THE UNIVERSITY of EDINBURGH

Edinburgh Research Explorer

The role of the library when computers can read

Citation for published version:
Terras, M 2022, The role of the library when computers can read: Critically adopting Handwritten Text Recognition (HTR) technologies to support research. in A Wheatley & S Hervieux (eds), *The Rise of AI: Implications and Applications of Artificial Intelligence in Academic Libraries.* ACRL - Association of College & Research Libraries, Atlanta, pp. 137-148.

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Publisher's PDF, also known as Version of record

Published In:
The Rise of AI

General rights
Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact openaccess@ed.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.

OPEN ACCESS

Download date: 20. Apr. 2024

Chapter 11

The Role of the Library When Computers Can Read:

Critically Adopting Handwritten Text Recognition (HTR) Technologies to Support Research

Melissa Terras

Introduction

Computational approaches to processing and searching images of historical manuscripts by handwritten text recognition (HTR) is one of the most promising machine learning approaches for academic research in the humanities, having the potential to transform access to our written past for the use of researchers, institutions, and the general public. This chapter surveys the current use of HTR in the library sector, highlighting major

CC BY

tools currently in use and activities being undertaken by academic and research libraries. Using Transkribus as a case study, the chapter provides examples of where libraries have successfully deployed HTR and focuses on emerging issues for incorporating the application and results of HTR into a digitisation workflow, including documentation, results delivery, and sustainability. The chapter considers how HTR can be best deployed to support researchers, including the need for transparency, training, and data infrastructure. Although HTR technology is now reasonably mature, academic libraries need to adopt this machine learning (ML) technique in a critical way, signposting the data in a way that explains its creation and allows its embedding into historical practice to best support their user communities.

Computationally Reading Documents: OCR, HTR, and Mass Digitisation

Libraries and archives have invested heavily in mass digitisation of their print and manuscript collections over the past thirty years.[1] The textual content of the resulting digital images, however, was only recently available to those with the resources to manually transcribe individual passages.[2] Manual transcription is an approach that does not scale across larger collections of documents given the costs associated with employing researchers or the setup and monitoring costs associated with working with volunteers (although "crowdsourcing" such volunteer labour using online mechanisms can be cost-effective in the long term).[3] It has long been an aim of computer scientists, librarians, archivists, and curators to be able to generate accurate machine-readable transcriptions of their holdings, with the accepted understanding that doing so will enable the "key functional elements of large databases of print—easily readable texts and full-text searching."[4] Including handwritten material "promises to yet again extend and revolutionize the study of historical handwritten documents,"[5] allowing both searching at a scale impossible to the offline human reader and the use of advanced mining, analysis, and visualisation techniques.[6] This brings with it the possibility of providing new untold insights to benefit the research community, utilising (and democratising access to) the vast volume of digital images of manuscripts now produced by the heritage sector.

Both optical character recognition (OCR, the conversion of images of typed or printed text into machine-encoded format) and handwritten text recognition (HTR, the use of computational technologies to interpret text from handwritten sources and to produce it in machine-encoded format) have long histories, stretching back as early as the nineteenth century.[7] OCR is now a standard approach, routinely embedded within digitisation workflows and digital library programs,[8] with the resulting datasets allowing for searching across massive repositories of digitised text. There are various OCR solutions available (which themselves use AI), including those which are commercial (ABBYY FineReader, Kofax OmniPage Standard, Adobe Acrobat DC)[9] and open source (Tesseract, now owned

by Google).[10] Issues with accuracy, however, remain, particularly with formats that deviate from clearly printed text, such as the inky spread of newsprint or more complex fonts and page presentation, and that can affect resulting further analysis of the material.[11] Recent developments in HTR, involving both machine and deep-learning approaches, mean it is now possible to generate machine-processable text directly from digitised images of handwritten (or complex print) material.[12] Improved accuracy rates have increased discoverability and the potential to undertake new and novel research at scale.[13] As a result, HTR is one of the few applications of artificial intelligence in the cultural and heritage sector that has reached relative maturity and is now being applied by digitisation units and scholarly projects across the academic library sector. Use is far from standardised and there has been next to no research on how best practices can be undertaken in storing, sharing, and explaining HTR-generated content in this field.

