LISE JAILLANT (ED.)

ARCHIVES, ACCESS AND ARTIFICIAL INTELLIGENCE

WORKING WITH BORN-DIGITAL AND DIGITIZED ARCHIVAL COLLECTIONS

DIGITAL HUMANITIES RESEARCH
BIELEFELD UNIVERSITY PRESS

Lise Jaillant (ed.)
Archives, Access and Artificial Intelligence

Digital Humanities Research | Volume 2

Editorial

Digital Humanities is an evolving, cross cutting field within the humanities employing computer based methods. Research in this field, therefore, is an interdisciplinary endeavor that often involves researchers from the humanities as well as from computer science. This collaboration influences the methods applied as well as the theories underlying and informing research within those different fields. These implications need to be addressed according to the traditions of different humanities' disciplines. Therefore, the edition addresses all humanities disciplines in which digital methods are employed. **Digital Humanities Research** furthers publications from all those disciplines addressing the methodological and theoretical implications of the application of digital research in the humanities.
The series is edited by Silke Schwandt, Anne Baillot, Andreas Fickers, Tobias Hodel and Peter Stadler.

Lise Jaillant has a background in publishing history and digital humanities. She is an expert on issues of Open Access and privacy with a focus on archives of digital information. She was the first researcher to access the emails of the writer Ian McEwan at the Harry Ransom Center in Texas. Her work has been recognised by a British Academy Rising Star award. She is currently leading two externally-funded international networks on artificial intelligence applied to digital archives: the UK/Irish network AURA (www.aura-network.net) and the UK/US network AEOLIAN (www.aeolian-network.net). For more information, see: www.lisejaillant.com.

Lise Jaillant (ed.)

Archives, Access and Artificial Intelligence

Working with Born-Digital and Digitized Archival Collections

[transcript]

Open access funding for this volume was provided at 72% as part of Lise Jaillant's grant funded by the Arts and Humanities Research Council (Ref AH/R00773X/1) and at 28% as part of Tobias Hodel's OA fund at the University of Bern. We are very grateful to these institutions for their support.

Bibliographic information published by the Deutsche Nationalbibliothek
The Deutsche Nationalbibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data are available in the Internet at http://dnb.d-nb.de

This work is licensed under the Creative Commons Attribution-Non Commercial 4.0 (BY-NC) license, which means that the text may be may be remixed, build upon and be distributed, provided credit is given to the author, but may not be used for commercial purposes.
For details go to: http://creativecommons.org/licenses/by-nc/4.0/
Permission to use the text for commercial purposes can be obtained by contacting rights@transcript-publishing.com
Creative Commons license terms for re-use do not apply to any content (such as graphs, figures, photos, excerpts, etc.) not original to the Open Access publication and further permission may be required from the rights holder. The obligation to research and clear permission lies solely with the party re-using the material.

© 2022 transcript Verlag, Bielefeld
Published by Bielefeld University Press, an Imprint of transcript Verlag.

http://www.bielefeld-university-press.de

Cover layout: Maria Arndt, Bielefeld
Printed by Majuskel Medienproduktion GmbH, Wetzlar
Print-ISBN 978-3-8376-5584-1
PDF-ISBN 978-3-8394-5584-5
https://doi.org/10.14361/9783839455845
ISSN of series: 2747-5476
eISSN of series: 2749-1986

Printed on permanent acid-free text paper.

Contents

Introduction
Lise Jaillant, Loughborough University, UK 7

Chapter 1: Artificial Intelligence and Discovering the Digitized Photoarchive
X.Y. Han, Cornell University | Vardan Papyan, University of Toronto | Ellen Prokop, National Gallery of Art, Washington, DC | David L. Donoho, Stanford University | C. Richard Johnson, Jr., Cornell University. 29

Chapter 2: Web Archives and the Problem of Access: Prototyping a Researcher Dashboard for the UK Government Web Archive
Mark Bell, The National Archives, London | Tom Storrar, The National Archives, London | Jane Winters, School of Advanced Study, University of London. 61

Chapter 3: Design Thinking, UX and Born-digital Archives: Solving the Problem of Dark Archives Closed to Users
Lise Jaillant, Loughborough University, UK 83

Chapter 4: Towards Critically Addressable Data for Digital Library User Studies
Paul Gooding, University of Glasgow 109

Chapter 5: Reviewing the Reviewers: Training Neural Networks to Read Peer Review Reports
Martin Paul Eve, Birkbeck, University of London | Robert Gadie, University of the Arts, London | Victoria Odeniyi, University of the Arts, London | Shahina Parvin, Brandon University, Canada and Jahangirnagar University, Bangladesh. 131

Chapter 6: Supervised and Unsupervised: Approaches to Machine Learning for Textual Entities
Tobias Hodel, University of Bern 157

Chapter 7: Inviting AI into the Archives: The Reception of Handwritten Recognition Technology into Historical Manuscript Transcription
Melissa Terras, University of Edinburgh 179

AFTERWORD: Towards a new Discipline of Computational Archival Science (CAS)
Richard Marciano, University of Maryland 205

Authors (by order of appearance in the volume) 219

Introduction

Lise Jaillant, Loughborough University, UK

Digital archives are transforming the humanities and the sciences. Digitized collections of newspapers and books have pushed scholars to develop new, data-rich methods. Born-digital records ("items created and managed in digital form"[1]) are now better preserved and managed thanks to the development of open-access and commercial software. Digital humanities have moved from the fringe to the center of academia. Yet, the path from the appraisal of records to their analysis is far from smooth.

Cultural heritage organizations face at least three main challenges. First, the volume of digital archives makes it extremely difficult for archivists to assess records. Applying Artificial Intelligence (AI) and machine learning (ML) to archives is still at an experimental stage, but AI/ ML could become an integral part of archival processes.[2] To manage the sheer bulk and potential sensitivity of records, archivists will also rely on creators to help them make appraisal and selection decisions at the point of deposit.

Second, most born-digital collections are currently closed due to a wide range of reasons (including technical issues, copyright, and data protection). Regardless of whether archives are digital or not, archivists need to balance individual rights and the public interest in the context of the General Data Protection Regulation (GDPR) in Europe. Nobody would reasonably claim that all born-digital data should be unlocked and openly accessible. Yet, it is important to recognize that "dark" archives contain vast amounts of data essential to scholars – including email correspondence, drafts of manuscripts, digital photos and videos. Within current legal frameworks, making born-digital archives more accessible is an urgent priority to fully make sense of our cultural heritage.

---

1 Ricky Erway, Defining "Born Digital," OCLC Research, November 2010, URL: https://www.oclc.org/content/dam/research/activities/hiddencollections/borndigital.pdf [last accessed: Mar. 29, 2021].

2 Artificial Intelligence (AI) is a large concept designating the creation of intelligent machines that can simulate human thinking capability and behaviour. Machine Learning (ML) is an application or subset of AI that allows machines to learn from data without being programmed directly. In practice, the terms "AI" and "ML" are often used interchangeably.

Third, data science and AI are becoming essential tools, but very few scholars (particularly in the humanities) have been trained to master these research methods, a skills gap which in turn has an impact on the training we offer to students. This is a central topic of a recent White Paper from the Alan Turing Institute on *The Challenges and Prospects of the Intersection of Humanities and Data Science*. According to Barbara McGillivray and her co-authors, "the challenge here is to find a way to train and upskill humanities researchers in quantitative and computational methods, while at the same time incorporating the basic principles from these methods throughout undergraduate and graduate degrees, so humanities graduates are well equipped to lead projects but also potentially undertake careers in research software engineering and data science for arts and humanities." The authors suggest setting up basic courses in data science and software engineering to "offer the foundational skills to support humanists in having structured and informed conversations with computer scientists and data scientists needed in interdisciplinary projects."[3]

Automation, Access and AI are becoming keywords to decipher our history. We do not suffer from a lack of records, but from too many records – often locked away in dark archives. Access to dark archives is central but needs to be complemented with data-rich methodologies. How can we shed light on born-digital and digitized archives? How can we give greater access to archives currently closed to the public? What is the role of automation and AI? *Archives, Access and AI* addresses these central questions and explores crossovers between various disciplines to improve the discoverability, accessibility and use of born-digital archives and other cultural assets.

1. *Applying AI to Archives*

Archivists have commented on the digital revolution and its impact on archives for the past three decades. But it was not until the mid-2000s that the scholarship on digital preservation started growing. In addition to the preservation of digital materials,[4] commentators have examined the impact of the digital revolution on

3 Barbara McGillivray et al., *The Challenges and Prospects of the Intersection of Humanities and Data Science: A White Paper from The Alan Turing Institute*, London 2020, see 21, doi:10.6084/M9.FIGSHARE.12732164.

4 From 2005 to 2007, the UK funder Jisc supported the PARADIGM (Personal Archives Accessible in Digital Media) project, undertaken by the Bodleian Library in Oxford and the John Rylands Library in Manchester. The overall aim was to examine the issues in preserving personal digital materials, and to produce best-practice guidelines. In 2007, the Arts and Humanities Research Council (AHRC) funded the two-year Digital Lives project, led by the British Library in partnership with University College London and the University of Bristol.

appraisal.[5] Focusing on the preservation of born-digital and digitized records, or on the selection of these records, is not enough. Access and the production of new knowledge are issues that need to move to the center of the scholarly debate. In particular, Artificial Intelligence can be used by archivists to identify sensitive records, but also by researchers to process large amounts of digital archival data. AI has the potential to transform archives, but it also brings new challenges (including ethical challenges).

The closure of libraries, archives and museums due to the COVID-19 pandemic has highlighted the urgent need to make archives and cultural heritage materials accessible in digital form. Yet too many born-digital and digitized collections remain closed to researchers and other users due to privacy concerns, copyright and other issues. Born-digital archives are rarely accessible to users. For example, the archival emails of the writer Will Self at the British Library are not listed on the Finding Aid describing the collection, and they are not available to users either onsite or offsite. At a time when emails have largely replaced letters, this severely limits the amount of content openly accessible in archival collections. Even when digital data is publicly available (as in the case of web archives), users often need to physically travel to repositories to consult web pages. In the case of digitized collections, copyright can also be a major obstacle to access. For instance, copyright-protected texts are not available for download from HathiTrust, a not-for-profit collaborative of academic and research libraries preserving 17+ million digitized items (including around 61% not in the public domain).

The primary aim of the project was to develop ways to secure the personal archives of individuals in the digital era. In 2008, the Andrew Mellon foundation funded the futureArch project at the Bodleian Library to find solutions to the problem of born-digital but also hybrid archives (composed partly of paper materials). In particular, Bodleian Electronic Archives and Manuscripts (BEAM) worked on digital preservation infrastructure and researcher interfaces for hybrid archives. The 2010s saw the development of guidelines to preserve email archives. See Christopher J. Prom, *Preserving Email - DPC Technology Watch Report*, Digital Preservation Coalition 2011 (rev. ed. 2019).

