Inviting AI into the archives

Chapter 7: Inviting AI into the Archives: The Reception of Handwritten Recognition Technology into Historical Manuscript Transcription

Melissa Terras, University of Edinburgh

ABSTRACT

Archives and libraries are increasingly investing in mass-digitization, but until recently transcriptions of manuscripts were costly to generate. Handwritten Text Recognition (HTR) technology is now transforming access to our written past, producing increasingly accurate transcriptions for use by individuals and institutions, or providing material for further analysis. However, there has been little consideration of how this will affect archival and historical method. Considering users of the Transkribus platform as a community of practice, this chapter will report on a survey of users of HTR, undertaken as a reception study regarding the practical, methodological, theoretical, and ethical issues raised when inviting machine learning into historical archives. Evidencing Transkribus use by a diverse community, it is suggested that the scale and scope of transcriptions generated by HTR will require new approaches to both history and public engagement, while providing recommendations on how to best support the community applying HTR to cultural heritage materials.

INTRODUCTION

Libraries and archives are now routinely investing in the digitization of their manuscript and early print collections to support access, and research. However, until recent technological developments, the content of digital images of historical texts has only been available to those who have the resources available to employ researchers, or manage volunteers, to undertake page-by-page transcription. The use of machine learning based Handwritten Text Recognition (HTR) to search, process, and generate transcriptions from mass-digitized content is now transforming access to our written past at scale. This has significant implications for the accessibility of the written records of global cultural heritage, making content available for use by individual researchers, scholarly editing projects, institutions,

the general public, and also meaning that these collections are open to further computational analysis.

There has been little consideration of how HTR will affect archival and historical method. This chapter reports from the Recognition and Enrichment of Archival Documents (READ) European Union Horizon 2020 project (2015-2019)[1] which developed advanced HTR based on artificial neural networks. READ developed a publicly available infrastructure: Transkribus, currently the primary user-facing platform for applying HTR to digitized content. Institutional and individual users of Transkribus are able to apply HTR to extract data from handwritten and printed texts, while simultaneously contributing to the improvement of the platform via machine learning principles. This chapter reports on a survey of registered users of Transkribus, conducting a reception study[2] on how HTR has been adopted by the library, archive, scholarly and genealogical community, and reflecting on practical, methodological, theoretical, and ethical issues raised when inviting AI into historical practice.

HANDWRITTEN TEXT RECOGNITION AND MASS DIGITISATION

Libraries and archives have undertaken mass-digitization of their collections for over thirty years,[3] however in the resulting digital images "one of the key functional elements of large databases of print—easily readable texts and full-text search-

1 This research was funded as part of the Recognition and Enrichment of Archival Documents (READ) project. This project has received funding from the European Union's Horizon 2020 research and innovation programme under Grant Agreement No. 674943. This research was previously funded as part of the tranScriptorium project. This project received funding from the European Union's Seventh Framework Programme for research, technological development and demonstration under Grant Agreement No. 600707. Melissa Terras is on the Board of the READ-COOP, but does not financially benefit from this relationship. Thanks are extended to the wider Transkribus team: in particular Guenter Muehlberger and Louise Seaward gave input on survey design, and Andy Stauder and Florian Stauder commented on this chapter, We thank users of the Transkribus platform, especially those who responded to this survey: "This survey is too long" – thank you for contributing.

2 Pertti Alasuutari, Three Phases of Reception Studies, in: Pertti Alasuutari (ed.), *Rethinking the Media Audience, the New Agenda*, London 1999, 1-8.

3 Lorna Hughes, *Digitizing Collections: Strategic Issues for the Information Manager*, London 2004; Melissa Terras, The Rise of Digitization: An Overview, in: Ruth Rikowski (ed.), *Digitization Perspectives*, Leiden, Netherlands, 2011, 1-20; Natasha Stroeker/René Vogels, Survey Report on Digitisation in European Cultural Heritage Institutions 2014, URL: https://www.egmus.eu/fileadmin/ENUMERATE/documents/ENUMERATE-Digitisation-Survey-2014.pdf [last accessed: April 2, 2021].

ing—is unavailable." [4] Transcribing the resulting digital images with human labor is resource intensive, whether that is via employed researchers, or in the management of volunteers in crowdsourcing initiatives.[5] The automatic generation of accurate, machine readable transcriptions of digital images of handwritten material has long been an ideal of both researchers and institutions, and it is understood that with successful Handwritten Text Recognition (HTR) "the next generation of digitized manuscripts promises to yet again extend and revolutionize the study of historical handwritten documents."[6] HTR generated transcriptions will allow the searching of vast manuscript repositories, extending the scale of the primary sources encountered by researchers, while also allowing textual information to be mined, visualized, and analyzed using various advanced Digital Humanities techniques.[7] The transformative aspects of this for research will therefore be manifold, providing "the basis for advanced semantic, linguistic, and geo-spatial computational analysis of historical primary source material."[8]

HTR, and its allied technique, Optical Character Recognition (OCR, the means by which images of text can be transformed into machine-processable format, focusing on the identification of single characters) both have long histories.[9] OCR, which generally involves segmenting characters in images of documents for individual recognition, has been routinely adopted by the library sector, with the resulting machine-processable texts primarily providing a finding aid,[10] for example for full text search in vast online libraries such as archive.org, or newspaper archives such as https://www.britishnewspaperarchive.co.uk. There are advanced OCR tools which can have success in transcribing clear handwritten text, such as

---

4 Laura Estill/Michelle Levy, Chapter 12: Evaluating Digital Remediations of Women's Manuscripts, in: *Digital Studies/Le champ numérique* 6 (2016), doi:10.16995/dscn.12.

5 Tim Causer/Melissa Terras, 'Many Hands Make Light Work. Many Hands Together Make Merry Work': Transcribe Bentham And Crowdsourcing Manuscript Collections, in Mia Ridge (ed.), *Crowdsourcing Our Cultural Heritage*, Farnham 2014, 57-88.

6 Laura Estill/Michelle Levy, Chapter 12: Evaluating Digital Remediations of Women's Manuscripts.

7 *The Programming Historian*, 2021, URL: https://programminghistorian.org [last accessed: April 2, 2021].

8 Guenter Muehlberger et al., Transforming Scholarship in the Archives through Handwritten Text Recognition: Transkribus as a Case Study, in: *Journal of Documentation* 75 (5/2019), 954-976, doi:10.1108/JD-07-2018-0114.

9 Herbert Schantz, *The History of OCR, Optical Character Recognition*, Manchester Center, VT, 1982; T. L. Dimond, Devices for Reading Handwritten Characters, in: *Papers and Discussions Presented at the December 9-13, 1957, Eastern Joint Computer Conference: Computers with Deadlines to Meet* (1957), 232-237.

10 Amarjot Singh/Ketan Bacchuwar/Akshay Bhasin, A Survey of OCR Applications, in: *International Journal of Machine Learning and Computing* 2 (3/2012), 314-318, doi:10.7763/IJMLC.2012.V2.137.