The Current HTR Landscape for Libraries

Libraries are at a point where they must choose if they want to trial this potentially useful technology within their heritage digitisation workflows. They can work in tandem with computer scientists and generate their own bespoke HTR approaches for the material in question (there being a long history of this research-led approach to reading historical texts),[14] or they can reuse the outputs of these projects. For example, the In Codice Ratio project is developing "tools to support content analysis and knowledge discovery from large collections of historical documents," concentrating on the collections of the Vatican Secret Archives.[15] Using a convolutional neural network classifier* and statistical language models to generate the most likely transcript, it also engages palaeographers in crowd-sourcing training data. Likewise, the Monk system has been developed by the University of Groningen using AI methods for accessing historical archive collections that are difficult to process by traditional OCR methods—for example, due to their historical character types or due to the fact that the material is handwritten. The system consists of two major components: (1) a setup for the storage and web-based annotation of scanned page images and parts thereof, and (2) a set of (handwriting and text) recognition algorithms as well as retrieval and search methods.[16]

These systems work with a variety of human languages and across temporalities: the PYTHIA project is the first ancient text restoration model that recovers missing characters from damaged Greek inscriptions using deep neural networks;[17] and The Center for Open Data in the Humanities in Tokyo is developing approaches and datasets for training and reading different types of Japanese script, such as deep-learning techniques

* A convolutional neural network, or CNN, is a particular type of deep neural network (an artificial neural network with multiple layers between input and output) that is mostly deployed in the analysis of digital images. Inspired by the design of the animal visual cortex, they provide an efficient pattern recognition approach. See Saad Albawi, Mohammed Tareq Abed, and Saad Al-Zawi, *Understanding of a convolutional neural network*, 2017.

to be used in conjunction with the Kuzushiji Dataset of Pre-Modern Japanese Text.[18] The projects mentioned here are only an indicative few, which are part of a large and expanding community of computational researchers developing their own HTR solutions as well as publishing their code and results for others to evaluate and reuse with emerging benchmarks and best-practice guidelines.[19] Establishing such partnerships, however, requires resources, expertise, and confidence in the sustainability of chosen systems. Libraries should have a robust plan as to how they will manage and maintain both the code and the resulting datasets if these in-house solutions are adopted and applied to mass-digitised content.

Libraries can turn to the solutions being offered by publishers and established technology companies, although the processes used to generate HTR can be opaque. For example, Adam Matthew Digital "is the first publisher to utilise artificial intelligence to offer handwritten text recognition (HTR) for its handwritten manuscript collections" at time of writing, offering full-text search of transcriptions of seven major archival collections that have been processed with their in-house HTR.[20] Institutions can now license access to their Quartex document management system, which is the "only platform with built-in HTR, making manuscripts searchable."[21] Likewise, Gale is offering full-text searching for two of its collections processed with HTR, although not providing access to a platform for others to upload their content.[22] Google Arts and Culture recently launched Fabricius, using AI to help translate hieroglyphs, and they are also continuing to expand API for developers wishing to detect handwriting in images.[23] This potential relationship with corporate publishers and technological giants is therefore one to closely watch as they continue to develop tools to apply HTR to mass-digitised content. As with relationships with all digital commercial entities, care should be given to copyright; image licensing; mechanisms for storing, sharing, and long-term archiving of both input and output data; and access to explanations of the algorithms involved to better understand how the data is processed.

Transkribus as an HTR Solution for Libraries

Between 2011 and 2019, a large consortium of EU-funded researchers led by the University of Innsbruck developed a machine-learning approach for automatic generation of transcripts from digitised images of historical handwritten text, which resulted in software—Transkribus*—which is now capable of generating transcriptions with up to 98 percent accuracy.[24] At the time of writing, Transkribus has forty-one thousand registered users, including individuals and major libraries, archives, and museums worldwide. Over one thousand users actively use the software each week, with more than a million images uploaded every month for processing and over six thousand HTR models now generated in total by the community. From July 1, 2019, the platform has been operated and further

* See: https://transkribus.eu/Transkribus/.

developed by the European cooperative READ-COOP, a mechanism planned to sustain and grow the Transkribus infrastructure beyond the end of its grant-financed period built around a cooperative economic model.[25]