5 See Ross Harvey/Dave Thompson, Automating the Appraisal of Digital Materials, in: *Library Hi Tech* 28 (2/2010), 313-322, doi:10.1108/07378831011047703; Kate Cumming/Anne Picot, Reinventing Appraisal, in: *Archives and Manuscripts* 42 (2/2014), 133-145, doi:10.1080/01576895.2014.926824; Anne Gilliland, Archival Appraisal: Practicing on Shifting Sands, in: Caroline Brown (ed.), *Archives and Recordkeeping: Theory into Practice*, London, 2014; William Vinh-Doyle, Appraising Email (Using Digital Forensics): Techniques and Challenges, in: *Archives and Manuscripts* 45 (1/2017), 18-30, doi:10.1080/01576895.2016.1270838; Victoria Sloyan, Born-Digital Archives at the Wellcome Library: Appraisal and Sensitivity Review of Two Hard Drives, in: *Archives and Records* 37 (1/2016), 20-36, doi:10.1080/23257962.2016.1144504; André Vellino et al., Assisting the Appraisal of E-Mail Records with Automatic Classification, in: *Records Management Journal* 26 (3/2016), 293-313, doi:10.1108/RMJ-02-2016-0006.

Archives, Access and AI is particularly timely. "Born-digital archives" are among the new research priorities highlighted in the UKRI (UK Research and Innovation) Infrastructure Roadmap Progress Report (2019): "The complexity of 'born-digital' archives […] and the challenges of archiving for discovery across many different formats raise significant questions about how to preserve, catalogue and make available these materials discoverable and accessible in a coherent fashion, in perpetuity." The report adds that "this is a fertile area for the arts, humanities and social sciences to explore natural crossovers with other research domains."[6] A recent UKRI report on UK's research and innovation infrastructure reiterates this priority on access to cultural collections using new technologies.[7]

Machine Learning applied to data in libraries and other cultural institutions is also at the center of current debates in the US, in Europe and elsewhere. Ryan Cordell recently wrote a Library of Congress report on ML,[8] which built on previous work in the same field – including Thomas Padilla's report on Data Science, ML and AI in libraries for the OCLC (Online Computer Library Center). In particular, Padilla gives examples of applications of AI/ ML to enhance descriptions of records at scale: "semantic metadata can be generated from video materials using computer vision; text material description can be enhanced via genre determination or full-text summarization using machine learning; audio material description can be enhanced using speech-to-text transcription; and previously unseen links can be created between research data assets that hold the potential to support unanticipated research questions."[9] *Archives, Access and AI* builds on this booming interest in new technologies applied to born-digital and digitized collections.

While many digital materials are completely "dark" and inaccessible to users, other records are open in theory, but difficult to find in practice. As Mark Bell, Tom Storrar and Jane Winters show in this edited collection, the UK Government Web Archive (UKGWA) is open to anyone with an internet connection, but discoverability is an issue for several reasons, including the inadequacy of keyword search to find relevant materials. The problem of searching huge amounts of records was also at the center of a 2019 article by Winters and Andrew Prescott. "With the rise of very large born-digital resources such as e-mail archives, Wikileak dumps and web

6 UKRI, Infrastructure Roadmap Progress Report, 2019, see 59.

7 UKRI, The UK's Research and Innovation Infrastructure: Opportunities to Grow our Capability, 2020, see 3, https://www.ukri.org/wp-content/uploads/2020/10/UKRI-201020-UKinfrastructure-opportunities-to-grow-our-capacity-FINAL.pdf [last accessed: Mar. 31, 2021].

8 Ryan Cordell, *Machine Learning + Libraries*, Washington D.C., 2020, https://labs.loc.gov/static/labs/work/reports/Cordell-LOC-ML-report.pdf?loclr=blogsig [last accessed: Mar. 29, 2021].

9 Thomas Padilla, *Responsible Operations: Data Science, Machine Learning, and AI in Libraries*, Dublin, OH, 2019, see 12, doi:10.25333/xk7z-9g97.

archives, the limitations of Google-type searching are becoming more evident."[10] For Winters and Prescott, it can be useful to re-examine the work of mid-twentieth century precursors of the web such as Vannevar Bush and Ted Nelson, who viewed the recording of links and information as the most effective way of processing very large quantities of information.

In addition to textual records, digital pictures can be very difficult to find without adequate metadata and cataloguing. The issue of discoverability is central for the Frick Collection in New York City, an art museum that houses the collection of industrialist Henry Clay Frick (1849 – 1919). Digital images and documentation for hundreds of thousands of art works have already been made freely available on the institution's online digital archive. An ambitious program of digitization is underway for the Frick Art Reference Library's Photoarchive – a research collection with more than 1.2 million reproductions of works of art in the Western tradition. As Ellen Prokop and her co-authors highlight in this edited volume, the rapid pace of digitization has led to a backlog of images that have been digitized but lack metadata and cataloguing information. In turn, this lack of information makes these digitized pictures unfindable and unusable. To solve this problem, the Frick Collection has worked with computer scientists to apply AI to the Photoarchive, automatically annotating images in the collection with the headings used in the archive's classification system. The team harnessed the power of Convolutional Neural Networks (CNNs), a deep learning technique used for classification and computer vision tasks – including image classification and face recognition.

Handwritten manuscripts also fall under the category of hidden collections, difficult to find, search and analyze at scale. While Optical Character Recognition (OCR) can decipher machine-generated texts to make them fully searchable, the technology does not work well for human-generated texts in digital form. Unlike typed characters, no handwritten characters are identical. Handwritten Text Recognition (HTR) is still at an early stage, but significant progress has been made in the past few years using Convolutional Neural Networks. In this edited volume, Tobias Hodel draws on the example of the European project Transkribus, a comprehensive platform for the digitization, AI-powered recognition, transcription and searching of historical documents. He shows that large amounts of data are used to train algorithms to recognize hand-written characters. As a consequence, "the resulting models are highly biased by the material they are trained on," as Hodel points out.

In recent projects on handwritten text recognition with deep learning, researchers often used the IAM Handwriting Dataset to train their models. This dataset contains 115K+ English words by 600+ authors. Since deep learning model

10 Jane Winters/Andrew Prescott, Negotiating the Born-digital: a Problem of Search, in: *Archives and Manuscripts* 47 (3/2019), 391-403, doi:10.1080/01576895.2019.1640753.

need at least 10^5 - 10^6 training examples in order to perform well, the IAM Handwriting Dataset meets those requirements. Using large sets of data is a prerequisite for HTR, but it can also lead to a "cycle of bias" as Hodel puts it. Understanding the source materials and the methods used are essential to engage critically with HTR and more broadly, with any AI-powered techniques applied to archival collections.

The case studies featured in this book will be useful to our intended audience – archivists, digital humanists and social scientists, computer scientists and anyone else interested in the issues faced by archival collections in the twenty-first century. In their 2019 article "More Human than Human? Artificial Intelligence in the Archive," Gregory Rolan and his colleagues note the "barriers to the uptake of AI technology for recordkeeping knowledge work." One explanation is "a lack of compelling case studies": "there are not many real-world examples within the academic or professional literature."[11] Intertwining practical case studies and theoretical insights, the edited collection aims to fill this gap in scholarship. It presents advanced work in Archives and AI, while also "zooming out" and looking at the big picture.

To solve the problem of access to digital archives, cross-disciplinary collaborations are absolutely essential. The big challenges of our time – from global warming to social inequalities – cannot be solved within a single discipline. The same applies to the challenge of "dark" archives. We cannot expect archivists or digital humanities to find a magical solution that will instantly make digital records more accessible. And it is not enough to encourage collaborations between disciplines that are very close (for example, history and literary studies). Instead, we need to take a radical step outside our comfort zone and set up collaborations across disciplines that seldom talk to each other. This is the main goal of the AURA network (Archives in the United Kingdom/ Republic of Ireland and AI), funded by the Arts and Humanities Research Council in the UK and the Irish Research Council in 2020-2021.[12]

Led by a management team with expertise in digital humanities, archives and computer science, AURA organized three workshops to bring people together and offer a forum for discussion and future collaborations: "Open Data versus Privacy" (Workshop 1); "AI and Archives: Current Challenges and Prospects of Born-digital archives" (Workshop 2); "AI and Archives: What comes next?" (Workshop 3). While

11 Gregory Rolan et al., More Human than Human? Artificial Intelligence in the Archive, in: *Archives and Manuscripts* 47 (2/2019), 179-203, see 186, doi:10.1080/01576895.2018.1502088. See also: Basma Makhlouf Shabou et al., Algorithmic Methods to Explore the Automation of the Appraisal of Structured and Unstructured Digital Data, in: *Records Management Journal* 30 (2/2020), 175-200, doi:10.1108/RMJ-09-2019-0049; Tim Hutchinson, Natural Language Processing and Machine Learning as Practical Toolsets for Archival Processing, in: *Records Management Journal* 30 (2/2020), 155-174, doi:10.1108/RMJ-09-2019-0055.

12 www.aura-network.net[last accessed: Mar. 29, 2021].

AURA focuses mostly on the UK and Ireland, AEOLIAN (Artificial Intelligence for Cultural Organizations) strengthens connections between British and American partners. Funded by the AHRC in the UK and the National Endowment for the Humanities in the US, AEOLIAN runs from 2021 to 2023 and consists in a series of meetings and case studies that bring together a team of experts to develop new approaches to improving access to and use of digital archives.[13]

Other initiatives have created partnerships between experts in various fields, benefiting cultural institutions in general and archival collections in particular. The AHRC-funded Computational Archival Science (CAS) research network and the Advanced Information Collaboratory explore the conjunction of big data methods and technologies with archival practice. In Germany, the Leipzig Computational Humanities group includes experts in the humanities and computer science. In the USA, the HathiTrust Research Centre and Stanford Literary Lab also foster computational research in the humanities. The global spread of these initiatives show that cross-disciplinary collaborations are not enough: we also need to bring the best people together, independently of their nationalities and professional affiliations.

Published by a German press and written by an international team of contributors, the six chapters of this book feature examples of collaborations between researchers in computer science and engineering, archivists and scholars in the digital humanities. Often, these collaborations are made possible thanks to external funding and are hosted in large cultural institutions in world cities. Applying AI to archives would be difficult for a small archival collection in, say, Loughborough (a market town in the North of Britain). The combination of prestigious metropolitan institutions, ground-breaking technology, and advanced expertise can be intimidating. But this does not have to be the case. AI is still a very imperfect technology, with applications to the archival sector at an early, experimental stage.

In a 2019 article entitled "Artificial Intelligence – the Revolution hasn't happened yet," Michael Jordan notes that when the term "AI" was coined in the late 1950s, it referred to the ambition to build hardware and software possessing human-level intelligence. He uses the phrase "human-imitative AI" to refer to this aspiration to create an entity that would resemble humans. AI was meant to focus on "the high-level or cognitive capability of humans to reason and to think," Jordan points out. "Sixty years later, however, high-level reasoning and thought remain elusive. The developments now being called AI arose mostly in the engineering fields associated with low-level pattern recognition and movement control, as well as in the field of statistics, the discipline focused on finding patterns in data and on making well-founded predictions, tests of hypotheses, and decisions."[14]

---

13 www.aeolian-network.net[last accessed: Mar. 29, 2021].

14 Michael I. Jordan, Artificial Intelligence – The Revolution Hasn't Happened Yet, in: *Harvard Data Science Review* 1 (1/2019), 1-9. doi:10.1162/99608f92.f06c6e61.

If we look at the specific case of archival collections, it is certainly true that AI has been used for low-level tasks: identifying sensitive information such as credit card numbers in emails, tagging pictures, transcribing handwriting for examples. These tasks could be done by any normal teenager with a minimum of training. The value of AI is not its ability to perform complex high-level tasks that require contextualization, theorization or creativity. Instead, the value of AI comes from its capacity to process huge amounts of data very rapidly – something that no human can do single-handedly.