Kraken[11] and Tesseract,[12] although these work best as long as the handwritten characters are spatially separated. Integration of machine learning with OCR has improved its accuracy,[13] although OCR still struggles with complex fonts, layouts, or media, such as smudged newsprint.[14]

Recently, HTR techniques, which use machine learning approaches such as deep neural networks to extract visual features, and recognize characters and words in a segmented line of text via the calculation of overlapping probabilities, have become more stable, accurate, and efficient.[15] This complexity demands an increase in computational power at a scale beyond the processing needs for OCR. Over the past five years HTR has been increasingly integrated into digitization programs and scholarly projects across the academic library sector, although its use is not standardized, and there has been "next to no research on how best practices can be undertaken in storing, sharing, and explaining HTR generated content" within libraries and their online repositories.[16] There have been usability studies on particular HTR features and workflows, limited to studies within the software design process,[17] rather than understanding use of HTR within the wider, external context. A survey of 15 libraries was carried out in 2020 regarding their use of HTR,[18] showing that a lack of transcription of manuscript material impedes researchers. The aim of this chapter is to engage with the community undertaking HTR, to understand their motivations, practices, concerns, and insights.

---

11 http://kraken.re [last accessed: April 2, 2021].

12 https://github.com/tesseract-ocr/tesseract [last accessed: April 2, 2021].

13 Lakhmi Jain/Beatrice Lazzerini (eds.), *Knowledge-Based Intelligent Techniques in Character Recognition*, Boca Raton, FL, 2020.

14 Ryan Cordell, "Q i-jtb the Raven": Taking Dirty OCR Seriously, in: *Book History* 20 (1/2017),188-225.

15 Byron Leite Dantas Bezerra (ed.), *Handwriting: recognition, development and analysis*, Hauppauge, NY, 2017.

16 Melissa Terras, The Role of the Library When Computers Can Read: Critically Adopting Handwritten Text Recognition (HTR) Technologies to Support Research, in: Amanda Wheatley/Sandy Hervieux (eds.), *The Rise of AI: Implications and Applications of Artificial Intelligence in Academic Libraries*, Atlanta, forthcoming 2021.

17 Luis A. Leiva et al., Evaluating an Interactive-Predictive Paradigm on Handwriting Transcription: A Case Study and Lessons Learned, in *2011 IEEE 35th Annual Computer Software and Applications Conference*, IEEE 2011, 610-617.

18 Nikolina Milioni, *Automatic Transcription of Historical Documents: Transkribus as a Tool for Libraries, Archives and Scholars*, Master's Degree Project, Department of ALM, Uppsala Universitet, 2020, URL: http://www.diva-portal.org/smash/get/diva2:1437985/FULLTEXT01.pdf [last accessed: April 2, 2021].

AVAILABLE HTR OPTIONS

There are various HTR solutions currently available, and a choice has to be regarding approach, whether: working with computer scientists and developing bespoke tools and infrastructure; using commercial solutions provided by large-scale publishers and technology platforms as part of the digitization process; or using software that has emerged from the research community, now serving the community as a cooperative provider.

There is a large, long standing community of computational and information engineering researchers working on HTR, at the cusp of image processing and machine learning. Methods, results, evaluation, and code are published, alongside emerging benchmarks and best practice approaches.[19] Many libraries, archives, and historians have partnered with computational researchers to undertake interdisciplinary projects on HTR, applying it to a variety of languages and temporalities. For example, the *In Codice Ratio* project is developing "tools to support content analysis and knowledge discovery from large collections of historical documents," working on the collections of the Vatican Secret Archives.[20] The Connecticut Digital Archive partnered with UConn School of Engineering (and others), using a neural network approach to produce transcripts from the John Quincy Adams Papers.[21] Brigham Young University have built in-house approaches to HTR in conjunction with the Computer Science department and Family History Technology Lab, extracting information from 1918 pandemic death certificates.[22] Such local interdisciplinary partnerships can be successful, however, they require careful management of resources, expertise, and results.[23] These systems can be adopted and adapted by others: the Monk system has been developed by the University of Groningen, in

19 Joan Andreu Sánchez et al., A Set of Benchmarks for Handwritten Text Recognition on Historical Documents, in: *Pattern Recognition* 94 (2019), 122-134.

20 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), 263-272.

21 Jean Nelson, UConn Library, School of Engineering to Expand Handwritten Text Recognition, in: *UConn Today* 09.07.2020, URL: https://today.uconn.edu/2020/07/uconn-library-school-engineering-expand-handwritten-text-recognition/# [last accessed: April 2, 2021].

22 Veronica Maciel, Database of 1918 Pandemic Deaths Inspires Answers for the Future, in: *The Daily Universe* 11.02.2021, URL: https://universe.byu.edu/2021/02/11/database-of-1918-pandemic-deaths-inspires-answers-for-the-future/ [last accessed: April 2, 2021].

23 Melissa Terras, Being The Other: Interdisciplinary Work in Computational Science and the Humanities, in Marilyn Deegan/Willard McCarthy (eds.), *Collaborative Research in the Digital Humanities*, Farnham 2012, 213-240.

collaboration with the Dutch National Archives, and the code is available for reuse[24] for application to documents from the 1400s.

Some established technology platforms, publishers, and search engines are actively developing HTR tools to support full-text search of library and archival content. For example, Adam Matthew Digital currently claims it "is the first publisher to utilize artificial intelligence to offer Handwritten Text Recognition (HTR) for its handwritten manuscript collections," offering full-text search of transcriptions of seven mass-digitized archival collections,[25] processed by Planet-AI's ArgusSearch.[26] Adam Matthew's Quartex document management system can be licensed for sharing and display of mass-digitized content, and this now has HTR embedded, increasing the searchability of hosted collections.[27] Full-text searching via HTR is also available in some of the publisher Gale's online archives.[28] Google recently launched Fabricius,[29] which uses machine learning to support the transcription and translation of Ancient Egyptian hieroglyphs, and Google continue to expand their API for developers wishing to detect handwriting in images.[30] Mitek (previously A2iA) offer handwriting recognition software for license.[31] However, commercial systems can be opaque, the algorithms and approaches used are seldom published, and care must be taken regarding copyright, image licensing, and long-term storage of digital assets, when brokering partnerships with such major technological entities.

---

24 Monk, Homepage, 2020, https://www.ai.rug.nl/ lambert/Monk-collections-english.html [last accessed: April 2, 2021].

25 Adam Matthew Digital, Artificial Intelligence Transforms Discoverability of Handwritten Manuscripts, 2020, URL: https://www.amdigital.co.uk/products/handwritten-text-recognition [last accessed: April 2, 2021].

26 Adam Matthew Digital, Historic 17th & 18th Century Manuscript Documents Made Fully Searchable Using Artificial Intelligence, 06.09.2017, URL: https://www.amdigital.co.uk/about/news/item/htr-planet [last accessed: April 2, 2021].

27 Quartex, Homepage, Features, 2020, URL: https://www.quartexcollections.com [last accessed: April 2, 2021].

28 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, URL: https://www.gale.com/intl/c/china-and-the-modern-world-hongkong-britain-china [last accessed: April 2, 2021].

29 Google Arts and Culture, What is Fabricius?, 2020, URL: https://artsexperiments.withgoogle.com/fabricius/en [last accessed: April 2, 2021].

30 Google Cloud, Detect Handwriting in Images, AI and Machine Learning Products, 2021, URL: https://cloud.google.com/vision/docs/handwriting [last accessed: April 2, 2021].

31 Mitek, Handwriting Recognition, Mitek's Handwriting Recognition Software offers an Unprecedented Level of Access, 2020, URL: https://www.a2ia.com/en/handwriting-recognition [last accessed: April 2, 2021].