Transkribus uses deep neural network machine learning technology. Once images of manuscripts are uploaded, layout analysis tools segment them into lines (the Transkribus graphical user interface (GUI) contains both automatic and manual segmentation tools, allowing user correction), before each line is transcribed. A training process—on approximately fifteen thousand transcribed words or seventy-five pages of handwritten script—generates what is known as an HTR *model* for recognising text written in one hand. Users can either apply a model created (and sometimes openly published) by another project or train up their own, which is then used to generate transcriptions.[26] In the best cases, HTR can produce automated transcriptions of handwritten material with a character error rate (CER) of below 5 percent, meaning 95 percent of characters are correct, and if used on printed material, that CER can reach 1 to 2 percent. Users can then work with the Transkribus GUI to correct and improve the generated transcripts to compile new, improved, models; this creates a feedback loop that improves the efficacy of the underlying neural network, benefiting future users. Once HTR has been completed, the user is free to take the resulting transcriptions and use them in any way they feel appropriate, such as inclusion in digital scholarly editions, used as source data for further linguistic or semantic analysis, or ingested into digital library content management systems as a finding aid used in conjunction with keyword searching.[27]

An early user of Transkribus was the Bentham Project at University College London, which trained models on the writing of the philosopher and jurist Jeremy Bentham (1748–1832) in order to transcribe his vast personal archive. The initial models were trained on crowdsourced transcriptions of Bentham's manuscript material.[28] With forty-one thousand users in at least fifty-three countries worldwide, there are a broad variety of institutions that have adopted the platform for a range of complex projects.[29] For example, transcribing a large set of Michel Foucault's reading notes, including citations, references, and comments; increasing the accessibility and usability of the archive of the cloister of the Poor Clares St.-Elisabethsdal in Boxtel (1390–1719); and retro-digitizing and automatically structuring the large bibliography collection of the *Internationale Bibliographie der Lexikographie* by Herbert Ernst Wiegand.[30] These examples indicate the range of potential uses on digitised academic library collections while also presenting a resource-saving opportunity and removing barriers caused by not having a budget to employ staff to transcribe these archival manuscript materials themselves.

The Transkribus platform is now at a crucial moment: having built a working platform, the grant-funded period has now ended, and to maintain the infrastructure, an income stream must be created. The platform will be transferred to a paid-for model in late 2020 and, therefore, library projects will have to include a budget for its use in their plans. There are other criticisms of Transkribus, the main one vocalized being the fact that not all of its processes, algorithms, and models are published, and so it is not fully subscribing to the Open Science principles.[31] Via the READ-COOP, Transkribus continues to work with its growing user base in order to provide continued access to this tool for the wider cultural

heritage community. There are regular meetings and lively Facebook user forums (both official and unofficial) to assist when applying HTR via Transkribus to digitised content. It has rapidly become the most recommended tool for HTR (and AI, by extension) within the cultural heritage sector.

Critically Embedding HTR Into Libraries

Although there are considerable savings in time and resources in using HTR to generate transcriptions of historical manuscripts, HTR is not a panacea. If it is to be successfully used to increase access to information within and usability of handwritten textual material, it needs to be embedded into both digitisation workflows within libraries and other institutions as well as public-facing digital library infrastructures. However, there is little consideration to date of how HTR can be built into service-level provision of digitisation within academic libraries that adopt it or the type of messaging and communication that will be necessary to allow users to embed HTR-generated data into their research practice.

HTR's use is dependent on the availability of digitised content, and that itself has long been known to be a costly and complex endeavour, which often has ethical and legal implications, including the necessary navigation of copyright and related permissions.[32] HTR has the potential to extend the diversity of materials available to researchers, but only if libraries engage with them at the digitisation phase, ensuring a plurality of voices and that archives are entering the digitisation pipeline, which has not been the case to date.[33]

While there are efficiencies in employing HTR to generate transcripts of images of handwritten text, rather than employing researchers to do so, the training of models and generating of transcripts by HTR is not fully automated. For optimal results, there needs to be full engagement with the feedback loops built into the process; therefore, it does still take resources to deploy HTR successfully. As well as finding resources, libraries need to plan ahead regarding how HTR can become embedded into existing digitisation and information workflows. It is essential that projects establish a data-management plan for the resulting HTR-generated transcripts. Integrating HTR into digitisation processes (managed by platforms such as Goobi)[34] and embedding the results of HTR within content management systems (CMS) are essential steps to reap the benefits of this technology and promote the discoverability of the results, thus ensuring that the data is preserved and following normal institutional digital-curation practices. While there are various approaches to handling OCR-generated data that can be adopted, there are not yet any *de facto* standards for even the basics in HTR data handling, including industry file naming standards for this content, never mind more sophisticated workflows, such as persistent identifier generation, links to other infrastructural frameworks such as IIIF,* or standards