Michael Jordan argues that success in human-imitative AI has been quite limited. We are very far from having artificially intelligent systems that can compete with humans at the higher levels of intelligence. A focus on human-imitative AI can distract us from a key challenge of our times: making sure that AI works for humans, that it makes our human lives better rather than worse. For Jordan, we are witnessing the creation of a new discipline: *human-centric engineering*. "Whereas civil engineering and chemical engineering built upon physics and chemistry, this new engineering discipline will build on ideas that the preceding century gave substance to, such as information, algorithm, data, uncertainty, computing, inference, and optimization."[15] Since the new discipline will focus on data from and about humans, it will need the perspectives of social scientists and humanists.

Applied to archival collections, *human-centric engineering* will focus on building artifacts and designing processes to make archives more accessible. The new discipline will bring together not only engineers, data scientists and computer scientists, but also archivists and scholars in the humanities and social sciences. Working collaboratively, these interdisciplinary teams will pay close attention to issues of privacy and biases. Unlocking archives should not come at any price, and the hype surrounding Open Data and AI must not distract us from the need to comply with data protection regulations and to address bias associated with black-box algorithms. In short, we need to mitigate the risks of malicious AI in our quest to unlock archival records.

2. *The Threat of Dark AI*

Using AI to make dark archives accessible is risky: sensitive information in government archives could inadvertently be released and fall into the hands of criminals; private information in email archives could be leaked, leading to distress for individuals and breaches of data protection laws; pictures in digital collections could be mis-labelled, leading to embarrassment and damage to the cultural institution's "brand." Automatic image labelling has a long history of producing embarrassing

15 Jordan, Artificial Intelligence, 3.

results – in 2015, Google Photos labelled a picture of two black people as "gorillas." Google Maps and Flickr have also suffered from race-related problems. Training datasets that contain biases result in problematic, sometimes appalling outcomes. In 2016, after engaging with users on Twitter, a Microsoft chat box began sharing racist, genocidal and misogynistic messages. And in 2020, the researcher Timnit Gebru said she was fired from Google after co-writing a paper on AI-generated language, which replicates unsavory biases found in online text. To unlock dark archives, we cannot rely on dark AI – defined as AI that is making human lives worse, not better.

Bias in large datasets is at the center of Thomas Padilla's and Ryan Cordell's recent reports on ML applied to libraries and archives. Cordell gives the example of the *Chronicling America* newspaper collection, a searchable database of US newspapers with descriptive information and selected digitized pages. *Chronicling America* is produced by the National Digital Newspaper Program (NDNP), a partnership between the National Endowment for the Humanities and the Library of Congress. The NDNP relies on institutions in each state to select and digitize approximately 100,000 newspaper pages representing that state's history and geographic coverage. The website gives the impression that digitized collections reflect the diversity of each state:

> Participants are expected to digitize primarily from microfilm holdings for reasons of efficiency and cost, encouraging selection of technically suitable film, bibliographic completeness, diversity and "orphaned" newspapers (newspapers that have ceased publication and lack active ownership) in order to decrease the likelihood of duplicative digitization by other organizations.[16]

Yet, far from reflecting diversity, "the data skews to newspapers serving the majority," Cordell argues.[17] For example, the collection privileges newspapers read by white middle-class audiences in the nineteenth century, rather than black and other minority-run papers. As Benjamin Fagan points out, *Chronicling America* does currently list forty-six black newspapers in its digital archive (c. 2.5% of a total of 1,799), but all were printed in 1865 or later.[18] There are no black newspapers among the digital copies of 215 newspapers published before 1865. NDNP participants prioritized geographic spread, inadvertently deemphasizing racial representation.

16 https://chroniclingamerica.loc.gov/about/[last accessed: Mar. 29, 2021].

17 Cordell, *Machine Learning + Libraries*, 14.

18 Benjamin Fagan, Chronicling White America, in: *American Periodicals: A Journal of History & Criticism* 26 (1/2016), 10-13, see 11, https://muse.jhu.edu/article/613375 [last accessed: Mar. 29, 2021].

"Consequently any ML projects based on *Chronicling America* will reflect those same oversights and exclusions," writes Cordell.[19]

Bias in large newspaper datasets is also at the center of the *Oceanic Exchanges* project (2017-2019). Melodee Beals and her co-investigators argue that the national focus of digitized newspapers collections obscures the fact that international news exchange was central to the nineteenth-century press. ML projects based on national newspapers risk missing important international links with other papers. To mitigate these risks, *Oceanic Exchanges* produced the *Atlas of Digitised Newspapers and Metadata*, an open access guide to digitized newspapers around the world. Highlighting the history of digitized newspapers and digitization choices, the atlas examines metadata available in these collections. In particular, it "explores how machine-readable information about an issue, volume, page, and author is stored in the digital file alongside the raw content or text."[20] This project sheds light on information that is often hidden. It invites researchers and other users to see digitized newspaper datasets as human constructions, rather than unproblematic data.

Managing bias is essential for libraries and archival collections. For Thomas Padilla, eliminating bias entirely is not an option: in the hope of cleaning the dataset, elimination risks introducing more bias. Instead, Padilla proposes a bias management strategy to reflect on and integrate bias within the cultural organization. One of the recommendations is to:

> Hold symposia focused on surfacing historic and contemporary approaches to managing bias with an explicit social and technical focus. The symposium should gather contributions from individuals working across library organizations and focus critical attention on the challenges libraries faced in managing bias while adopting technologies like computation, the internet, and currently with data science, machine learning, and AI. Findings and potential next steps should be published openly.[21]

Symposia and other communication strategies would promote self-conscious attention and criticism at every stage of an ML project. But, as Cordell argues, this may not be enough. "To create ML projects that reflect data justice, […] libraries cannot pretend to be neutral or objective in relationship to race, class, gender, sexuality, or culture, but instead must consciously strive to forefront marginalised voices."[22]

The notion of data justice is closely related to the black box problem, explain Catherine D'Ignazio and Lauren Klein in *Data Feminism*. Machine learning algo-

19 Cordell, *Machine Learning + Libraries*, 14.

20 https://www.digitisednewspapers.net/dhawards/[last accessed: Mar. 29, 2021].

21 Padilla, *Responsible Operations*, 10.

22 Cordell, *Machine Learning + Libraries*, 15.

rithms are so complex that they are often described as incomprehensible black boxes: you put data in, and you get something out, but what happens inside the box is a mystery. For D'Ignazio and Klein, data justice aims to "ensure that past inequities are not distilled into black-boxed algorithms." While terms such as *ethics* "locate the source of the problem in individuals or technical systems," *justice* acknowledges "structural power differentials" and works "toward dismantling them."[23] Data justice seeks to address the structural inequalities in the training datasets that lead algorithms to produce less-favorable outcomes for women and ethnic minorities.

In *Algorithms of Oppression*, Safiya Umoja Noble argues that Google search algorithms privilege white people and discriminate against ethnic minorities, particularly women. These algorithms in turn reinforce existing prejudices against women of color with search results presenting black women as "angry" or "sassy." Noble points out that the search operations are invisible: users of Google and other search engines have no access to the algorithms and deep machine learning systems, developed to index masses of information and move some to the first page of results. While the process of searching is hidden, the information users see on their screens becomes a reality and has an impact on decision making. For Noble, Artificial Intelligence will become a major human rights issue in the twenty-first century. "We are only beginning to understand the long-term consequences of these decision-making tools in both masking and deepening social inequality."[24]

Noble, Cordell and others have stressed the role that libraries and other cultural institutions can play to balance the power of tech giants. For Noble, "the public is increasingly reliant on search engines in lieu of libraries, librarians, teachers, researchers, and other knowledge keepers and resources." Yet, it does not make sense "to outsource all of our knowledge needs to commercial search engines" that will return biased results.[25] Librarians and other knowledge keepers can mitigate bias and offer an alternative to commercial interests. Likewise, Cordell argues that "by centering ethics, transparency, diversity, privacy and inclusion, libraries can take a leadership role in one of the central cultural debates of the twenty-first century."[26] Many entrepreneurs in the tech industry still live by Facebook's early motto: "Move fast and break things." But the same motto cannot apply to cultural institutions that value continuity over disruption.

---

23 Catherine D'Ignazio/Lauren F. Klein, *Data Feminism*, Cambridge, MA, 2020, https://data-feminism.mitpress.mit.edu/ [last accessed: Mar. 29, 2021].

24 Safiya Umoja Noble, *Algorithms of Oppression: How Search Engines Reinforce Racism*, New York City, 2018, 1.

25 Noble, *Algorithms of Oppression*, 16.

26 Cordell, *Machine Learning + Libraries*, 1.

> As institutions of memory and community, libraries cannot be bound to destructive ideologies of technological implementation but must instead model alternative engagements with ML focused on building rather than breaking. Libraries can become ideal sites for cultivating responsible and responsive ML, as that term describes a constellation of technologies that explain data within the contexts of its collection, aggregation, and association.[27]

This debate over the place of libraries in the digital age is not new. In his 1994 article "Electronic Records, Paper Minds," Terry Cook reminded information professionals that their role was to guide users from masses of information onto specific knowledge. At a time of big digital data, this role had become particularly challenging. If librarians and other knowledge keepers failed, they will "be replaced by software packages that can handle facts, and data, and information very efficiently, without any mediation by archivists or anyone else."[28]

In January 1994, shortly before Cook's article appeared, two electrical engineering graduate students at Stanford University created a guide to the World Wide Web. Their website, which was named Yahoo in March 1994, was a directory of other websites, organized in a hierarchy, as opposed to a searchable index of pages. Commercial search engines then evolved to rank results by counting how many times the search terms appeared on the page. The creation of Google in 1998 marked a turning point, with the development algorithms that analyzed links and relationships between websites to determine their importance. By the late 1990s and the massification of internet access, people increasingly relied on commercial search engines to find information, rather than on traditional knowledge keepers.

Tech giants took the advantage over libraries three decades ago and have remained at the forefront of information searching. It is not surprising that the field of AI/ ML applied to cultural institutions is so heavily invested by Google and the like. For the "LIFE Tags" project, Google organized over 4 million images from the LIFE magazine archives into an interactive encyclopedia. Machine learning automatically applied tags to digitized images, which greatly simplified the archiving work since the archive spans approximately 1,800 meters in three different warehouses. LIFE Tags allows users to easily navigate the magazine archives using keywords. For this project, Google drew on a deep neural network used in Google Photo search that has been trained on millions of images. As we have seen, this technology is not neutral: the training dataset can be biased, which in turn can lead to problematic labels (including racist labels).

---

27 Cordell, *Machine Learning + Libraries*, 2.

28 Terry Cook, Electronic Records, Paper Minds: The Revolution in Information Management and Archives in the Post-Custodial and Post-Modernist Era, in: *Archives and Manuscripts* 22 (2/1994), 300-328, see 306.

Google also worked with MoMA (Museum of Modern Art) in New York City, to automatically identify art works in archival photos of exhibitions. Since 1929, MoMA has kept thousands of photos of its exhibitions. However, these images were difficult to find and use as they lacked metadata and other information on the works displayed in exhibitions. Google's algorithms automatically identified 27,000 works of art and made MoMA collections more accessible. Again, the issue comes from black-boxed algorithms: the museum has no control over the algorithms used by Google, and over the results produced by these algorithms.