A current half-way house between these two approaches is the software Transkribus:[32] developed by researchers in an EU funded project, and available for institutional or individual users, via a paid-for, co-operative model. It is this platform (and generous input from its users) which provides the discussion presented here.

TRANSKRIBUS AND READ-COOP

Work on the Transkribus platform began in 2013, with funding from the European Commission's Seventh Framework Programme (FP7), as part of the tranScriptorium project (2013–2015), which brought together a large computer science research community working on deep learning approaches to HTR, coordinated by the University of Innsbruck. The HTR client was initially launched in early 2014,[33] providing free access to the tranScriptorium HTR infrastructure, becoming known as Transkribus from February 2015. The successor project Recognition and Enrichment of Archival Documents (READ, 2016-19), funded under the EU Horizon 2020 scheme, aimed to develop functionality and usability. From July 1, 2019, Transkribus has been operated and further expanded by the READ-COOP,[34] a mechanism to sustain and grow the Transkribus infrastructure beyond the end of its grant-funded period, built around a European cooperative society governance model.[35] Transkribus switched from a "free at the point of delivery" mode to a purchasable, credit-based financial model on October 19, 2020.[36]

The functionality of Transkribus is fully documented elsewhere:[37] briefly, uploaded images are segmented into lines (via both automatic and manual segmentation tools, allowing user correction), before a transcription is generated. A training process is undertaken on a subset of material: the end user corrects any errors the system has made (75 pages or 15,000 transcribed words gives adequate results). This then generates an HTR model, which can give more accurate results across

32 https://transkribus.eu/Transkribus/

33 Philip Kahle et al., *Deliverable 4.2, READ Platform And Service Maintenance, Deliverable Submitted To The European Commission*, 2017, URL: https://read.Transkribus.eu/wp-content/uploads/2017/12/D4.2.pdf [last accessed: April 2, 2021].

34 READ-COOP, Revolutionizing Access to Handwritten Documents, 2021, URL: https://readcoop.eu [last accessed: April 2, 2021].

35 European Commission, The European Cooperative Society (SCE), n. d., URL: https://ec.europa.eu/growth/sectors/social-economy/cooperatives/european-cooperative-society_en [last accessed: April 2, 2021].

36 READ-COOP, FAQ, When Will the Payment Model for Transkribus Start?, 2020, URL: https://readcoop.eu/Transkribus/credits/faq/ [last accessed: April 2, 2021].

37 Guenter Muehlberger et al., Transforming Scholarship in the Archives through Handwritten Text Recognition.

similar mass-digitized content, also creating a feedback loop increasing the efficacy of the underlying neural network, and improving the output of the system for future users. Some models from other projects are made available for public use,[38] and users can apply these to their own material: Transkribus has now been trained to recognize text in a variety of languages, including English, Italian, Dutch, Latin, Swedish, Finnish, Danish, Old German, Polish, Bangla, Hebrew, Church Slavonic, and Arabic, with different models being created for distinct time periods. Best case results from Transkribus generate a Character Error Rate of below 5% on handwritten material, and below 1% on print material. The resulting transcriptions can be exported in a variety of formats including plain text, XML (including ALTO and TEI), PDF, and Microsoft's proprietary .docx.

Registered user numbers of Transkribus have increased dramatically over the past few years: from 2,200 in 2015; 4,800 in 2016; 8,500 in 2017; 17,000 in 2018; 30,000 in 2019; to 45,000 in 2020 (usage reports show between 800 and 1400 active users per week). In February 20201 the platform achieved 50,000 users, including major memory institutions worldwide: the Rijksmuseum; Rahvusarhiiv (National Archives of Estonia); the Arkivverket (National Archives of Norway); Kansallisarkisto (The National Archives of Finland); and the British Library. Universities using Transkribus include the University of Cambridge, Université du Québec à Rimouski, and the Universitat Politècnica de València (Technical University Valencia).[39] The end of year report showed that in 2020 there were 18,6264 unique active users, uploading 19.9m images in 660,077 jobs, with 8443 new models generated and 490m words transcribed.[40]

METHOD

In order to elucidate active use of HTR, we have undertaken various surveys of the user community. A survey of 25 short questions was issued after the Transkribus User Conference 2018,[41] by Dr Louise Seaward.[42] The 72 responses allowed the development team to ascertain overall opinions of Transkribus, with 80% of users

38 READ-COOP, Public Models in Transkribus, 2021, URL: https://readcoop.eu/Transkribus/public-models/ [last accessed: April 2, 2021].

39 READ-COOP, Members of READ-COOP SCE, 2021, URL: https://readcoop.eu/members/ [last accessed: April 2, 2021].

40 Transkribus, Latest report from 2021-01-01 with detailed user data over a period of 365 day(s), generated by Transkribus@uibk.ac.at.

41 READ-COOP, Transkribus User Conference 2018, URL: https://readcoop.eu/Transkribus-user-conference-2018/ [last accessed: April 2, 2021].

42 Then working with READ, as part of the Transcribe Bentham project, at University College London.

saying they were quite or very satisfied. The survey also highlighted aspects of the system which users were finding difficult, such as training HTR models and managing documents, leading to more targeted support.[43] However, this tools assessment survey was not a "reception study" which aims to study digital media uptake and use among a particular constituency,[44] or community of practice.[45]

The Transkribus delivery team developed a more encompassing survey containing 50 detailed questions (many with non-mandatory subsections) that would take approximately 30 minutes to complete. Ethical approval was granted by the University of Edinburgh. The survey was live between March 26 and April 24, 2019, hosted by Jisc Online Surveys, a GDPR (General Data Protection Regulation) compliant research questionnaire platform. It was shared with all 20,000 registered users (at the time) via the email newsletter, and also posted on Facebook on both the official Transkribus Platform group[46] (then 185 members), and the unofficial Transkribus Users group[47] (then 250 members: a public forum where the majority of community discussions still happen). We did not share with a wider online audience, preferring to target active users of Transkribus via these relatively closed fora. The study did not store personal information, or any protected characteristics data from participants: results are fully anonymized.

This case-study, Reflection-in-Action questionnaire approach allowed us to identify "features of the practice situation – complexity, uncertainty, instability, uniqueness and value conflict."[48] The results were synthesized using a Content Analysis recursive methodology.[49] Any otherwise unattributed quotes given subsequently are taken from anonymous survey responses, with only minor editing undertaken of obvious typographical errors.

---

43 Materials have since been compiled in the Transkribus Resource Base: https://readcoop.eu/transkribus/resources/ [last accessed: April 2, 2021].

44 Pertti Alasuutari, Three Phases of Reception Studies, 3.

45 Etienne Wenger, *Communities of Practice: Learning, Meaning, and Identity*, Cambridge 1999.

46 Transkribus Platform, Official Facebook Group, 2021, URL: https://www.facebook.com/Transkribus/groups/ [last accessed: April 2, 2021].

47 Transkribus Users, Facebook Group, 2021, URL: https://www.facebook.com/groups/614090738935143 [last accessed: April 2, 2021].