* The International Interoperability Framework (IIIF) aims to define a set of common standards and application programming interfaces that support interoperability between image repositories, promote best practice in this area, and become more adopted in the GLAM sector; see https://iiif.io.

for openly publishing datasets (although projects such as EScripta are starting to explore these opportunities, and should be watched).[35]

In addition, most user-facing CMSs are not yet configured to allow this additional computer-generated textual information to sit alongside high-resolution images and to enable discoverability by full-text searching. It is also unclear where, in now-standard library and archival metadata structures, these AI-generated transcripts should be stored or if they are covered by non-print legal deposit regulations for mandated ingestion into national digital library repositories. It is not known how best to manage the reporting of any errors, in order to correct and improve HTR-generated content, in a way that is both transparent and scalable (although there are parallels with mass OCR correction).[36] Libraries, therefore, need to carefully consider both the ingestion, processing, and output of HTR processes to ensure that it is employed in a useful and sustainable manner alongside existing infrastructure.

In addition, there are issues regarding information literacy. There is a need to highlight to users of library and archival systems which information fields have been generated (and checked) by trained professionals in the sector, which have been automatically created by algorithms, and which may be crowdsourced from other users as part of the feedback process. Researchers need to be made aware of how datasets are created from digitisation to dataset, and there has been little work on the transparency of HTR tools and techniques, a criticism that can be levelled at the major technological providers (including Transkribus). Academic libraries must address these issues in a clearly explained manner in order to support their research users adequately. There has been next to no research on the users of HTR-generated historical content and the implications this technology has for scholarly work. If we are to see HTR-generated datasets used as new source material to underpin novel research, we must be able to explain their provenance so that the resulting datasets can be trusted as a scholarly source. Libraries will have to provide training and support for researchers to understand appropriate methodologies to get the best results from the emerging, relatively large-scale datasets and to provide new infrastructures to host and support these new "collections as data."[37] The claims made that HTR has the potential to revolutionise scholarship should be viewed with interest by libraries. They will have a supporting function in the creation, hosting, and delivery of such large-scale transcripts and in supporting the user community to make the most of the opportunities therein.

Beyond scholarly applications, there are wider environmental issues that need to be addressed with the use of high-performance computing. It can take several days for a server farm to train an HTR model, which has implications for energy usage and carbon emissions. Many institutions are starting to think holistically about their use of such resources and mitigate against them with schemes such as carbon-offsetting, which will need to be considered and factored into project plans.[38]

As a result, HTR may be a maturing technology from an algorithmic perspective, but procedurally, libraries have to adopt it critically to help these issues settle into best practices for the sector. Given the newness of the approach and the tools involved, we are in a parallel position to where libraries were in the 1990s when digitisation at scale first became affordable, possible, and practical, and where the reports of individual libraries

on their attempts to operationalise this approach were essential to establishing sector best practices. At this stage in the HTR journey, it behooves projects and institutions adopting this technology to consciously report on the decisions, practices, and protocols that are necessary within the digitisation and usability pipeline. This will allow others to learn from previous implementations to help the sector work toward a user-centric best practice of this promising AI technology and to aid research into the human notes of the past from a conscious, informed position.

Conclusion

Academic libraries have an important role to play now that computers can read handwritten text in selecting diverse material to be digitised and supporting a plurality of voices in our digital cultural heritage landscape. Libraries must decide where to apply HTR to generate searchable and processable outputs from these sources; establish best practices to ensure that these data sources are sustainable, findable, and useable; and support the research community in accessing, analysing, and reporting on their content. In order to make the most of the opportunities recently presented by HTR technology, those in the cultural heritage sector using it (however experimentally) should report, communicate, and discuss both methods and results with others similarly interested to contribute to a convergence of approaches which will, eventually, become standardised best practice. It is essential that this new machine-learning technology be harnessed by those wishing to increase access to content held within manuscript material that will be of interest to researchers. Doing so means a level of understanding, engagement, and control to establish where HTR can contribute to the work of academic researchers in their understanding of the past and to make processes transparent and understandable. Libraries are expertly placed to be the nexus that can support and encourage the novel research questions, approaches, and, ultimately, outputs that give handwritten text recognition its transformational potential. As this approach continues to develop, libraries can also frame and explain how the use of this technology may change and expand existing historical scholarly practice given certain ventures into these new AI-generated vistas.

