining and Digital Humanities*. Special Issue: Historical Documents and automatic text recognition, pp. 1-26. doi: 10.46298/jdmdh.10403.

Rosenzweig, R. (2003), "Scarcity or abundance? Preserving the past in a Digital Era. The *American historical review*", *The American Historical Review*, Vol. 108 No. 3, pp. 735-762, doi: 10.1086/ahr/108.3.735.

Rüller, S., Aal, K., Tolmie, P., Hartmann, A., Rohde, M. and Wulf, V. (2022), "Speculative design as a collaborative practice: ameliorating the consequences of illiteracy through digital touch", *ACM Transactions on. Computer-Human Interaction*, Vol. 29 No. 3, pp. 1-58, doi: 10.1145/3487917.

Schantz, H.F. (1982), *History of OCR, Optical Character Recognition*, Manchester Center, Vermont, Recognition Technologies User Association.

Schomaker, L. (2019), "Lifelong learning for text retrieval and recognition in historical handwritten document collections", in Fischer, A., Liwicki, M. and Ingold, R. (Eds), *Handwritten Historical Document Analysis, Recognition and Retrieval – State of the Art and Future Trends*, World Scientific, London, pp. 221-248.

Secker, J. and Morrison, C. (2016), *Copyright and E-Learning: A Guide for Practitioners*, 2nd ed., facet publishing, London.

Seely Brown, J. and Duguid, P. (2000), *The Social Life of Information*, Harvard Business School Press, Boston.

Skluzacek, T.J., Chard, K. and Foster, I. (2022), "Automated metadata extraction: challenges and opportunities", *2022 IEEE 18th International Conference on e-Science and Grid Computing*, Salt Lake City, 11-14 October, pp. 495-500, available at: https://ieeexplore.ieee.org/document/9973723

Smith, J. (2021), "Blog post, Palladium: appraisal and sensitivity review of the Carcanet email archive", John Rylands Research Institute and Library, available at: https://rylandscollections.com/2021/05/28/palladium-appraisal-and-sensitivity-review-of-the-carcanet-email-archive/

Stauder, A. (2022), "Recording. The next generation of Transkribus", *4th Transkribus User Conference*, Innsbruck, 29-30 September, available at: https://www.youtube.com/watch?v=bv9Gie-hd88&li

Stauder, A. (2023), "Invitation: ChatGPT and transkribus - members meeting", Personal communication, READ-COOP members list, 15 May, 2023, 08:22:11 GMT.

Stoler, A.L. (2002), "Colonial archives and the Arts of governance", *Archival Science*, Vol. 2 No. 5, pp. 87-109, doi: 10.1007/BF02435632.

Stoler, A.L. (2008), *Along the Archival Grain*, Princeton University Press, Princeton, NJ.

Strauss, T., Weidemann, M. and Labahn, R. (2017), "D7.11 Language Models - improving transcriptions by external language resource", Innsbruck: Recognition and Enrichment of Archival Documents (READ), available at: https://readcoop.eu/wp-content/uploads/2017/12/D7.11_final.pdf

Strubell, E., Ganesh, A. and McCallum, A. (2019), "Energy and policy considerations for deep learning in NLP", *57th Annual Meeting of the Association for Computational Linguistics (ACL)*, July, Florence, available at: https://arxiv.org/abs/1906.02243

Saving Ukrainian Cultural Heritage Online (SUCHO) (2023), available at: www.sucho.org

Tanner, S., Muñoz, T. and Hemy Ros, P. (2009), "Measuring mass text digitization quality and usefulness. Lessons learned from assessing the OCR accuracy of the British library's 19th century online newspaper archive", *D-Lib Magazine*, Vol. 15 Nos 7-8, doi: 10.1045/july2009-munoz.

Terras, M. (2022), "Inviting AI into the archives: the reception of handwritten recognition technology into historical manuscript transcription", in Jaillaint, L. (Ed.), *Archives, Access and Artificial Intelligence: Working with Born-Digital and Digitized Archival Collections*, Bielefeld University Press, Bielefeld, pp. 179-204.

Terras, M., Nockels, J., Gooding, P., Muehlberger, G. and Stauder, A. (Forthcoming 2024), "On automating standardised editions: the affordances of handwritten text recognition platforms for scholarly editing", *Scholarly Editing*, pp. 1-38.