48 Donald A. Schön, *The Reflective Practitioner: How Professionals Think In Action*, New York 1983, see 18.

49 Klaus Krippendorf, *Content Analysis: An Introduction to its Methodology*, Thousand Oaks, CA, 2018.

RESULTS

Response Rate

There were 155 survey responses. There were approximately 800 different active accounts in the survey period (out of 20,000 registered accounts): giving a survey response rate of 19% of active users. Online surveys tend to have low response rates, particularly for detailed surveys:[50] the reasonable response rate by the Transkribus user community indicates their personal investment in the platform, although keen individuals would have had more motivation to respond. There was one vexatious return: a user who took the time to write "YOUR TOOL DOES NOT WORK" in every available field. They had tried to apply Transkribus to Ancient Greek, but had not made any attempt to train it, and public models for Greek were not available until 2020. There were therefore 154 useful, considerate responses to the survey, and 103 (67%) of respondents answered every question: answers tended to tail off towards the questionnaire end.

User Information

53% of respondents were associated with an academic institution, and 47% were not: Transkribus is also used outside the academic research community. Of the 112 who provided a job title, there were 18 researchers, 13 professors, 11 archivists, 6 students, 6 retirees, 3 librarians, 3 teachers, and 3 volunteers, with a variety of others including project management and consultancy, and an architect, dentist, electrician, head of tourism, software engineer, translator, and web developer: HTR has a wide appeal. Returns represented an international user community with a European majority (88%), also including respondents from North America (7%, including 7 from the USA, and 3 from Canada), Central and South America (3%, with 1 respondent each from Brazil, Colombia, Mexico, and Uruguay), and Australasia (2%, with 3 respondents from New Zealand). There were no respondents from Africa or Asia. Respondents from Europe included those from Germany (26%), Austria (11%), Switzerland (10%), The Netherlands (8%), United Kingdom (8%), France (6%), Italy (5%), Denmark (3%), Spain (3%), Hungary (2%), with one respondent each from Belgium, Finland, Greece, Ireland, Norway, Poland, Portugal, Slovakia, Sweden and Turkey.[51]

50 Joel R. Evans/Anil Mathur, The Value of Online Surveys: a Look Back and a Look Ahead, in: *Internet Research* 28 (4/2018), 854-887, see 859.

51 An analysis of 21,260 logins to the platform from 822 unique users over the period of the survey, by the domain name of their email linked to their account, likewise showed a dominance of European users, with Germany (23%), Switzerland (11%), the Netherlands (10%) Austria (9%), France (7%) and Denmark (7%), roughly mapping onto survey responses, and

The majority of respondents had some professional experience with palaeography, or reading historical documents themselves: 30% described themselves as expert; 30% intermediate; only 29% saying they were a beginner; and 11% having no previous experience. 24% had been working with Transkribus for over a year, 41% for between one month and year, and 36% for less than a month: the majority of respondents had reasonable experience with the system. 21% described themselves as expert users, 43% as intermediate users, and 36% were beginners. Respondents worked with Transkribus weekly (31%) monthly (20%), every few months (7%) or occasionally (27%): only 10% of respondents worked with Transkribus every working day, and 5% of respondents had never used it themselves. The majority of respondents (64%) were working with Transkribus in an individual capacity on their own projects, with only 33% using it as part of a program of work in an organization or institution. 3% were working on both institutional and personal projects. This therefore reveals an important user community of individuals outside institutional settings that the platform must ensure to support, going forward (and expands the list of "Expected user of HTR" beyond only the previously identified "individual researchers with experience in handwritten documents" and "volunteers which collaborate in large transcription projects").[52]

Collections Information

Given that the majority of users are individuals working on their own projects, it is not surprising that many of the collections being analyzed were relatively small: 62% were below 1000 pages, and of these, 27% were less than 100 pages. A total of 38% of respondents were processing collections above 1000 pages, with only 17% working with collections over 10,000 pages. 4 respondents were working on projects processing over 500,000 pages: the largest project was planning to process 4 million. In total, the respondents were planning to process more than 8 million pages of handwritten material, containing an estimated one billion words.

Respondents obtained their digital images from a variety of sources (often more than one). Digitization of primary historical material remains a bottleneck for applying HTR, and projects are dependent on prior activity ("For my ongoing project:

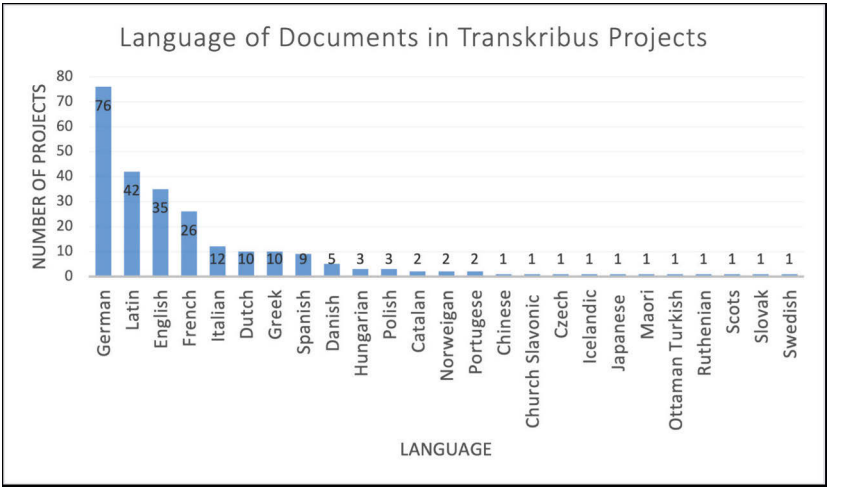
giving confidence in its coverage. There were active logins from 53 countries during that period, including Australia (1%), and fewer than 1% of users from Russia, Chile, Côte d'Ivoire, Cocos Islands, South Africa, India, People's Republic of China, Argentina, Iraq, Turkmenistan, El Salvador, Laos, and Montenegro, showing that the survey did not elicit responses from all constituencies.

52 Serena Ammirati et al., In Codice Ratio: Scalable Transcription of Historical Handwritten Documents, in: *Symposium on Advanced Database Systems, Proceedings of the 25th Italian Symposium on Advanced Database Systems* (2017), URL: http://ceur-ws.org/Vol-2037/paper_11.pdf [last accessed: April 2, 2021], see 4.

All digitized. For my next project: Have to digitize myself"). More than half of respondents were using digital images within their own personal ownership, with 58% of images created by the respondents ("Only about half of my sources have been digitized already, but I can almost always (and do) take photos of them for myself, at archives"), with a further 11% using images from another's personal collection. 62% were using images from an institutional collection, such as a library or archive, and 14% had downloaded images to use from large digital cultural heritage aggregators (such as Europeana, and Flickr Commons). Only 5% were using images from a commercial digitization provider ("many handwritten documents of interest for genealogical research are available through commercial providers or personal sources. However, quality of the digitized documents varies considerably, some commercial sources are so poor as to be unusable by any HTR or OCR software"). 4% of respondents downloaded images from other websites, including Google Books, local archives, and church offices ("it was unexpectedly easy getting the digitized texts"). Many (10%) did not know who had digitized the documents they were working with, although 43% believed they were funded by a memory institution, and 23% believed digitization had been funded via a research grant. For larger projects linked to institutions, recent developments in HTR are affecting institutional choices regarding digitization ("we ran our own campaign… However, using a tool like Transkribus has forced us to better organize our digitization… to obtain better pictures for better OCR" and "Transkribus is a driver; an impetus to digitizing more textual material in contrast to formats like photographs and maps, etc. It has also highlighted the quality required for digitization and transcription. Some digital images that we have done need to be reimaged").

There were a variety of languages being processed, from the medieval period onward, as seen in Fig 7.1, showing the flexibility of the system, and how it has been trained to encompass a diverse range of languages.