Conflict of interest statement: Melissa Terras is on the Board of Directors of the READ COOP. She does not and will not benefit financially from this, either now or in the future, due to the cooperative financial structures in place.

Endnotes

1. Melissa Terras, "The Rise of Digitization: An Overview," in *Digitisation Perspectives*, ed. Ruth Rikowski (Leiden, Netherlands: Brill, 2011), 1–20; Natasha Stroeker and René Vogels, "Survey report on digitisation in European cultural heritage institutions 2012," Panteia, Netherlands, on behalf of the ENUMERATE Thematic Network, 2014.
2. John Edwards, "Easily adaptable handwriting recognition in historical manuscripts" (PhD Dissertation, University of California, Berkeley, CA, 2007).
3. Tim Causer, Kris Grint, Anna-Maria Sichani, and Melissa Terras, "'Making such bargain': Transcribe Bentham and the quality and cost-effectiveness of crowdsourced transcription," *Digital Scholarship in the Humanities* 33, no. 3 (2018): 467–87.

4. Laura Estill and Michelle Levy, "Chapter 12: Evaluating Digital Remediations of Women's Manuscripts," Digital Studies/Le champ numérique, Vol. 6, 2016, http://doi.org/10.16995/dscn.12.
5. Estill and Levy, "Chapter 12: Evaluating Digital Remediations."
6. Kristen Schuster and Stuart Dunn, *Routledge International Handbook of Research Methods in Digital Humanities* (Oxfordshire, UK: Routledge, 2020).
7. Herbert Schantz, *The History of OCR, Optical Character Recognition*, Manchester Center, VT: Recognition Technologies Users Association, 1982.
8. Karmen Sotošek, "Best practice examples in library digitisation," Europeana Travel, D2.2, 011, https://pro-beta.europeana.eu/files/Europeana_Professional/Projects/Project_list/EuropeanaTravel/Deliverables/D2.2%20Best%20practice%20examples%20in%20library%20digitisation.pdf.
9. ABBYY FineReader, 2020, https://pdf.abbyy.com/; Kofax Omnipage Standard, 2020, https://www.kofax.com/Products/omnipage/standard; Adobe Acrobat DC, Adobe, 2020, https://acrobat.adobe.com/uk/en/acrobat/pricing.html.
10. Tesseract-OCR, 2020, https://github.com/tesseract-ocr.
11. Simon Tanner, Trevor Muñoz, and Pich Hemy Ros, "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* 15, no. 7/8 (2009), http://www.dlib.org/dlib/july09/munoz/07munoz.html; Ryan Cordell, "'Q i-jtb the Raven': Taking Dirty OCR Seriously," *Book History* 20, No. 1 (2017): 188–225.
12. Byron Bezerra et al., *Handwriting: Recognition, Development and Analysis* (New York: Nova Science Publishers, 2017).
13. Estill and Levy, "Chapter 12: Evaluating Digital Remediations."
14. Melissa M. Terras, *Image to Interpretation: An Intelligent System to Aid Historians in Reading the Vindolanda Texts* (Oxford University Press, 2006).
15. Donatella Firmani et al., "Towards Knowledge Discovery from the Vatican Secret Archives. In Codice Ratio-Episode 1: Machine Transcription of the Manuscripts," in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2018, 263–72.
16. Monk, *Lambert Schomaker, Lifelong learning for text retrieval and recognition in historical handwritten document collections*, 2019, https://www.ai.rug.nl/~lambert/Monk-collections-english.html.
17. Assael Yannis, Thea Sommerschield, and Jonathan Prag, "Restoring ancient text using deep learning: a case study on Greek epigraphy," arXiv preprint arXiv:1910.06262, 2019.
18. Tarin Clanuwat, Alex Lamb, and Asanobu Kitamoto, "Kuronet: Pre-modern Japanese kuzushiji character recognition with deep learning," in *2019 International Conference on Document Analysis and Recognition (ICDAR)*, 2019, 607–14.
19. Joan Andreu Sánchez, Verónica Romero, Alejandro H. Toselli, Mauricio Villegas, and Enrique Vidal, "A set of benchmarks for handwritten text recognition on historical documents," *Pattern Recognition* 94 (2019): 122–34.
20. "Artificial intelligence transforms discoverability of handwritten manuscripts," Adam Matthew Digital, 2020, https://www.amdigital.co.uk/products/handwritten-text-recognition.
21. "Features," Quartex, 2020, https://www.quartexcollections.com.
22. "An essential primary source archive for researching the history of Hong Kong in the context of Modern China and the British Empire in Asia," Gale, 2020, https://www.gale.com/intl/c/china-and-the-modern-world-hongkong-britain-china.
23. "What is Fabricius?," Google Arts and Culture, 2020; "Detect handwriting in images," Google Cloud, 2020.
24. Guenter Muehlberger et al., "Transforming scholarship in the archives through handwritten text recognition: Transkribus as a case study," *Journal of Documentation* (2019), https://doi.org/10.1108/JD-07-2018-0114.
25. READ-COOP, "Revolutionizing Access to Handwritten Documents," 2020, https://readcoop.eu.
26. Philip Kahle, Sebastian Colutto, Günter Hackl, and Günter Mühlberger, "Transkribus-A service platform for transcription, recognition and retrieval of historical documents," in *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, vol. 4 (2017), 19–24.
27. Muehlberger et al., "Transforming scholarship."