Instead of a black box model controlled by tech giants, a more open model is possible.

Fig 0.1: Implementing an open model for ML projects in libraries and cultural organizations. Courtesy of the author.

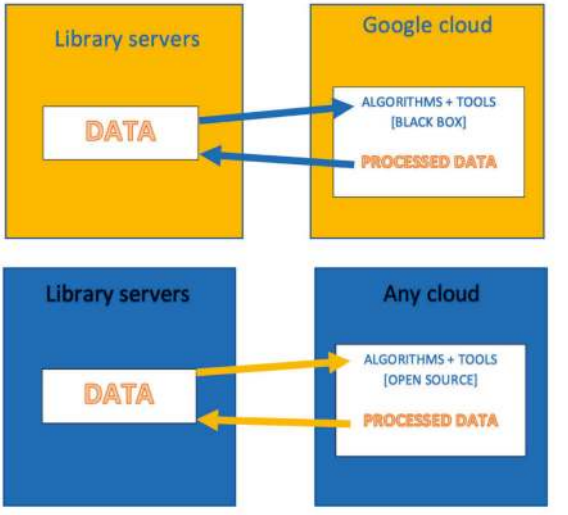


Fig 0.1 gives a simplified account of what goes on when libraries and cultural organizations partner with Google and the like. They send their data to the tech partner and receive processed data in return. What happens in between is largely controlled by the tech company. It is possible to take back control. Instead of relying on tech giants to use their own algorithms (such as Google's Image Content-based Annotation algorithm to generate labels based on image pixels), cultural organizations could work with multi-disciplinary teams to generate their own algorithms. There is no need to re-invent the wheel: many open-source AI software are

easily available, including tools released by Google, Microsoft, Facebook and other tech companies. Teams composed of librarians and archivists, humanities scholars, computer scientists and software engineers would be well-equipped to conduct ML projects currently outsourced to tech giants. There are obvious obstacles, including funding and the availability of expertise, but these concerns should be viewed as challenges rather than showstoppers. Funding bodies are increasingly pushing for cross-disciplinary projects crossing the divide between the sciences and the humanities. "Reuniting the humanities with sciences might protect their future," notes a recent *Economist* article on Digital Humanities.[29]

3. *AI for Good*

What does "AI for good" mean for archival collections? Combined with other technologies, AI has the potential to make digital archives more comprehensive and to guarantee the authenticity of records. Let's start with the issue of comprehensiveness. If you are buying a house, you do not want to request a property title and discover that the record has disappeared from the archive, or never was there in the first place. Yet, anyone who has done archival work knows that archives are rarely complete. Documents are thrown away deliberately or by mistake, either by the record creator or by the archivist. Appraisal, i.e. the selection of records, is central aspect of the archival process. In *Modern Archives: Principles and Techniques* (1956), the American archivist T. R. Schellenberg argued that a record has "primary value" to the creator but may also have "secondary value" (as evidence or information) to historians and future users of the records outside of the originating institution. "Archivists should be empowered to review all records that government agencies propose to destroy."[30] For Schellenberg, only records that have secondary value over the long term should be kept in archival collections.

Reviewing records manually has become almost impossible in the digital age. Archivists have to deal with huge amounts of records, but also records that are scattered in various places – both within and outside the creator's organizational system. The deployment of cheap cloud technologies, smartphones and other mobile devices have led to a rise in "shadow IT," i.e. IT systems deployed and supported by providers outside the organization's central IT and by definition not aligned to the central IT strategy and direction. For example, civil servants may be tempted to use private emails accounts, WhatsApp, Skype, Facebook, Twitter and other platforms

29 "How Data Analysis Can Enrich the Liberal Arts," in: *Economist*, 19.12.2020, https://www.economist.com/christmas-specials/2020/12/19/how-data-analysis-can-enrich-the-liberal-arts [last accessed: Mar. 29, 2021].

30 Theodore R. Schellenberg, *Modern Archives: Principles and Techniques*, Chicago 1956, see 32.

to easily share information with their colleagues, instead of using government IT tools. Not only does shadow IT increase the risk of data leaks, but it also makes it impossible to comply with record management obligations.

In the UK, these obligations are outlined in the Freedom of Information Act 2000, the Public Records Act 1958, and the Data Protection Act 2018 following the introduction of the GDPR in Europe. In a nutshell, the Freedom of Information Act provides public access to information held by public authorities. To foster a culture of open government accountable to citizens, disclosure of information should be the default. In other words, information should be kept private only when there is a good reason, and it is permitted by the FOI Act. Moreover, the amended Public Records Act states that records of UK central government selected for permanent preservation shall be transferred not later than twenty years after their creation to The National Archives (rather than the previous thirty years). Finally, the Data Protection Act gives people more control over use of their data and provides them with new rights to move or delete personal data.

For Knowledge and Information Management (KIM) government professionals, shadow IT is a major problem. To respond to Freedom of Information requests, Data Protection subject access and public inquiries, KIM professionals need access to the relevant records. If these records are scattered outside official channels, they often become undiscoverable and inaccessible – making government vulnerable to accusations of secrecy and malpractice that can potentially lead to prosecution.

Even when information remains within government IT systems, it can be extremely difficult to find. This issue was at the center of the *Better Information for Better Government* report that the UK Cabinet Office released in 2017. "While little information has been lost altogether, much of what has accumulated over the past fifteen to twenty years is poorly organised, scattered across different systems and almost impossible to search effectively."[31] At Year 7 following their creation, archival materials are transferred to an internal archive where they stay for thirteen years, before their transfer to The National Archives at Year 20. "We need to know what's here, we need to be able to find it," one KIM professional points out. "When we take it in our archive, we need to be able to index it, to classify it and catalogue it properly" in order to find information easily and respond to possible Freedom of Information requests and other access requests.[32]

Artificial Intelligence has a role to play in bringing scattered records together, making them findable and usable. However, it is a controversial role since AI makes

31 Cabinet Office (UK), *Better Information for Better Government*, London 2017, https://www.gov.uk/government/publications/better-information-for-better-government [last accessed: Mar. 29, 2021].

32 Conversation with author, Nov. 2, 2020.

an intervention on the records. *Respect des fonds* is a key principle in archival theory that goes back to at least the nineteenth century.[33] According to this principle, archival materials need to be grouped according to their *fonds* or origins. Archivists should maintain records using the creator's organizational system instead of imposing a new order. Established at a time of growing archival records, the principle allowed archivists to save time and avoid any attempts to re-arrange documents created by the same agency, individual, or organization. These attempts would be at best futile, and at worst would tamper with the collection.

In the digital age, the purist and non-interventionist viewpoint of *respect des fonds* has come under attack. Is it better for archivists to passively accept digital information acquired into archives and store it essentially as it comes, without any modification? For some Knowledge and Information management professionals, the principle is no longer adequate at a time when digital information created by organizations is inherently chaotic and unorganized. To easily find information and respond to requests, it is necessary to (re)organise records by grouping, metadata tagging and the like, and in so doing actively interpret them and construct them into thematic archives.

On a small scale, (re)-organizing archival collections can be done manually. One archivist built a digital collection on the 2012 Olympics in London by directly approaching people involved in the preparation and delivery of the events and asking them to send their digital records. The resulting thematic archive does not conform to *respect des fonds* since it was actively created by the archivist rather than passively received. But it served an important purpose: bringing scattered documents together and making them easy to find, search and use.

According to this more interventionist principle, AI's role would be to automatically add metadata, extract names and topics. As Tobias Hodel explains in this edited collection, topic modeling is a statistical method used in machine learning and Natural Language Processing to discover clusters of words or "topics" that occur in a dataset. The approach uses unsupervised machine learning: algorithms identify what words appear together frequently, resulting in the extraction of topics. AI would not only improve the discoverability of records, but it would also make them more accessible.

By automatically identifying sensitive records, AI would allow non-sensitive records to be opened up and made available to researchers and other users. Automatic sensitivity review is still at early stage, but it has the potential to shed light on "dark" archives. This is particularly important for departments that deal with a lot of sensitive and confidential information – such as the UK Cabinet Office as

33 Michel Duchein, Theoretical Principles and Practical Problems of Respect Des Fonds in Archival Science, in: *Archivaria* (16/1983), 64-82.

opposed to, say, the Department for Education (DfE). The risk appetite of the Cabinet Office is very low because many of its records are very secret and sensitive. In contrast, because DfE's policy making is about schools and education and is very public, the risk of leaking sensitive information is much lower.

Because AI is deployed on huge amounts of data, it would result in vast re-organized archives free of sensitive materials. UK central government departments currently only send a very small proportion of their records to The National Archives (around 5%). Is it advisable to continue this approach? Or would it be better to exploit the potential of digital and AI to make available a larger corpus? The government needs to avoid the release of sensitive material, but it also needs to encourage access and transparency – principles that are at the heart of the Freedom of Information Act 2000 and the National Data Strategy (NDS) policy paper released in September 2020, which aims to "unlock the power of data for the UK." "Data is a non-depletable resource in theory, but its use is limited by barriers to its access – such as when data is hoarded, when access rights are unclear or when organisations do not make good use of the data they already have," declares the NDS.[34]

Although Artificial intelligence can be used to re-organize the archive and add metadata, the technology is fraught with risks. When record creators work with commercial partners to apply AI to their archives, they rarely invite external archivists to the table. Yet, selected records will eventually end up in the external archive. Does the archivist have a role to play in defining the algorithms and code needed for decision making? asked Anthea Seles of the International Council on Archives (ICA) in 2019.[35] She outlined four main challenges and issues. First, archivists will be responsible for the conservation of these algorithms in archives used by historians and other users. Second, archivists are not currently considered stakeholders in discussions related to the development and implementation of AI technologies. Third, archivists currently do not have the capacity and skills to play their role as advisors on good records management to ensure the longevity and sustainability of these new archival documents. Fourth, archivists will need to understand not only how to advise on the conservation of AI algorithms, but also how to deal with important ethical issues – including the black box issue. Even with all the necessary elements are kept, it is often difficult to understand how an algorithm came to a decision. When AI is applied to archives, the risk is to bias the historical document and consequently history as well as our collective memory.

Archivists rightly ask for a seat at the table and an opportunity to shape the discussions on AI applied to archives. "Automation is no longer a choice, it is a necessity but that does not mean that the archivist (the human) is not relevant in the

34 https://www.gov.uk/government/publications/uk-national-data-strategy/national-data-strategy[last accessed: Mar. 29, 2021].

35 Anthea Seles, Norwegian Triennial Archival Conference, April 8, 2019.

process," Seles points out. Advocating for algorithmic transparency and accountability are becoming key roles for archivists. In 2017, the Association for Computing Machinery issued a statement outlining seven principles to make public policy more transparent and accountable. First, the principle of *awareness* implies that all stakeholders should be aware of the bias problem and potential harm associated with automation. Second, a right to *access and redress* should allow individuals and groups that are adversely affected by algorithmically informed decisions to question these decisions. Third, the principle of *accountability* would hold institutions responsible for decisions made by the algorithms that they use. Fourth, producing *explanations* should be required from institutions that use algorithmic decision-making, in order to understand how specific decisions are made. Fifth, *data provenance* should be documented, including a description of the way in which the training data was collected. Sixth, the principle of *auditability* should encourage institutions to record models, algorithms, data, and decisions so that they can be audited. Seventh, the principles of *validation and testing* should encourage institutions to validate and document their models, but also to routinely perform tests to make sure that models do not generate harm.