Fig 7.1: The languages being analyzed using Transkribus in respondents' projects. Many were using material in more than one language, and time period (for example, US/UK English, or Middle High, Middle Low, and Modern German). Although some languages, such as Chinese, are not processable by Transkribus, a user explained "there are a number of tools for pre-processing that can be relevant, like page segmentation or other features." These, plus user-generated models covering particular epochs and languages, have expanded the reach of the tool beyond the core European languages initially covered in the first release of the tool.



When asked to provide project details, a quarter of respondents were using the outputs of Transkribus for genealogical research, which could be deeply personal ("(re)construct family and neighborhood histories" and "Just give to a friend, it's important to him"). The majority were using Transkribus to support research purposes, including (most popular first): to publish online as a finding and research aid ("attempt to link it with our online catalogue so that the transcription can be viewed online" and "the library will make it publicly available on the Internet, with the images, transcribed text and searching tools"); to use as a primary source ("extract facts, keep the full text as reference"); further use in critical and scholarly editions, including TEI-XML ("assess > edit > publish > archive portions of it in a TEI XML edition" and "a TEI-XML database, enriched in linguistic tagging and available to the public on our website"); corpus linguistics ("post processing: NLP and normalization. We will use it for linguistic and historical research"); to facilitate geographical analysis via GIS; and to publish and share the results openly elsewhere ("open repository as reusable data" and "bring it to Wikisource to be used world-

wide"). A number of respondents specifically said they were creating a novel data source to analyze for their PhD study, in history or digital humanities.

There were a wide range of document types mentioned including: personal diaries; parish registers; burial registers; municipality records; court records; legal documents including wills, deeds, and charters; military records including prisoner of war data, and letters; police reports; minutes and records of various meetings; curatorial records within museums; public health information including asylum records, and post-mortem inventories; scientific records including meteorological, climate, and phenological data; and the transcription of manuscripts from the collections of renowned historical figures. Respondents believed that Transkribus was appropriate to use for these tasks for a variety of reasons including: other tools were not successful ("OCR does not work with these kind of documents"); the volume of documents requiring transcription meant the process would have to be automated if to be transcribed at all, and there was much interest in them; the process was relatively easy; resources were not available to transcribe these manually ("we are a small workforce so having HTR complete even a portion of the transcription process is helpful to us" and "it is very difficult to know if a document will contain any useful information before deciphering it which takes a lot of time"); the collections were relatively homogenous and were appropriate for this approach ("same writer over several years, standard layout"); and the fact that Transkribus was currently "the only one which is able to do that successfully." When asked if the documents would be transcribed without access to HTR infrastructure, 33% said there would be no resources to do so, 40% said that it could happen, but it would take time; and 26.5% said that some of the collection had already been transcribed manually. Only 8% gave a firm "yes" that transcription would happen without HTR being available: clearly the infrastructure is already transforming access to handwritten historical sources given available institutional and individual resources for transcription by other methods.

The Potential Efficiency of HTR

21% of respondents said Transkribus had delivered a significant increase in efficiency to their projects, with a further 23% stating that it was a useful increase in efficiency. Many were still training and trialing the software: 36% were hoping for further efficiencies in the future. However, 12% said that it had not sped up the processes of generating transcripts from historical texts. This is a matter of scale, depending on the size of collections being analyzed, and also one of training and experience with HTR. Only 4% of respondents said results generated from Transkribus were very accurate and required little correction (there have been improvements since to the Character Error Rate). 34% said results were quite accurate; 16% said results were disappointing, and 8% found results unusable. 21% acknowledged

that results were variable dependent on individual texts. The remainder were still planning to use HTR and could not comment. 19% of respondents said HTR was an essential tool for their project, and it could not be done without it. 12% said it was part of their routine workflow, with a further 52% saying it had the potential to become a routine part of digitization. However, 15% described Transkribus as "an interesting experiment but not terribly useful." These comments capture a technology maturing, and entering the digitization pipeline as an increasingly essential tool, albeit with further potential.

Only six respondents had tried to calculate how and where HTR would speed up the transcription process ("transcription time can be reduced by a factor of 3 or 4," "the time required for data entry would be reduced to a small fraction (∼10%) of what it is now"). This obviously has cost saving potential ("with a small loss of accuracy (+3-5% CER) more than 80% of costs can be saved compared by manual transcription by students"), although the HTR process may replace some rounds of human transcription ("We hope… to first upload and transcribe a document using the software, then pass the transcription onto a person who can go back in and correct…. At the moment transcribing using human transcribers takes at least two rounds of transcription to get a finished transcript which is obviously costly in time").

Transkribus Features

When asked what the most useful features were, respondents commented (from most popular): automatic line detection, layout analysis, and segmentation; HTR training; Key Word Spotting; tagging functionality; export functions (including TEI); the ability to manually correct and edit results; the ability to share ongoing work with others including working on group transcription; table region generation; Unicode integration; and the ability to create standardized transcripts. Suggested improvements included: better HTR handling of abbreviations; the possibility of automated export to online client;[53] and having the handbook and interface available in different languages.

One quarter of respondents noted unexpected benefits to using the platform, including: the results ("I'm stunned at the quality. And I am an OCR expert for many years"); the supportive Transkribus community and connections to other projects; successful use in the classroom; and the "discovery of needles unsought for in haystacks of historical documents." Disappointments, though, included: difficulty onboarding beginners; technical issues including installing updates; java integration; the occasional slowness of the client; problems with the interface design ("yes, bugs, features for usability & user-interface"); the need for further training

53 This feature will be launched in 2021.

and handbook materials ("steep learning curve with user interface, having to tell co-workers how things work and what they need not look for"); and the level of system training needed ("it takes a lot of transcriptions to get a good HTR model").[54]

Considering HTR

Respondents were asked what three words came to mind when considering HTR. "Help" and "Access" were the most common thoughts, along with expressions regarding innovation, speed, and excitement:

Fig 7.2: Wordcloud generated from respondent's answers to "When you think of Handwritten Text Recognition (HTR), what three words come to mind?"

When asked what the benefits of HTR were for research, 63% of respondents mentioned speeding up transcription which "allows for transcribing data that otherwise would not have been transcribed." 50% said the success of HTR will encourage more historical resources to be digitized. 46% said that the ability to easily search within vast collections of images of texts ("some documents are quite repetitive (notarial acts), yet they can contain very interesting information. If there is a way to analyze these texts faster, we can focus on the more relevant information"). A

54 Materials have since been compiled in the Transkribus Resource Base: https://readcoop.eu/transkribus/resources/ [last accessed: April 2, 2021].

further 46% said that it allowed easier sharing of document contents ("hopefully entanglement of interesting details out of the jungle of handwritings"). 42% believed that HTR demonstrated that technology can be employed for social and non-commercial good, and 41% believed that HTR has the potential to increase the quality of existing transcriptions. 37% believed it would change the scope of the historical documents we can access ("it enhances the research possibilities (and questions)"). 30% believed HTR would democratize knowledge, with 25% believing that a benefit was being able to build upon previously digitized content ("making available the knowledge that is locked in digital images"), and that it "keeps us ambitious even when funding is scarce."