28. Tim Causer and Melissa Terras, "'Many hands make light work. Many hands together make merry work': Transcribe Bentham and Crowdsourcing manuscript collections," in *Crowdsourcing Our Cultural Heritage*, ed. Mia Ridge (Farnham, UK: Ashgate Publishing, 2014), 57–88.
29. Melissa Terras, Tweet, 2019.
30. Vincent Ventresque, Arianna Sforzini, and Marie-Laure Massot, "Transcribing Foucault's handwriting with Transkribus," *Journal of Data Mining & Digital Humanities* (2019), https://jdmdh.episciences.org/5218/pdf; Geertrui Van Synghel, "'Sine clausura': Unlocking the archive of the cloister of the Poor Clares St.-Elisabethsdal in Boxtel (1390-1719)," *Franciscan Studies* 77, no. 1 (2019): 89–110; David Lindemann, Mohamed Khemakhem, and Laurent Romary, "Retro-digitizing and Automatically Structuring a Large Bibliography Collection," European Association for Digital Humanities (EADH) Conference, EADH, December 2018, Galway, Ireland, https://hal.archives-ouvertes.fr/hal-01941534/document.
31. "What is Open Science? Introduction," Foster, 2020, https://www.fosteropenscience.eu/node/1420.
32. Lorna Hughes, *Digitizing Collections: Strategic Issues for the Information Manager* (London: Facet Publishing, 2004).
33. Nanna Thylstrup, *The Politics of Mass Digitization* (Cambridge: MIT Press, 2019).
34. "Goobi for the workflow," Intranda, 2020, https://www.intranda.com/en/digiverso/goobi/workflow/.
35. Peter Stokes, Benjamin Kiessling, Robin, Tissot, Daniel Stökl, and Ben Ezra, "eScripta: A New Digital Platform for the Study of Historical Texts and Writing," Digital Humanities 2019, Utrecht University, July 2019, https://dev.clariah.nl/files/dh2019/boa/0322.html.
36. Rose Holley, "Many hands make light work: Public collaborative OCR text correction in Australian historic newspapers," National Library of Australia, 2009, https://www.nla.gov.au/content/many-hands-make-light-work-public-collaborative-ocr-text-correction-in-australian-historic.
37. Thomas Padilla, Laurie Allen, and Hannah Frost, "Always Already Computational: Collections as Data," Zenodo, 2019, http://doi.org/10.5281/zenodo.3152935.
38. Loïc Lannelongue, Jason Grealey, and Michael Inouye, "Green Algorithms: Quantifying the carbon emissions of computation," arXiv preprint arXiv:2007.07610, 2020.

Bibliography

ABBYY FineReader PDF. 2020. https://pdf.abbyy.com/.