Like many humanities scholars, archivists often fear that they lack the skills and knowledge to productively participate in these debates. But there is no need to be a computer scientist or software engineer to have a productive discussion on AI, transparency and accountability. The private sector offers a valuable model to bring together professionals with various specialisms and expertise. When an IT consulting company is contracted to implement a new system or deliver a solution to a data problem, the first thing they do is to bring together a committee with representatives of the client's internal IT and operational services. The committee defines standards to apply to the data, and starts with a pilot (for example, a small amount of data to process). Following an agile process, consultants then work to deliver the project in close discussion with their client's stakeholders from various teams. In the case of government archives, it should be possible to bring to the same table AI specialists (either consultants or internal experts), government record creators and experts from different services including KIM professionals, and external archivists. A central objective would be to improve the organization and comprehensiveness of the digital archive, while also pushing for algorithmic transparency and accountability.

In addition to comprehensiveness, authenticity is central to archival collections. One of the most important roles of archives in our societies is to preserve authentic documents, before making them available to users. If you are buying a house, you will need to access authentic records about previous ownership. And if a historian consults government records of the nineteenth century, they need to trust that the records have not been tampered with. Guaranteeing the integrity of digital records is a key objective for The National Archives UK and other institutions. As

technology evolves and software used to read certain formats becomes obsolete, digital records often need to be ported from one format to another. Although these records are easy to copy and modify, their content must remain unaltered while stored in the archive.

The ARCHANGEL project (2017-2019) addressed the challenges around trust, integrity and authenticity of born-digital archival materials by exploring the possibilities offered by blockchain and machine learning.[36] Blockchain is the technology that underpins Bitcoin and other cryptocurrencies, but it has the potential for application to other sectors. With blockchains, data can be added to the chain, but it cannot be overwritten, amended or deleted. A blockchain is therefore a growing list of records, called blocks – with each block containing a cryptographic hash[37] of the previous block, a timestamp and other information. Moreover, the technology is distributed, i.e. no central organization has sole possession or control over of the data. Finally, it is transparent, with all entries in the chain visible to all trusted members who have a copy. Combined with machine learning, blockchain offers a digital fingerprint for archival materials, making it possible to verify their authenticity.

ARCHANGEL prototyped the creation of hashes using machine learning methods, particularly for image and video records. ML can identify the causes of glitches and noise in these records – which could either be caused by transcoding and format-shifting, or by any undesirable process, such as corruption of the files in storage or tampering. Machine learning complements the ARCHANGEL blockchain, which enables archival collections to upload metadata that uniquely identified specific records. In the case of sensitive records, metadata can itself be confidential and sensitive, making it inappropriate to add it to the blockchain. One solution is to add an archival reference and the record's checksum instead (a unique computer-generated string that changes if the file is altered). That data is then sealed into a block that cannot be changed or deleted without detection. Finally, a copy of the data is shared with all trusted members of the network. The ARCHANGEL example shows that AI can work *for* rather than *against* archival collections.

4. *Structure*

Archives, Access and AI is organized in two parts, with three chapters in each section. The first part on "Selection, Appraisal, Discoverability and Access" starts with the example of AI applied to the Photoarchive at the Frick Art Reference Library

36 http://www.archangel.ac.uk/ [last accessed: Mar 29 2021]

37 A cryptographic hash is an algorithm that takes an arbitrary block of data and returns a fixed-size bit string.

in New York. The project led to the automatic creation of metadata that improved the discoverability of the archive. This collection has been made more accessible and usable thanks to a cross-disciplinary team with expertise in computer science, art history and other fields. Chapter 2 on web archives also moves away from single disciplines to solve the problems associated with large-scale digital collections. Bringing together archive professionals and a digital humanist, the project aims to make born-digital archives more accessible by prototyping a researcher dashboard for the UK government web archive. Chapter 3 focuses on design thinking, a human-centered method to solving business and social problems. It argues that design thinking is a productive way to solve the problems of *access* and *use* of archival collections in the digital age. Researchers should work closely with archivists to shape access policies that will facilitate the use of AI and other innovative methodologies.

The second part on "Using the Archives: AI and New Knowledge" starts with a chapter on digital library user studies. Investigating the impact of e-legal deposit on UK academic deposit libraries in Chapter 4, Paul Gooding argues that transparent workflows and data documentation should be central to user studies. Currently, library user studies often rely on tools such as Google Analytics and suffer from a black box problem. The quality of library patron data is also an issue, with potentially biased data leading to problematic results. In Chapter 5, Martin Paul Eve and his co-authors examine another large-scale dataset: academic peer review reports in the sciences. The confidentiality of these reports and the difficulties of access make research extremely complicated. Neural networks can be used to machine-read these archives and make them more accessible, but the process is fraught with difficulties. In Chapter 6, Tobias Hodel focuses on Handwritten Text Recognition (HTR). Supervised deep learning approaches have led to astonishing results in deciphering handwriting, but also to new problems – including in terms of transparency and accountability. Hodel also presents an overview of unsupervised machine learning, including topic modeling used on huge amounts of textual data. Continuing the discussion on HTR in Chapter 7, Melissa Terras shows that this technology is now transforming access to our written past. Drawing on a survey of users of HTR on the Transkribus platform, Terras highlights issues raised when inviting machine learning into historical archives. Transcriptions generated by HTR will require new approaches to both history and public engagement, Terras argues, before providing recommendations on how to best support the community applying HTR to cultural heritage materials. Finally, an afterword by Richard Marciano offers further thoughts on the intersection of technology and archives to produce new ways of preserving and making accessible our collective past.

Bibliography

Cabinet Office (UK), *Better Information for Better Government*, London 2017, https://www.gov.uk/government/publications/better-information-for-better-government [last accessed: Mar. 29, 2021].

Cook, Terry, Electronic Records, Paper Minds: The Revolution in Information Management and Archives in the Post-Custodial and Post-Modernist Era, in: *Archives and Manuscripts* 22 (2/1994), 300-328.

Cordell, Ryan, *Machine Learning + Libraries*, Washington D.C., 2020, https://labs.loc.gov/static/labs/work/reports/Cordell-LOC-ML-report.pdf?loclr=blogsig [last accessed: Mar. 29, 2021].

Cumming, Kate/Picot, Anne, Reinventing Appraisal, in: *Archives and Manuscripts* 42 (2/2014), 133-145, doi:10.1080/01576895.2014.926824.

D'Ignazio, Catherine/Klein, Lauren F., *Data Feminism*, Cambridge, MA, 2020, https://data-feminism.mitpress.mit.edu/ [last accessed: Mar. 29, 2021].

Duchein, Michel, Theoretical Principles and Practical Problems of Respect Des Fonds in Archival Science, in: *Archivaria* (16/1983), 64-82.

Erway, Ricky, Defining "Born Digital," OCLC Research, November 2010, URL: https://www.oclc.org/content/dam/research/activities/hiddencollections/borndigital.pdf [last accessed: Mar. 29, 2021].

Fagan, Benjamin, Chronicling White America, in: *American Periodicals: A Journal of History & Criticism* 26 (1/2016), 10-13, https://muse.jhu.edu/article/613375 [last accessed: Mar. 29, 2021].

Gilliland, Anne, Archival Appraisal: Practicing on Shifting Sands, in: Caroline Brown (ed.), *Archives and Recordkeeping: Theory into Practice*, London, 2014.

Harvey, Ross/Thompson, Dave, Automating the Appraisal of Digital Materials, in: *Library Hi Tech* 28 (2/2010), 313-322, doi:10.1108/07378831011047703.

Hutchinson, Tim, Natural Language Processing and Machine Learning as Practical Toolsets for Archival Processing, in: *Records Management Journal* 30 (2/2020), 155-174, doi:10.1108/RMJ-09-2019-0055.

Jordan, Michael I., Artificial Intelligence – The Revolution Hasn't Happened Yet, in: *Harvard Data Science Review* 1 (1/2019), 1-9. doi:10.1162/99608f92.f06c6e61.

McGillivray, Barbara, et al., *The Challenges and Prospects of the Intersection of Humanities and Data Science: A White Paper from The Alan Turing Institute*, London 2020, doi:10.6084/M9.FIGSHARE.12732164.

Noble, Safiya Umoja, *Algorithms of Oppression: How Search Engines Reinforce Racism*, New York City, 2018.

Padilla, Thomas, *Responsible Operations: Data Science, Machine Learning, and AI in Libraries*, Dublin, OH, 2019, doi:10.25333/xk7z-9g97.

Prom, Christopher J., *Preserving Email – DPC Technology Watch Report*, Digital Preservation Coalition 2011 (rev. ed. 2019).

Rolan, Gregory, et al., More Human than Human? Artificial Intelligence in the Archive, in: *Archives and Manuscripts* 47 (2/2019), 179-203, doi:10.1080/01576895.2018.1502088.

Schellenberg, Theodore R., *Modern Archives: Principles and Techniques*, Chicago 1956.

Shabou, Basma Makhlouf, et al., Algorithmic Methods to Explore the Automation of the Appraisal of Structured and Unstructured Digital Data, in: *Records Management Journal* 30 (2/2020), 175-200, doi:10.1108/RMJ-09-2019-0049.

Sloyan, Victoria, Born-Digital Archives at the Wellcome Library: Appraisal and Sensitivity Review of Two Hard Drives, in: *Archives and Records* 37 (1/2016), 20-36, doi:10.1080/23257962.2016.1144504.

UKRI, The UK's Research and Innovation Infrastructure: Opportunities to Grow our Capability, 2020, https://www.ukri.org/wp-content/uploads/2020/10/UKRI-201020-UKinfrastructure-opportunities-to-grow-our-capacity-FINAL.pdf [last accessed: Mar. 31, 2021].

Vellino, André, et al., Assisting the Appraisal of E-Mail Records with Automatic Classification, in: *Records Management Journal* 26 (3/2016), 293-313, doi:10.1108/RMJ-02-2016-0006.

Vinh-Doyle, William, Appraising Email (Using Digital Forensics): Techniques and Challenges, in: *Archives and Manuscripts* 45 (1/2017), 18-30, doi:10.1080/01576895.2016.1270838.

Winters, Jane/Prescott, Andrew, Negotiating the Born-digital: a Problem of Search, in: *Archives and Manuscripts* 47 (3/2019), 391-403, doi:10.1080/01576895.2019.1640753.

Chapter 1: Artificial Intelligence and Discovering the Digitized Photoarchive

X.Y. Han, Cornell University | Vardan Papyan, University of Toronto | Ellen Prokop, National Gallery of Art, Washington, DC | David L. Donoho, Stanford University | C. Richard Johnson, Jr., Cornell University.