Respondents were asked about how using HTR had changed their own approach to reading historical documents. 59% said it had encouraged them to be optimistic about the future of digital research tools and digital infrastructure, 48% said it has made it quicker and more efficient to generate transcripts, and 31% said HTR made it possible to generate transcripts from a wider variety of documents than they would be able to previously. 20% said that HTR had improved their computer skills, and confidence with digital tools, 19% said it had made it possible to undertake new research. 38% said it improved their own palaeography skills and only 10% said that it had reduced the need to use them. Respondents were mostly emphatic that HTR would not replace skills in reading historical texts: "Palaeography must still be taught so that the researcher can look critically at the transcription. Plus palaeography is not only about reading a text, recognizing the type of writing provides other type of information." It was repeatedly pointed out that palaeographic skills were necessary to train models, and that "a critical review of machine performance will always be necessary": "I just cannot stress enough that I don't think machines should ever be trusted, so historians still must be sufficiently trained in palaeography to not rely on machine transcriptions."

When asked how HTR complements their own skills, the most popular comment was on its efficiency ("Speed!"), but individual comments also highlighted a greater awareness of the varieties of handwriting, that HTR can transcribe documents in languages they do not speak ("It can predict a word I have difficulty interpreting"), and that the scale of transcription, even though not always correct, "provides a good amount of context, making reading /transcribing faster. Scanning through documents is quite a nice thing."

Only 10% of respondents said that they fully understood the technology behind HTR: "I think we need to at least in theory to understand the technology (and epistemological consequences)." Many expressed frustration at this, and an interest in knowing more about the process ("I would like to out of curiosity, and scientific integrity"), and that in doing so it may improve their use of HTR ("This might also speed up the transcribing-process since I might know which mistakes are ignorable and of little relevance to the machine") and

> knowing a little about how the different algorithms work might help to adjust expectations. although a lot can already be done with HTR, a lot cannot (yet?). Tools in Transkribus are often "black boxes" which may or may not work on certain documents. One has to invest quite some trial-and-error time to check out what works best in a certain context. For the (interested) user not trained in computer vision or pattern recognition it would be helpful to get more background information about what happens in the box, to understand why it fails in case A but works for case B.

It is clear that by understanding the HTR process more fully, scholars will be able to gauge its "epistemic affordances": the abilities, possibilities, and limitations of this environment when used in knowledge creation.[55]

When asked about the future of HTR and historical texts, most respondents predicted it "will vastly increase accessibility" with "a wider availability of sources" and although "transcribing is not the same as understanding, it will have a big influence on access," primarily surrounding the range of primary source material that can be searched, processed, and analyzed. This may "open up sources that were inaccessible due to their volume or complexity and we can start studying topics that are sparsely addressed." The public engagement aspects of HTR should not be overlooked ("I think it is a great tool for historians on the one hand, but also a great tool for 'hobby historians' like me… It brings history and historical information closer to the people"). However, it was acknowledged that HTR still "needs development in its usability and accuracy before these expectations can be fully met."

Data processing approaches used with HTR outputs, such as text mining, text analysis, and linked data, were mentioned as transformative, given in the "vast corpus" there is "the possibility to cross-reference, finding new links." This will "increase our knowledge and allow history to be more relevant," while enhancing "the understanding of circulation of texts/knowledge/expertise in the past" while also producing a "higher quality of source editions and research based on them." However, this will also require that "adequate support infrastructures will be designed to display and disseminate what has been achieved/accomplished" including web-platforms to share results, models, and datasets; IIIF interoperability to import document scans hosted elsewhere;[56] adequate, affordable data storage platforms for sharing outputs transparently, as well as new peer-review mechanisms to be

---

55 Lina Markauskaite/Peter Goodyear, Epistemic Thinking, in: *Epistemic Fluency and Professional Education*, Dordrecht 2017, 167-194, see 185.

56 This is currently under development, see Florian Krull/Guenter Muehlberger/Melissa Terras, Transkribus and IIIF: Beneficial Possibilities between Image Sharing and Handwritten Text Recognition Frameworks, in: *IIIF Conference - Göttingen, Germany, 24 June – 28 June 2019*, URL: https://iiif.io/event/2019/goettingen/program/26/#Transkribus-and-iiif-beneficial-possibilities-between-image-shar [last accessed: April 2, 2021].

able to judge the quality of scholarly editions produced in this way. It was stressed that historians need training in advanced digital methods (and large-scale project management) to respond to these opportunities: "at the moment it's a geek topic."

At the time of the survey, Transkribus was available without charge: a third of respondents had not considered its sustainability, and just half were concerned about ongoing availability of the tool. Only 9% were not concerned about future sustainability of the infrastructure. Respondents' feelings were mixed on issues of copyright and intellectual property surrounding the developing infrastructure. A third were not concerned, a third were unsure, and a third mentioned issues in transparency, accountability, licensing, and open publishing of models. Transkribus is only partially open source (the processing components are closed, although the client is fully open source), and many voiced concerns that their labor was contributing to a closed system ("I think it's important to make the technical workflows and implementations in Transkribus available to the public"). These are complex issues to navigate, regarding both the business model underpinning the sustainability of the system, and the Intellectual Property Rights of systems themselves.[57] There are also ethical issues to resolve, particularly when related to cultural heritage and ethnographic collections: "who owns the model created by the digital images owned by the user? [for the ethnic minority community whose collection is being analyzed it matters] that the images are not stored overseas, or in the cloud [so that] traditional knowledge can be worked with and protected." These align with broader discussions on data ethics in cultural heritage,[58] including GDPR, privacy legislation, ethics of care,[59] and how this intersects with developments in artificial intelligence, testing the limits of existing legal concepts and approaches.

DISCUSSION

Throughout survey responses, aspects of speed and efficiency of HTR were stressed, expanding the volume of historical documents that are now searchable, accessible, and available for further processing, beyond what was previously possible via human labor. The promise already realized is increased access to the historical record. Support, collaboration, and data-sharing (including HTR models, the corpora they were based on, and the transcriptions generated) will further extend that potential, as well providing training in Digital Humanities options for further analysis of

57 Burkhard Schafer et al., A Fourth Law of Robotics? Copyright and the Law and Ethics of Machine Co-Production, in: *Artificial Intelligence and Law* 23 (3/2015), 217-240.

58 Sarah Colley, Ethics And Digital Heritage, in: Tracy Ireland/John Schofield (eds.), *The Ethics Of Cultural Heritage*, New York, NY, 2015, 13-32.

59 Temi Odumosu, The Crying Child: On Colonial Archives, Digitization, and Ethics of Care in the Cultural Commons, in: *Current Anthropology* 61 (S22/2020), S289-S302.

the large-scale generated data (such as the tutorials provided by the Programming Historian),[60] and ongoing expansion of the Transkribus Resource Base,[61] which is compiling how to guides, FAQs, video tutorials, and publications. HTR platforms should be mindful of developing discussions regarding data-ethics and archival content, reminding users of their responsibilities, and navigating the interaction with complex data models created by machine learning methods.

Arguments used by our respondents are reminiscent of reception discussions surrounding mass digitization in the early 2000s, for example the comments on Google Books digitization made by Dan Cohen in 2010, when the initiative was reaching maturity:

> The existence of modern search technology should push us to improve historical research. It should tell us that our analog, necessarily partial methods have had hidden from us the potential of taking a more comprehensive view, aided by less capricious retrieval mechanisms which, despite what detractors might say, are often more objective than leafing rapidly through paper folios on a time-delimited jaunt to an archive.[62]

It has been said of historical material that "the range of interest in the archives is so extensive that everything is potentially of interest."[63] As one of our respondents commented, "it is very difficult to know if a document will contain any useful information before deciphering it which takes a lot of time." Through discussion with our community, it has become clear that HTR has the potential to provide the means of "discovery of needles unsought for in haystacks of historical documents," and it is *this* aspect of HTR that will have most effect on present historical method and approach.