Adam Matthew Digital. "Artificial intelligence transforms discoverability of handwritten manuscripts." 2020. https://www.amdigital.co.uk/products/handwritten-text-recognition.

Assael, Yannis, Thea Sommerschield, and Jonathan Prag. "Restoring ancient text using deep learning: a case study on Greek epigraphy." arXiv preprint arXiv:1910.06262. 2019.

Adobe. "Adobe Acrobat DC Plans and Pricing." 2020. https://acrobat.adobe.com/uk/en/acrobat/pricing.html.

Albawi, Saad, Tareq Abed Mohammed, and Saad Al-Zawi. "Understanding of a convolutional neural network." In *2017 International Conference on Engineering and Technology (ICET)* (2017): 1–6.

Bezerra, Byron, et al., *Handwriting: Recognition, Development and Analysis*. New York: Nova Science Publishers, 2017.

Causer, Tim, and Melissa Terras. "'Many hands make light work. Many hands together make merry work': Transcribe Bentham and Crowdsourcing manuscript collections." In *Crowdsourcing Our Cultural Heritage*, edited by Mia Ridge. Farnham, UK: Ashgate, 2014, 57–88.

Causer, Tim, Kris Grint, Anna-Maria Sichani, and Melissa Terras. "'Making such bargain': Transcribe Bentham and the quality and cost-effectiveness of crowdsourced transcription." *Digital Scholarship in the Humanities* 33, no. 3 (2018): 467–87.

Clanuwat, Tarin, Alex Lamb, and Asanobu Kitamoto. "Kuronet: Pre-modern Japanese kuzushiji character recognition with deep learning." In *2019 International Conference on Document Analysis and Recognition (ICDAR)*. IEEE (2019), 607–14.

Cordell, Ryan. "'Q i-jtb the Raven': Taking Dirty OCR Seriously." *Book History* 20, No. 1 (2017): 188–225.

Edwards, John Alexander. "Easily adaptable handwriting recognition in historical manuscripts." PhD Dissertation, University of California, Berkeley, CA, 2007. http://www2.eecs.berkeley.edu/Pubs/TechRpts/2007/EECS-2007-76.pdf.

Estill, Laura, and Michelle Levy. "Chapter 12: Evaluating digital remediations of women's manuscripts." Digital Studies/Le champ numérique, Vol. 6. 2016. http://doi.org/10.16995/dscn.12.

Firmani, Donatella, Marco Maiorino, Paolo Merialdo, and Elena Nieddu. "Towards Knowledge Discovery from the Vatican Secret Archives. In Codice Ratio-Episode 1: Machine Transcription of the Manuscripts." In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (2018), 263–72.

Foster, "What is Open Science? Introduction." 2020. https://www.fosteropenscience.eu/node/1420.

Gale. "An essential primary source archive for researching the history of Hong Kong in the context of Modern China and the British Empire in Asia." 2020. https://www.gale.com/intl/c/china-and-the-modern-world-hongkong-britain-china.

Google Arts and Culture. "What is Fabricius?" 2020. https://artsexperiments.withgoogle.com/fabricius/en.

Google Cloud. "Detect handwriting in images." AI and Machine Learning Products. 2020. https://cloud.google.com/vision/docs/handwriting.

Holley, Rose. "Many hands make light work: Public collaborative OCR text correction in Australian historic newspapers." National Library of Australia. 2009. https://www.nla.gov.au/content/many-hands-make-light-work-public-collaborative-ocr-text-correction-in-australian-historic.

Hughes, Lorna M. *Digitizing Collections: Strategic Issues for the Information Manager*. London: Facet Publishing, 2004.

Intranda. "Goobi for the workflow." 2020. https://www.intranda.com/en/digiverso/goobi/workflow/.

Kahle, Philip, Sebastian Colutto, Günter Hackl, and Günter Mühlberger. "Transkribus—A service platform for transcription, recognition and retrieval of historical documents." In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, vol. 4, IEEE (2017), 19–24.

Kofax Ominpage Standard. 2020. https://www.kofax.com/Products/omnipage/standard.

Lindemann, David, Mohamed Khemakhem, and Laurent Romary. "Retro-digitizing and Automatically Structuring a Large Bibliography Collection." European Association for Digital Humanities (EADH) Conference, EADH, December 2018, Galway, Ireland. https://hal.archives-ouvertes.fr/hal-01941534/document.