Abstract[1]

This chapter introduces the technical aspects of a useful model for effective interaction between the fields of computer science and cultural heritage preservation. Machine learning researchers at Cornell University, Stanford University, and the University of Toronto are collaborating with staff members of the Frick Art Reference Library (FARL), New York, to explore how computer vision might enhance the accessibility and discoverability of the Library's digital resources. Focusing on FARL's Photoarchive—a research collection of 1.2 million reproductions of works of art in the Western tradition—we are seeking to leverage recent tools in artificial intelligence (AI) to automatically annotate images in the collection with the headings used in the Photoarchive's local, iconography-based classification system. This is being achieved by engineering the syntax of the local classification system into the training and predictive process of deep convolutional neural networks, the cornerstone of modern AI advancements. Thus, the machine leaning researchers are adapting state-of-the-art AI techniques by drawing on and incorporating the expertise of art historians. We demonstrate promising performance metrics and offer informative scientific insights that have the potential to create a valuable tool for metadata creation and image retrieval, an end-product that will address the real-world challenge faced by FARL staff who must manage a growing backlog of images that are digitized but not yet classified.

Section 1. Introduction and Background

Computer vision and computational art history are naturally synergistic fields. With the internet's growing presence in modern life, art museums and cultural heritage institutions are digitizing their collections to connect to a wider and more

1 X.Y. Han, Ellen Prokop, and Vardan Papyan contributed equally to this chapter and are listed alphabetically.

inclusive audience through their websites and social media platforms and thus are releasing terabytes of high-quality, annotated digital images online. Meanwhile, state-of-the-art deep neural networks have achieved near human-level performance in the identification of the subject matter and formal qualities of digitized images, a performance predicated on the availability of large and fully labeled training datasets such as those produced by museum and art library staff. Therefore, while the union of computer vision and art history may at first appear surprising, it offers great benefits to both disciplines. In this chapter, we document a collaboration between computer vision researchers and art librarians to harness recent advancements in artificial intelligence (AI) and machine learning (ML)[2] to develop an algorithm that has the potential to become a valuable tool for metadata creation and image retrieval. While we do develop new, specialized technical tools for this task (see Section 2.4), we emphasize the notability of this project is instead defined by the successful integration of the expertise of both the art history and computer science communities—through a collaborative back-and-forth that has now been ongoing for more than three years (since 2017)—into a pipeline that possesses mutually-acknowledged practicality.

Section 1.1 The Frick Art Reference Library's Photoarchive

The Frick Art Reference Library (FARL) was established in 1920 by the philanthropist Helen Clay Frick (1888–1984) (Fig 1.1) as a memorial to her father, the industrialist, financier, and art collector Henry Clay Frick (1849–1919).[3]

2 The distinction between AI and ML is vague—even to computer scientists. AI often refers to more modern, end-to-end algorithms in which the input is the raw data and the output is a recommended decision. In contrast, ML often refers to a broader class of computational algorithms that may first require human pre-processing of the raw data (often based on mathematical principles) before applying a main algorithm that may either output the final decision or just a subtask of the decision. Yet this distinction merely describes the contexts in which such terms are applied; in practice, the terms are often used interchangeably.

3 Martha Frick Symington Sanger, *Henry Clay Frick: An Intimate Portrait*, New York 1998, 499.

Fig 1.1: Portrait of Helen Clay Frick in her office at the Frick Art Reference Library, 1939, photographer unknown. Frick Family Photographs. Courtesy of The Frick Collection/Frick Art Reference Library Archives.

The founding collection of this research institution is the Photoarchive, a study collection of more than 1.2 million reproductions of works of art in the Western tradition from the fourth through the twentieth centuries, a resource modeled on the famous Library of Reproductions assembled by Sir Robert Witt (1872–1952) and his wife and housed at their home at 32 Portman Square in London.[4] The Witts' archive served as a "central and comprehensive storehouse [of reproductions] for easy and rapid reference and research" and was open by appointment to "scholars, critics, writers, collectors, [and] dealers."[5] When Helen Clay Frick visited the Witts' archive in the summer of 1920, she immediately recognized its value for art historical scholarship and determined to assemble a similar research collection in North America.[6] With the foundation of her archive, Helen Clay Frick sought to advance the study of art history in the United States, the development of which had been

4 Katharine McCook Knox, *The Story of the Frick Art Reference Library: The Early Years*, New York 1979, 6–7. For information on Sir Robert Witt, see: Sir Robert Witt, Dictionary of Art Historians, URL: https://arthistorians.info/wittr [last accessed: April 2, 2021].

5 Knox, *The Story of the Frick Art Reference Library*, 7.

6 Knox, *The Story of the Frick Art Reference Library*, 13.

impeded by several factors, the most significant of which was that high-quality photographs of works of art were often prohibitively expensive for many students and scholars.[7] During the nineteenth and early twentieth centuries, North American art historians often had to travel to complete their research, which was an option that only a few could afford. Photoarchives such as those founded by Helen Clay Frick, which made possible the consultation of hundreds of high-quality reproductions with related documentation at one time, helped not only to promote the discipline in the United States but also to motivate critical developments in the field, shifting the focus of study from artist biographies to comparative analysis.[8]

To expand the accessibility of this research collection, FARL staff began digitizing the Photoarchive in the late 1990s, a project that will continue through at least 2025. By the fall of 2020, digital images and documentation for more than 317,500 works of art have been made freely available on the institution's online digital archive, The Frick Digital Collections.[9] In December 2022, approximately 230,000 additional images are scheduled to be uploaded. Therefore, in the next two years, The Frick Digital Collections will potentially host images and metadata representing more than 547,500 works of art.

Unfortunately, cataloguing and metadata creation are not keeping pace with digitization and the backlog of images that have been digitized but are not yet catalogued is growing rapidly. FARL's photoarchivists decided that the accessibility and discoverability of the Photoarchive's holdings is the institution's priority and thus determined that all digital images will be released online with only minimal documentation. This includes the artist, title of the work of art, date of execution, and the institution's local, iconography-based classification system, which not only provides a fixed filing location for each reproduction in the physical archive but also increases the online discoverability of specific subjects and themes. Staff will enhance the online catalogue once the entire research collection has been digitized. Yet even applying minimal information for each digitized image is a time-consuming process.

FARL's photoarchivists investigated crowdsourcing as one means to augment the rate of metadata creation. Preliminary experiments with crowdsourcing, however, were unsatisfactory. For many volunteers, applying the Library's local classification system (described in Section 1.2) proved too restrictive and they either

---

7 Knox, *The Story of the Frick Art Reference Library*, xi.

8 For additional information about FARL's Photoarchive and its impact on the development of art history in the United States, see: Ellen Prokop, Digital Art History for the Masses? The Role of the Public Digital Art History Lab, in: *Život umjetnosti: Journal for Modern and Contemporary Art and Architecture* 105 (2/2019), 196–213.

9 The Frick Digital Collections are available at: https://digitalcollections.frick.org/ [last accessed: April 2, 2021].

abandoned the project or resorted to labeling the images with their own terms. These tags as developed and applied by volunteers certainly increase the discoverability of images in the digital realm but they do not provide standardized search results, which is of paramount concern for scholars who require all known examples of a certain subject or compositional element when conducting their research. Another solution was necessary. Therefore, Library staff launched a pilot project in collaboration with a team of computer vision researchers from Cornell University, Stanford University, and the University of Toronto[10] to bring both efficiency and standardization to the cataloguing process using new advancements in AI.[11] The computer vision researchers were intrigued by both the Photoarchive's holdings, which offered a unique dataset, and the classification system used to organize this research collection (described below), which afforded an exceptional intellectual challenge.

Section 1.2 Organization of the Photoarchive

As noted above, FARL's Photoarchive is modeled on Sir Robert Witt's photo study collection, which was deeded in 1944 to the University of London and later incorporated into the research libraries of The Courtauld Institute of Art.[12] The Witt Library as it is currently known features hundreds of thousands of photographs and published reproductions of works of art mounted on archival-quality paper. The artist, title, and date of the work of art are noted on the front of the sheets; on occasion, additional information regarding the attribution of the work or its provenance—that is, its record of ownership—is included on the reverse. These sheets or "photo study mounts" are stored in large boxes organized alphabetically by artist

10 When the collaboration began in 2017, X.Y. Han, Vardan Papyan, and David L. Donoho were at Stanford University while C. Richard Johnson, Jr. was at Cornell University and serving as FARL's Senior Research Advisor. Since that year, Han has moved to Cornell and Papyan to the University of Toronto—leading to the wide geographic spread of this project.

11 During the term of the collaboration, Han was supported in part by National Science Foundation Grants DMS-1407813, DMS-1418362, and DMS-1811614, and donations to Cornell University from private donors; Papyan was supported by National Science Foundation Grants DMS-1407813, DMS-1418362, and DMS-1811614, the Koret Foundation, and other private donors; Donoho was supported by National Science Foundation Grants DMS-1407813, DMS-1418362, and DMS-1811614, and by donations to Stanford University from Anne T. and Robert M. Bass; Johnson was supported in part by the National Science Foundation Grant CCF-1822007 and donations to Cornell University from Geoffrey and Susan Hedrick and other private donors. FARL donated staff time to the initiative; any travel-related expenses were supported in part by funds from private donors.

12 Witt Library, The Courtauld, URL: https://courtauld.ac.uk/study/resources/image-libraries/witt-library [last accessed: April 2, 2021].

and grouped in national schools; within the boxes, the artist's oeuvre is subdivided by subject to aid discoverability.[13]

When assembling FARL's Photoarchive, Library staff purchased black-and-white photographs from agents and dealers or cut reproductions from sales catalogues and mounted these images on nine-by-twelve-inch sheets of archival-quality grey cardboard. On the front of the sheets, they recorded the artist or national school, the title or subject, the current location, medium, and dimensions of the work of art. On the reverse, they sought to provide more complete documentation than the Witts and included information regarding the object's date of execution, attribution history, exhibition history, conservation history, provenance, and physical characteristics. Additional data such as the source of the mounted photograph or reproduction, documentation of other sources of reproductions, and a bibliography might also be noted. Like the photo study mounts in the Witts' archive, the FARL mounts were grouped by national school, filed alphabetically by artist, and then subdivided by subject. (If the artist was unknown, the work was filed under the national school and subject only.) Helen Clay Frick and her staff, however, applied a numerical classification system to the subject categories, one that incorporated the artist's national school. This improvement to the Witt's system resulted in a stable filing position for each mount and allowed for the discoverability of specific subjects and themes.[14]

Thus, the half-length portrait of Captain Thomas Sprigg (ca. 1765–1810) by the American artist Joshua Johnson (ca. 1763–1824) is not filed under the artist's name in a folder labeled "portraits" as it would be in the Witt Library but catalogued under the artist's name with the classification heading "121-6" (Fig 1.2 and 1.3).

13 Knox, *The Story of the Frick Art Reference Library*, 17–18.

14 Knox, *The Story of the Frick Art Reference Library*, 17.

Fig 1.2: Photo study mount of Joshua Johnson's Captain Thomas Sprigg (ca. 1805–1810), obverse. Courtesy of the Frick Art Reference Library Photoarchive.

The breakdown of this heading is as follows: the first "1" in the series designates the American School; the following "21" denotes a portrait of a man; and the "6" included after the dash indicates a half-length subject not wearing a hat facing right (as opposed to one facing left, which would be indicated with a "7," or one mounted on a horse, which would be indicated with a "3"). A group portrait of a man and two boys such as Johnson's portrait of John Jacob Anderson and his sons John and Edward presently in the collection of the Brooklyn Museum[15] is classified as "124-7" ("American School: Portrait Groups: Men and Children"); however, Johnson's charming portrait of the three sons of the Westwood family preserved in the collection of the National Gallery of Art in Washington, DC[16] is classified as "124-4" ("American School: Portrait Groups: Children"). The engineering process to adapt modern AI techniques to automatically classify images labeled with the Library's in-house classification system is described in the next section.