However, effective rollout of HTR is dependent on mass-digitized content, and the selection practices of organizations (due to limited resources)[64] has meant that certain collections have been prioritized for mass-digitization.[65] There is a limit to the possible impact of self-funded digitization by individuals, such as detailed by our respondents. To make the most of HTR, then, requires a move away from selection criteria for collections digitization, and a move towards batch digitization of

60 https://programminghistorian.org[last accessed: April 2, 2021].

61 https://readcoop.eu/transkribus/resources/[last accessed: April 2, 2021].

62 Dan Cohen, Is Google Good for History, 07.01.2010, https://dancohen.org/2010/01/07/is-google-good-for-history/ [last accessed: April 2, 2021].

63 Michael Moss/David Thomas/Tim Gollins, The Reconfiguration of the Archive as Data to be Mined, in: *Archivaria* 86 (2018), 118-151, see 139.

64 Lorna Hughes, *Digitizing Collections*, 2004, 31-53.

65 Tessa Hauswedell et al., Of Global Reach yet of Situated Contexts: An Examination of the Implicit and Explicit Selection Criteria that Shape Digital Archives of Historical Newspapers, in: *Archival Science* 20 (2020), 139–165, doi:10.1007/s10502-020-09332-1.

complete archives, which will require more resources (both human, and computational). To fully unlock the potential of HTR (and unlock its cost-saving potential), the spend on digitization has to increase.

Since this survey was undertaken, the tools available via Transkribus have been developed and expanded, and our user base has more than doubled. It will be important to undertake such analysis of our user community every two to three years, to ascertain their needs, but also to record how HTR is integrating with historical method, and embedding itself into the hobby, research, library and archive community. This survey was also undertaken before Transkribus became a charging system: future work will analyze how this has changed the Transkribus user-base, and their activities.

More work is needed on how to model the economic benefits of using HTR, but this should also be offset against costs now incurred to access the system, and the wider environmental costs of processing large volumes of data, and creating new models ("the time to create training data and run models over a very large set of images may take too much time and computing power").[66] Additionally, calculations such as "more than 80% of costs can be saved (compared by manual transcription by students)" indicate that employment and skills development opportunities are being disrupted by this technology, and this should be viewed with care to ensure the next generation of palaeographers, archivists and digital humanists have opportunities to develop the crucial skillsets needed to work alongside and with HTR, rather than being replaced entirely by it.

CONCLUSION

This chapter details the first reception study on how the historical community, broadly framed, have been adopting Handwritten Text Recognition within their practice, focusing particularly on users of Transkribus. Doing so has indicated the range of application of HTR, but also the breadth of interest in it, by a diverse international community. Important themes to emerge include the belief that HTR is already increasing the speed of transcription and expanding the volume of primary historical content available for further analysis, and that there is a hunger for this technology, with individuals willing to invest considerable effort in learning the tools. HTR has the potential to provide a step-change in historical method, focus, and related findings in changing the scope from selecting particular manuscript material for digitization and transcription (due to the costs involved), to transcribing and analyzing complete archival collections. However, this still has dependencies on access to digitized images of collections. Recommendations for memory

66 Thomas Griffin, *Why We should Care About The Environmental Impact of AI*, 2020.

institutions include: digitization processes should produce (and make available) high quality images[67] to facilitate HTR, and also digitize as widely and completely as possible. Recommendations for Transkribus (and other HTR providers) include: further information on how HTR operates should be provided, to allow researchers to understand "epistemic affordances"; users should be pointed to resources and training on how best to utilize the results of HTR; and individual users and projects should be encouraged to share their models, results (and where possible, data), to benefit the wider historical community (although the tension that Transkribus itself is not open-source should be acknowledged). Finally, it is important to document the changes this machine learning approach is making to both the historical record and the community using it. This first, detailed evidence of concerns, opinions, and considerations from those using HTR should have regular follow up studies to understand our changing information environment, as well as benefit the ongoing development of Transkribus as an HTR service.

67 Although Transkribus can work effectively with images of 300 DPI, serious problems arise from processing images scanned from microfilm instead of the originals, bitonal TIFF files, images taken at an angle, out-of-focus pictures, and other technical inadequacies due to inconsistency, or un-or poorly trained labour.

Bibliography

Adam Matthew Digital, Artificial Intelligence Transforms Discoverability of Handwritten Manuscripts, 2020, URL: https://www.amdigital.co.uk/products/handwritten-text-recognition [last accessed: April 2, 2021].

Adam Matthew Digital, Historic 17th & 18th Century Manuscript Documents Made Fully Searchable Using Artificial Intelligence, 06.09.2017, URL: https://www.amdigital.co.uk/about/news/item/htr-planet [last accessed: April 2, 2021].

Alasuutari, Pertti, Three Phases of Reception Studies, in: Pertti Alasuutari (ed.), *Rethinking the Media Audience, the New Agenda*, London 1999, 1-8.

Ammirati, Serena, et al., In Codice Ratio: Scalable Transcription of Historical Handwritten Documents, in: *Symposium on Advanced Database Systems, Proceedings of the 25th Italian Symposium on Advanced Database Systems* (2017), URL: http://ceur-ws.org/Vol-2037/paper_11.pdf [last accessed: April 2, 2021].

Bezerra, Byron Leite Dantas (ed.), *Handwriting: recognition, development and analysis*, Hauppauge, NY, 2017.

Causer, Tim/Terras, Melissa, 'Many Hands Make Light Work. Many Hands Together Make Merry Work': Transcribe Bentham And Crowdsourcing Manuscript Collections, in Mia Ridge (ed.), *Crowdsourcing Our Cultural Heritage*, Farnham 2014, 57-88.

Cohen, Dan, Is Google Good for History, 07.01.2010, https://dancohen.org/2010/01/07/is-google-good-for-history/ [last accessed: April 2, 2021].

Colley, Sarah, Ethics And Digital Heritage, in: Tracy Ireland/John Schofield (eds.), *The Ethics Of Cultural Heritage*, New York, NY, 2015, 13-32.

Cordell, Ryan, "Q i-jtb the Raven": Taking Dirty OCR Seriously, in: *Book History* 20 (1/2017),188-225.

Dimond, T. L., Devices for Reading Handwritten Characters, in: *Papers and Discussions Presented at the December 9-13, 1957, Eastern Joint Computer Conference: Computers with Deadlines to Meet* (1957), 232-237.

Estill, Laura/Levy, Michelle, Chapter 12: Evaluating Digital Remediations of Women's Manuscripts, in: *Digital Studies/Le champ numérique* 6 (2016), doi:10.16995/dscn.12.

European Commission, The European Cooperative Society (SCE), n. d., URL: https://ec.europa.eu/growth/sectors/social-economy/cooperatives/european-cooperative-society_en [last accessed: April 2, 2021].

Evans, Joel R./Mathur, Anil, The Value of Online Surveys: a Look Back and a Look Ahead, in: *Internet Research* 28 (4/2018), 854-887.

Firmani, Donatella, 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), 263-272.