Lannelongue, Loïc, Jason Grealey, and Michael Inouye. "Green Algorithms: Quantifying the carbon emissions of computation." arXiv preprint arXiv:2007.07610, 2020.

Monk. "Indexed books and overview." https://www.ai.rug.nl/~lambert/Monk-collections-english.html.

Muehlberger, Guenter, Louise Seaward, Melissa Terras, et al. "Transforming scholarship in the archives through handwritten text recognition: Transkribus as a case study." *Journal of Documentation* (2019). https://doi.org/10.1108/JD-07-2018-0114.

Padilla, Thomas, Laurie Allen, Hannah Frost, et al. "Always Already Computational: Collections as Data." Zenodo. 2019. http://doi.org/10.5281/zenodo.3152935.

Quartex, "Features." 2020. https://www.quartexcollections.com.

READ-COOP. "Revolutionizing Access to Handwritten Documents." 2020. https://readcoop.eu.

Sánchez, Joan Andreu, Verónica Romero, Alejandro H. Toselli, Mauricio Villegas, and Enrique Vidal. "A set of benchmarks for handwritten text recognition on historical documents." *Pattern Recognition* 94 (2019): 122–34.

Schantz, Herbert H. "The history of OCR, optical character recognition." Manchester Center, VT: Recognition Technologies Users Association. 1982.

Schomaker, Lambert. "Lifelong learning for text retrieval and recognition in historical handwritten document collections." Forthcoming in *Handwritten Historical Document Analysis, Recognition, and Retrieval, State of the Art and Future Trends*. arXiv preprint arXiv:1912.05156. 2019.

Schuster, Kristen, and Stuart Dunn, eds. *Routledge International Handbook of Research Methods in Digital Humanities*. Oxfordshire, UK: Routledge, 2020.

Stroeker, Natasha, and René Vogels. "Survey report on digitisation in European cultural heritage institutions 2012." Panteia, Netherlands. On behalf of the ENUMERATE Thematic Network, 2014.

Sotošek, Karmen Štular. "Best practice examples in library digitisation." Europeana Travel, D2.2. 2011. https://pro-beta.europeana.eu/files/Europeana_Professional/Projects/Project_list/EuropeanaTravel/Deliverables/D2.2%20Best%20practice%20examples%20in%20library%20digitisation.pdf.

Stokes, Peter, Benjamin Kiessling, Robin, Tissot, Daniel Stökl, and Ben Ezra. "eScripta: A New Digital Platform for the Study of Historical Texts and Writing." Digital Humanities 2019, Utrecht University. July 2019. https://dev.clariah.nl/files/dh2019/boa/0322.html.

Tanner, Simon, Trevor Muñoz, and Pich Hemy Ros. "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* 15, no. 7/8 (2009). http://www.dlib.org/dlib/july09/munoz/07munoz.html.

Terras, Melissa M. "The Rise of Digitization: An Overview," in *Digitisation Perspectives*, ed. Ruth Rikowski. Leiden, Netherlands: Brill, 2011, 1–20.

———. *Image to Interpretation: An Intelligent System to Aid Historians in Reading the Vindolanda Texts*. Oxford University Press, 2006.

———. Tweet: "So @Transkribus now has 22k users. 3k of those logged in over the past 3 months, in 31k sessions. Altho the location of 37% of those are indeterminable (gmail, hotmail, etc), I plotted the rest: according to their domains, we have active users in 53 countries worldwide!" July 10, 2019. https://twitter.com/melissaterras/status/1148859050735624192/photo/3.

Tesseract-OCR. 2020. https://github.com/tesseract-ocr.

Thylstrup, Nanna Bonde. *The Politics of Mass Digitization*. Cambridge: MIT Press, 2019.

Transkribus. 2020. https://transkribus.eu/Transkribus/.

Van Synghel, Geertrui. "'Sine clausura': Unlocking the archive of the cloister of the Poor Clares St.-Elisabethsdal in Boxtel (1390–1719)." *Franciscan Studies* 77, no. 1 (2019), 89–110.

Ventresque, Vincent, Arianna Sforzini, and Marie-Laure Massot. "Transcribing Foucault's handwriting with Transkribus." *Journal of Data Mining & Digital Humanities* (2019). https://jdmdh.episciences.org/5218/pdf.