15 Joshua Johnson, John Jacob Anderson and Sons, John and Edward, The Brooklyn Museum, URL: https://www.brooklynmuseum.org/opencollection/objects/2168 [last accessed: April 2, 2021].

16 Joshua Johnson, The Westwood Children, National Gallery of Art, URL: https://www.nga.gov/collection/art-object-page.45955.html [last accessed: April 2, 2021].

Fig 1.3: Photo study mount of Joshua Johnson's Captain Thomas Sprigg (ca. 1805–1810), reverse, with classification number at upper right. Courtesy of the Frick Art Reference Library Photoarchive.

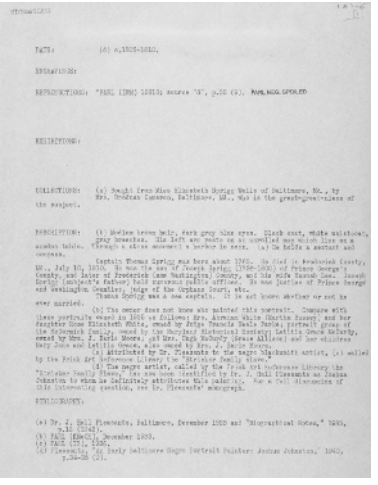


Fig 1.4: Example of hierarchical structure of the FARL classification system for the "Mythological" heading. The left shows the original heading in the FARL system. The right shows the same heading represented as a tree diagram.

15 Mythological 15-1 Mythological: Single figures 15-2 Mythological: Single figures with amoretti 15-3 Mythological: Groups: Four figures and fewer 15-4 Mythological: Groups: More than four figures 15-5 Mythological: Amoretti	Mythological: +---> Single figures: | +---> With amoretti | | \---> ~ | \---> ~ +---> Groups: | +---> Four figures and fewer | | \---> ~ | +---> More than four figures | | \---> ~ | \---> ~ +---> Amoretti | \---> ~ \---> ~

Section 1.3 AI for Image Classification

This pilot project builds upon the deep convolutional neural networks algorithm that has made rapid leaps in performance in the past decade: the study of this powerful algorithm is often called "deep learning."[17] By passing images through numerous computational layers, deep networks extract relevant features from images in order to predict the category to which the image belongs. Their behavior is controlled by a set of parameters—often numbering in the millions. In order to achieve the best performance, empirical algorithms must adjust these parameters by "training" them on a large collection (usually with at least tens of thousands of images) of pre-labeled images. More details are found in Section 2.

When large training sets exist, deep nets have become valuable scientific tools for performing complicated predictive tasks. For example, they have been employed by doctors to identify cancer cells[18] and by physicists to discover new subatomic particles.[19] The FARL's Photoarchive is perfectly positioned to benefit from this technology. When the collaboration was launched, FARL staff had digitized and catalogued 57,803 reproductions of American paintings, the majority of which were portraits and represented 642 unique classification headings and subheadings.[20] These tens of thousands of labeled-and-digitized images represent an ideal training dataset, and the trained network can be deployed to expedite the annotation of the hundreds of thousands of digitized images in the Photoarchive that have yet to be labeled.

Moreover, the Photoarchive presents an interesting engineering challenge since its dataset differs from conventional deep learning classification tasks in two ways. First, each classification heading consists of a series of constituent descriptors. Second, the Photoarchive classification headings follow a hierarchical structure (see Fig 1.4). For example, when photoarchivists applied the classification heading "121-6" or "American School: Portraits: Men: With hands (without hats): Head to right" to the half-length portrait of Captain Sprigg introduced above, they considered the descriptor "Head to right" only after the components "Portraits" and "Men" had already applied. (See Section 1.2.)

To tackle this problem, FARL staff and the artificial intelligence researchers at Cornell University, Stanford University, and the University of Toronto collabo-

17 Ian Goodfellow/Yoshua Bengio/Aaron Courville, *Deep Learning*, Cambridge 2016.

18 Dayong Wang et al., Deep Learning for Identifying Metastatic Breast Cancer, *arXiv preprint*, URL: arxiv:1606.05718 [last accessed: April 2, 2021].

19 Pierre Baldi/Peter Sadowski/Daniel Whiteson, Searching for Exotic Particles in High-Energy Physics with Deep Learning, in: *Nature Communications* 5 (1/2014), 1–9.

20 These reproductions were among the first archival resources to be digitized by Photoarchive staff because the Library owns the copyright to these photographs.

rated to develop a deep learning framework specialized to the Photoarchive's local classification system. The photoarchivists developed a decision tree capturing the classification system's syntax, and the deep learning researchers incorporated this syntax into both the objective with which the network parameters are optimized as well as the algorithm with which the network's decision is made.

In the latest version, we modified the popular ResNet152 network[21]—with 152 layers and 6 million parameters—and trained it on 46,242 classified reproductions of American paintings from the collection provided by the FARL's Photoarchive described above (see Section 3.1). More empirical and engineering details are provided in Section 3.

After training, the network was fed images from the unlabeled portion of the Library's Photoarchive and the network predicted a classification heading for each one. (For more details about the dataset, see Section 3.) These images were then annotated with the predictions and shown to the Library's photoarchivists through an application developed using the crowd-sourcing platform, Zooniverse.[22]

Section 1.4 Human Validation with Zooniverse

Zooniverse is a popular "citizen science" website where research teams post raw data in need of human annotation and processing. Through the Zooniverse desktop and mobile app, volunteers from the general public or specially-selected groups can then assist with metadata creation. When developing this app, we sought to produce an intuitive interface that would allow archivists to review hundreds of images quickly and easily; thus, the program mimics the notorious dating app Tinder. Library staff downloaded the app on their computer desktops or smartphones (Fig 1.5 and 1.6) and reviewed the algorithm's predictions. If the classification heading applied to an image was correct, it was considered a "match" and the staff member swiped right. If it was incorrect, staff swiped left and the image was sent to a folder for review.

21 Kaiming He et al., Deep Residual Learning for Image Recognition, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2016), 770–78.

22 Zooniverse, URL: https://www.zooniverse.org [last accessed: April 2, 2021].

Fig 1.5: Screenshot of the application developed using the crowd-sourcing platform Zooniverse to vet the algorithm's predictions as it appears on a computer desktop. Courtesy of the authors.

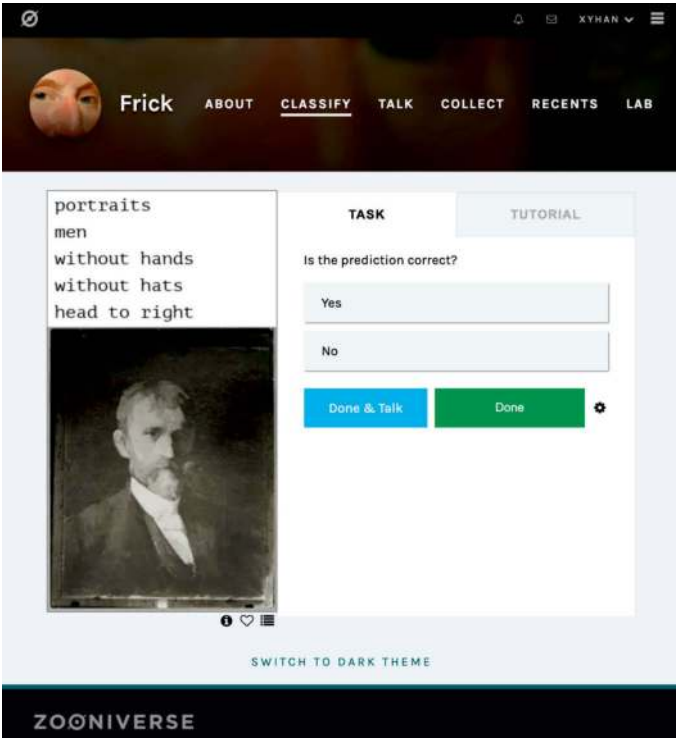


Fig 1.6: Screenshot of the application developed using the crowd-sourcing platform Zooniverse to vet the algorithm's predictions as it appears on a smartphone. Courtesy of the authors.

In testing the latest version of the network, photoarchivists vetted 8,661 images in the year August 2019–August 2020 and agreed with the network on the classification heading for 5,829 (67%) of the images. Yet even the incorrect predictions were, in general, "almost correct" with only one descriptor incorrect or missing. For example, the algorithm applied the classified heading "124-2: American School: Portrait Groups: Men" when it should have applied the heading "124-7: American School: Portrait Groups: Men and Children." In the future, we predict that the network can "learn" to identify these cases when given additional training examples of the headings on which it erred. In Fig 1.9, we show a small selection of outputs from

the network that demonstrate the types of one-term omission and extra inclusion mistakes that the network tends to make.

Fig 1.9: An example of typical outputs from the Zooniverse app on images where the network applies the incorrect label. We see that errors tend to be one-term omissions or extra inclusions. Moreover, notice that, to ensure the quality of the annotations, each image is vetted by multiple FARL personnel (each with a unique "user_name"). Their decision is indicated by "annotation" and their correction is given in "comment_body." The "expert_status" column indicates whether the user is a FARL staff member ("expert") or intern ("non-expert"). Attesting to the reliability of the process, most annotators tend to be in agreement on the required correction. Courtesy of the authors.

image_name	network_prediction	annotation	comment_body	user_name	expert_status
31071001176B6_001.jpg	portraits: women: with hands: head to right:	No	Portraits: Women: Without hands: Head to right.	JohnMcQ	expert
31071001176B6_001.jpg	portraits: women: with hands: head to right:	No	Portraits: Women: Without hands: Head to right.	sarahbigler	expert
31071001206S1_001.jpg	portraits: men: without hands: without hats: head to right	No	Portraits: Men: With hands (without hats): Head to right.	JohnMcQ	expert
31071001206S1_001.jpg	portraits: men: without hands: without hats: head to right	No	Portraits: Men: With hands (without hats): Head to right.	kerrip	expert
31071001206S1_001.jpg	portraits: men: without hands: without hats: head to right	No	Portraits: Men: With hands (without hats): Head to right.	sarahbigler	expert
31071001206S1_001.jpg	portraits: men: without hands: without hats: head to right	No	Portraits: Men: With hands (without hats): Head to right.	Ldla657	non-expert
31071001272B2_001.jpg	miniatures	No	Miniatures: Portraits: Men.	JohnMcQ	expert
31071001272B2_001.jpg	miniatures	No	Miniatures: Portraits: Men.	sarahbigler	expert
31071001285B8_001.jpg	portraits: men: with hands: without hats: head to right	No	Portraits: Men: With hands: With hats: Head to right.	levadas	expert
31071001285B8_001.jpg	portraits: men: with hands: without hats: head to right	No	Portraits: Men: With hands: With hats: Head to right.	sarahbigler	expert
31071001314S2_001.jpg	portraits: women: without hands: head to left:	No	Portraits: Women: Without hands: With hats.	levadas	expert
31071001314S2_001.jpg	portraits: women: without hands: head to left:	No	Portraits: Women: Without hands: With hats.	kerrip	expert
31071001314S2_001.jpg	portraits: women: without hands: head to left:	No	Portraits: Women: Without hands: With hats.	sarahbigler	expert
31071001314S2_001.jpg	portraits: women: without hands: head to left:	No	Portraits: Women: Without hands: With hats.	__genie	non-expert

As the digitization of the Photoarchive proceeds, additional image sets will become available for further testing, the results of which will also be vetted by Library staff using the app described above. The team anticipates that the network will be ready to be deployed within two years; thus, automatic image classifiers will become available for cataloguing purposes by 2022.