Thomas, W.G. III (2004), "Computing and the historical imagination", in Schreibman, S., Siemens, R. and Unsworth, J. (Eds), *A Companion to the Digital Humanities*, Wiley & Sons, New York, pp. 56-68.

Thylstrup, N.B. (2019), *The Politics of Mass Digitization*, MIT Press, Cambridge.

Toiganbayeva, N., Kasem, M., Abdimanap, G., Bostanbekov, K., Abdallah, A., Alimova, A. and Nurseitov, D. (2022), "KOHTD: Kazakh offline handwritten text dataset. Signal processing", *Image Communication*, Vol. 108, pp. 1-28, doi: 10.48550/arXiv.2110.04075.

Tschan, R. (2002), "A comparison of Jenkinson and Schellenberg on appraisal", *The American Archivist*, Vol. 65 No. 2, pp. 176-195, doi: 10.17723/aarc.65.2.920w65g3217706l1, available at: https://www.jstor.org/stable/40294205

Turkel, W.J., Kee, K. and Roberts, S. (2012), "A method for navigating the infinite archive", in Weller, T. (Ed.), *History in the Digital Age*, Routledge, London, pp. 57-72.

Turner, H. (2020), *Cataloguing Culture: Legacies of Colonialism in Museum Documentation*, University of British Columbia, Vancouver.

Underwood, T. (2017), "A genealogy of distant reading", *Digital Humanities Quarterly*, Vol. 11 No. 2, pp. 1-43, available at: http://www.digitalhumanities.org/dhq/vol/11/2/000317/000317.html

Unsworth, J. and Tupman, C. (2016), "Interview with John Unsworth, April 2011, carried out and transcribed by Charlotte Tupman", in Deegan, M. and McCarty, W. (Eds), *Collaborative Research in the Digital Humanities*, Routledge, London, pp. 231-240.

Valy, D., Verleysen, M. and Chhun, S. (2020), "Data augmentation and text recognition on Khmer historical manuscripts", *17th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, Dortmund, 7-10 September, 2020, doi: 10.1109/ICFHR2020.2020.00024\.

Van Lange, M. (2023), *Emotion Imprints of War: A Computer Assisted Analysis of Emotions in Dutch Parliamentary Debates, 1945-1989*, Bielefeld University Press, Bielefeld.

Vu, M.T., Le, V.L. and Beurton-Aimar, M. (2021), "IHR-NomDB: the old degraded Vietnamese handwritten script archive database", in Elisa, B., Wen, G., Steffan, B. and Yong, M. (Eds), *Document Analysis and Recognition - ICDAR 2021, Lecture Notes in Computer Science*, Springer International Publishing, Cham, pp. 85-99.

Whitelaw, M. (2015), "Generous interfaces for digital cultural collections", *Digital Humanities Quarterly*, Vol. 9 No. 1, pp. 1-16, available at: https://www.digitalhumanities.org/dhq/vol/9/1/000205/000205.html

Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L.B., Bourne, P.E., Bouwman, J., Brookes, A.J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C.T., Finkers, R., Gonzalez-Beltran, A., Gray, A.J.G., Groth, P., Goble, C., Grethe, J.S., Heringa, J., 't Hoen, P.A.C., Hooft, R., Kuhn, T., Kok, R., Kok, J., Lusher, S.J., Martone, M.E., Mons, A., Packer, A.L., Persson, B., Rocca-Serra, P., Roos, M., van Schaik, R., Sansone, S.-A., Schultes, E., Sengstag, T., Slater, T., Strawn, G., Swertz, M.A., Thompson, M., van der Lei, J., van Mulligen, E., Velterop, J., Waagmeester, A., Wittenburg, P., Wolstencroft, K., Zhao, J. and Mons, B. (2016), "The FAIR Guiding Principles for scientific data management and stewardship", *Scientific Data*, Vol. 3 No. 1, pp. 1-9, doi: 10.1038/sdata.2016.18.

Yale, E. (2015), "The history of archives: the state of the discipline", *Book History*, Vol. 18 No. 15, pp. 332-359, doi: 10.1353/bh.2015.0007.

Zaagsma, G. (2019), "Digital history and the politics of digitization", *Digital Scholarship in the Humanities*, Vol. 38 No. 2, pp. 830-851, doi: 10.1093/llc/fqac050/6702047.

Further reading

GitHub (2023), "Text recognition for zooniverse", available at: https://github.com/danhan52/text_recognition

Corresponding author

Joseph Nockels can be contacted at: j.h.nockels@sms.ed.ac.uk