Jain, Lakhmi C./Lazzerini, Beatrice (eds.), *Knowledge-Based Intelligent Techniques in Character Recognition*, Boca Raton, FL, 2020.

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, URL: https://www.gale.com/intl/c/china-and-the-modern-world-hongkong-britain-china [last accessed: April 2, 2021].

Google Arts and Culture, What is Fabricius?, 2020, URL: https://artsexperiments.withgoogle.com/fabricius/en [last accessed: April 2, 2021].

Google Cloud, Detect Handwriting in Images, AI and Machine Learning Products, 2021, URL: https://cloud.google.com/vision/docs/handwriting [last accessed: April 2, 2021].

Hauswedell, Tessa, et al., Of Global Reach yet of Situated Contexts: An Examination of the Implicit and Explicit Selection Criteria that Shape Digital Archives of Historical Newspapers, in: *Archival Science* 20 (2020), 139–165, doi:10.1007/s10502-020-09332-1.

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

Kahle, Philip, et al., *Deliverable 4.2, READ Platform And Service Maintenance, Deliverable Submitted To The European Commission*, 2017, URL: https://read.Transkribus.eu/wp-content/uploads/2017/12/D4.2.pdf [last accessed: April 2, 2021].

Krippendorff, Klaus, *Content Analysis: An Introduction to its Methodology*, Thousand Oaks, CA, 2018.

Krull, Florian/Muehlberger, Guenter/Terras, Melissa, Transkribus and IIIF: Beneficial Possibilities between Image Sharing and Handwritten Text Recognition Frameworks, in: *IIIF Conference - Göttingen, Germany, 24 June – 28 June 2019*, URL: https://iiif.io/event/2019/goettingen/program/26/#Transkribus-and-iiif-beneficial-possibilities-between-image-shar [last accessed: April 2, 2021].

Leiva, Luis A., et al., Evaluating an Interactive-Predictive Paradigm on Handwriting Transcription: A Case Study and Lessons Learned, in *2011 IEEE 35th Annual Computer Software and Applications Conference*, IEEE 2011, 610-617.

Maciel, Veronica, Database of 1918 Pandemic Deaths Inspires Answers for the Future, in: *The Daily Universe* 11.02.2021, URL: https://universe.byu.edu/2021/02/11/database-of-1918-pandemic-deaths-inspires-answers-for-the-future/ [last accessed: April 2, 2021].

Markauskaite, Lina/Goodyear, Peter, Epistemic Thinking, in: *Epistemic Fluency and Professional Education*, Dordrecht 2017, 167-194.

Milioni, Nikolina, *Automatic Transcription of Historical Documents: Transkribus as a Tool for Libraries, Archives and Scholars*, Master's Degree Project, Department of ALM, Uppsala Universitet, 2020, URL: http://www.diva-portal.org/smash/get/diva2:1437985/FULLTEXT01.pdf [last accessed: April 2, 2021].

Mitek, Handwriting Recognition, Mitek's Handwriting Recognition Software offers an Unprecedented Level of Access, 2020, URL: https://www.a2ia.com/en/handwriting-recognition [last accessed: April 2, 2021].

Monk, Homepage, 2020, https://www.ai.rug.nl/ lambert/Monk-collections-english.html [last accessed: April 2, 2021].

Moss, Michael/Thomas, David/Gollins, Tim, The Reconfiguration of the Archive as Data to be Mined, in: *Archivaria* 86 (2018), 118-151.

Muehlberger, Guenter, et al., Transforming Scholarship in the Archives through Handwritten Text Recognition: Transkribus as a Case Study, in: *Journal of Documentation* 75 (5/2019), 954-976, doi:10.1108/JD-07-2018-0114.

Nelson, Jean, UConn Library, School of Engineering to Expand Handwritten Text Recognition, in: *UConn Today* 09.07.2020, URL: https://today.uconn.edu/2020/07/uconn-library-school-engineering-expand-handwritten-text-recognition/# [last accessed: April 2, 2021].

Odumosu, Temi, The Crying Child: On Colonial Archives, Digitization, and Ethics of Care in the Cultural Commons, in: *Current Anthropology* 61 (S22/2020), S289-S302.

The Programming Historian, 2021, URL: https://programminghistorian.org [last accessed: April 2, 2021].

Quartex, Homepage, Features, 2020, URL: https://www.quartexcollections.com [last accessed: April 2, 2021].

READ-COOP, Transkribus User Conference 2018, URL: https://readcoop.eu/Transkribus-user-conference-2018/ [last accessed: April 2, 2021].

READ-COOP, FAQ, When Will the Payment Model for Transkribus Start?, 2020, URL: https://readcoop.eu/Transkribus/credits/faq/ [last accessed: April 2, 2021].

READ-COOP, Members of READ-COOP SCE, 2021, URL: https://readcoop.eu/members/ [last accessed: April 2, 2021].

READ-COOP, Public Models in Transkribus, 2021, URL: https://readcoop.eu/Transkribus/public-models/ [last accessed: April 2, 2021].

READ-COOP, Revolutionizing Access to Handwritten Documents, 2021, URL: https://readcoop.eu [last accessed: April 2, 2021].

Sánchez, Joan Andreu, et al., A Set of Benchmarks for Handwritten Text Recognition on Historical Documents, in: *Pattern Recognition* 94 (2019), 122-134.

Schafer, Burkhard, et al., A Fourth Law of Robotics? Copyright and the Law and Ethics of Machine Co-Production, in: *Artificial Intelligence and Law* 23 (3/2015), 217-240.

Schantz, Herbert, *The History of OCR, Optical Character Recognition*, Manchester Center, Vt., 1982.

Schön, Donald A., *The Reflective Practitioner: How Professionals Think In Action*, New York 1983.

Singh, Amarjot/Bacchuwar, Ketan/Bhasin, Akshay, A Survey of OCR Applications, in: *International Journal of Machine Learning and Computing* 2 (3/2012), 314-318, doi:10.7763/IJMLC.2012.V2.137.

Stroeker, Natasha/Vogels, René, Survey Report on Digitisation in European Cultural Heritage Institutions 2014, URL: https://www.egmus.eu/fileadmin/ENUMERATE/documents/ENUMERATE-Digitisation-Survey-2014.pdf [last accessed: April 2, 2021].

Terras, Melissa, Being The Other: Interdisciplinary Work in Computational Science and the Humanities, in Marilyn Deegan/Willard McCarthy (eds.), *Collaborative Research in the Digital Humanities*, Farnham 2012, 213-240.

Terras, Melissa M., The Rise of Digitization: An Overview, in: Ruth Rikowski (ed.), *Digitization Perspectives*, Leiden, Netherlands, 2011, 1-20.

Terras, Melissa M., The Role of the Library When Computers Can Read: Critically Adopting Handwritten Text Recognition (HTR) Technologies to Support Research, in: Amanda Wheatley/Sandy Hervieux (eds.), *The Rise of AI: Implications and Applications of Artificial Intelligence in Academic Libraries*, Atlanta, forthcoming 2021.

Transkribus platform, Official Facebook Group, 2021, URL: https://www.facebook.com/Transkribus/groups/ [last accessed: April 2, 2021].

Transribus Users, Facebook Group, 2021, URL: https://www.facebook.com/groups/614090738935143 [last accessed: April 2, 2021].

Wenger, Etienne, *Communities of Practice: Learning, Meaning, and Identity*, Cambridge 1999.