This project contributes to the growing field of works applying computer vision techniques to art classification and cultural heritage preservation. We discuss some of these related works in Section 4. Yet, this work is set apart by the back-and-forth collaborative process between FARL staff and the researchers at Cornell University, Stanford University, and the University of Toronto, which is creating a high-performing image classifier that follows the syntax of the FARL in-house classification system. In the following sections, we describe in detail the design of the image classifier as well as scientific measurements of its performance.

Section 2. Image Classification with Deep Nets

In image classification, we are given a training dataset of n example pairs, $\mathcal{D} \equiv \left\{\left(x_i, y_i\right)\right\}_{i=1}^{n}$, where x_i denotes the i-th image and y_i, called the *target*, is typically a term from a controlled vocabulary. For example, for the FARL classification system, we had total of 57,803 images and the descriptions were chosen from the

FARL vocabulary of 378 terms, such as "portrait," "landscape," etc. Such a dataset documents each of the descriptions that human experts would provide were they to describe the corresponding images. We seek to learn from this data a computational procedure that can successfully reproduce the descriptions from human experts on this dataset and hopefully on other similar datasets. Because the image descriptions come from a controlled vocabulary, experts, in describing an image, in effect classify it into one of several categories. So, a procedure which reproduces expert judgements can be called an image classifier.

Today's standard for image classification—a modern deep network—involves a *feature extraction* pipeline consisting of many *layers* jointly trained to distill relevant features of the input image. The output of this pipeline is a vector of *scores*, which determine the network's final description. Conceptually, each such pipeline computes, for some input image x, a function $f(x)$ giving those scores. Neural networks map the scores into descriptions by applying some fixed decision function $d(\cdot)$; each score vector $f(x)$ will induce the specific description

$$\hat{y}(x) \equiv d(f(x)),$$

called the *network prediction*.

In practice, the many layers of a deep network contain adjustable parameters, and we must *train* or *learn* these parameters to get a useful performance. Let θ denote the collection of adjustable parameters; f_θ will denote the scoring function f, when those parameters take the specific value θ.

Machine learning algorithms attempts to minimize the cross-training-set average of a *loss function*, $L(\cdot,\cdot)$, on $\mathcal{D}$:

$$\text{Network Training:}\ \min_\theta \frac{1}{n}\sum_{i=1}^{n} L\left(f_\theta(x_i), y_i\right).$$

Generally, L is chosen such that the *training loss*, $\frac{1}{n}\sum_{i=1}^{n} L\left(f_\theta(x_i), y_i\right)$, is large when the network makes many classification errors and small when the network performs well. Moreover, L should have desirable mathematical properties such as smoothness, boundedness, and continuity.

The minimization described in the display above can then be performed using empirical algorithms such as stochastic gradient descent (SGD).[23] These algorithms need an initial set of parameters, θ_0, on which it makes iterative adjustments. One could choose θ_0 randomly: this is called training *from scratch*. However, a common practice is initializing θ using parameters of a network trained on a

23 Léon Bottou, On-Line Algorithms and Stochastic Approximations, in: D. Saad (ed.), *Online Learning and Neural Networks*, Cambridge 1998; Léon Bottou/Olivier Bousquet, The Tradeoffs of Large Scale Learning, in: J. C. Platt et al. (eds.), *Advances in Neural Information Processing Systems 20 (NIPS 2007)*, 2007, 161–168.

different task and applying the optimization algorithm from the pretrained starting point. This practice, called *fine-tuning*, reduces the computational cost of training a network, and allows one to take advantage of image filters and feature extractors (created for a different task using a much larger dataset) that may not be achievable if one were to train from scratch on $\mathcal{D}$ (because n is too small).[24]

For example, today there are massive image datasets with millions and even billions of images for which network models have been trained; even if the images and descriptions in such existing trainings are quite different from those in our specific dataset. Under fine-tuning, we borrow an existing pipeline architecture and fully trained model to get our initialization $\mathcal{D}$, $u\theta_0$. One notable advantage of the deep net frameworks developed in Section 2.3 is that they are amenable to fine-tuning despite having undergone significant structural changes from the original setting.

Hence, we see that training requires only the specification of descriptions, prediction rule $\hat{y}$, and loss L to represent the task of interest. We have not yet fully described how the loss and prediction functions are chosen. This depends on details of the allowable descriptions of images: they could be fixed phrases chosen from a list; they could be combinations of such phrases; there could even be a grammar of allowable phrase combinations.

Section 2.1 Single-label Classification

The *single-label* classification problem has C different possible descriptions, and we must assign one of those C descriptions to each image. The collection of all images with the same description is called a *class*. Most modern deep neural networks were originally developed for classifying images in the benchmark ImageNet dataset[25] with $C = 1000$ classes (fish, bird, tree, etc.). In the context of works of art, the classes might represent the artists; the period; or a description of the objects represented in the scene. Most prior work for fine art classification solves single label problems. (See Section 4.)

Formally, for each image x, its description can be identified with its class's serial number, so the target can be represented as

$$y \in \{1, \ldots, C\},$$

and we call y the *true class* to which the image belongs. The feature vector $f_\theta(x)$ contains C numbers $f_\theta \equiv \left(f_\theta^c\right)_{c=1}^{C}$ with larger numbers meaning "more likely" and

24 Jason Yosinski et al., How Transferable Are Features in Deep Neural Networks?, in: Z. Ghahramani et al. (eds.), *Advances in Neural Information Processing Systems* 27, 2014, 3320–3328.

25 Jia Deng at al., Imagenet: A Large-Scale Hierarchical Image Database, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Miami, FL, 2009, 248–255.

smaller numbers "less likely." The decision of the network chooses the class (i.e. the description) having the highest score: In other words, $\hat{y}$ is the class corresponding to the largest score:

$$\hat{y}(x) = \operatorname{argmax}_{c \in \{1,\ldots,C\}} f_\theta^c(x).$$

For the single-label classification task, modern machine learning practice typically employs the *cross-entropy* (CE) loss:

$$L_{CE}(f_\theta(x), y) \equiv -\log\left(\frac{\exp\left(f_\theta^y(x)\right)}{\sum_{c=1}^{C}\exp\left(f_\theta^c(x)\right)}\right).$$

Observe that L is small when the score of the correct class is much larger than the score of the incorrect classes. Additionally, note that the argument of the logarithm is always between 0 and 1. This argument, called the *softmax probability*, will be denoted by the function

$$p(s, y) \equiv \frac{\exp\left(s^y\right)}{\sum_{c=1}^{C}\exp\left(s^c\right)}.$$

Section 2.2 Simple Multilabel Classification

In the multilabel setting, *multiple* labels from a collection of C potential labels may have been used on any particular image, and one tasks the network to identify all labels simultaneously. For example, eligible labels might include "with water" and "with bridges," and we might want both labels identified when they are both present in an image. In contrast, some images might correctly be described only as "with bridges" or only as "with water."

In this setting, the true labels identify *subsets* $y \subseteq \{1, \ldots, C\}$. Such a subset y is typically represented as $y \in \{0, 1\}^C$ i.e. a length-C binary vector such that the c-th element, denoted y, is 1 or 0 according to whether the c-th label applies or not, respectively.

The network feature extractor, f_θ, remains the same, but the network's prediction,

$$\hat{y}(x) \equiv \left(\hat{y}^c(x)\right)_{c=1}^{C},$$

is now a binary vector with components

$$\hat{y}^c(x) = \mathbf{1}\left\{p(f_\theta(x), c) > \gamma\right\};$$

Here, p is again the softmax probability, $\gamma \in [0, 1]$ is a *threshold parameter* chosen by the network engineer beforehand, and $\mathbf{1}\{\cdot\}$ is 1 or 0 depending on whether its argument is true. In Section 3.3, we discuss the choice of γ.

To train such a network, the CE loss that we discussed in the single label setting possesses a natural extension to the multilabel case called the *binary cross-entropy (BCE) loss:*

$$L_{BCE}\left(f_\theta\left(x\right), y\right) = -\left[\sum_{c=1}^{C} y^c \log p\left(f_\theta\left(x\right), c\right) + \sum_{c=1}^{C}\left(1 - y^c\right) \log\left(1 - p\left(f_\theta\left(x\right), c\right)\right)\right]$$

Training in the multilabel setting with the BCE loss is the machine learning community standard.

Section 2.3 Hierarchical Multilabel Classification

Multilabel classification problems sometimes possess an additional layer of structure that we call *hierarchical multilabel classification*: this is the case with the FARL classification system. In this problem, experts will never combine multiple labels in a completely arbitrary way; for example, it makes no sense to label the subject of an image as both "head to left" and "head to right." In fact, the possible descriptions in the FARL dataset can be identified with the nodes of a hierarchical structure.

Abstractly, the Frick classification system can be considered domain-specific language consisting of a set, $\mathcal{T}$, of C different *terms*—subsets of which can be grouped together into *phrases*. Yet, the grouping of terms into phrases must follow a syntax i.e. only certain combinations of terms in certain orders are allowed. Mathematically, we can represent phrases as ordered tuples of terms. Letting $\mathcal{P}$ represent the collection of all legal phrases, their relationship can be represented as

$$\mathcal{P} \subset \left\{\left(t_1, t_2, \ldots\right) : t_i \in \mathcal{T}\ \forall\ i\right\}.$$

In practical classification systems, the collection of all phrases has a hierarchical structure, where the topmost elements (the leftmost in the tuple) are the terms that can function as phrases all on their own, and the lower phrases in the hierarchy are syntactically legal continuations of higher phrases in the hierarchy.

In the context of FARL's classification system, components such as "portraits," "men," and "with hands" are examples of terms; full headings such as "Portraits: Men: With hands" and "Landscapes: With water" are examples of phrases; these phrases would be encoded as the tuples (portrait, men, with hands) and (landscapes, with water), respectively.

In the hierarchical multilabel setting the targets are then $y \in \mathcal{P}$. To incorporate the classification system's syntax into our prediction and loss functions, we first define the concepts of prefixes and syntactical continuations.

Definition (Prefix). If $r = (t_1, \ldots, t_l)$ is a valid phrase consisting of all l terms, then each tuple $r_k = (t_1, \ldots, t_k)$ where $1 \leq k \leq l$ is a *prefix* of r. We let $\mathcal{P}_+$ denote the set of all prefixes of phrases in $\mathcal{P}$.

Observe that the full-phrases are prefixes themselves, i.e., $\mathcal{P} \subseteq \mathcal{P}_+$.

Definition (Syntactical continuation function). For a set of phrases, $\mathcal{P}$, built from terms, $\mathcal{T}$, the *syntactical continuation* function, $\mathcal{S}\colon \mathcal{P}_+ \to 2^{\mathcal{T}}$, lists all terms that may follow the prefix i.e.

$$\mathcal{S}(r) = \left\{ t \in \mathcal{T} \colon\ (r, t) \in \mathcal{P}_+ \right\},$$

where (r, t) denotes the tuple created by appending the term, t, to the tuple r.

Since $\mathcal{T}$ has C elements and the score vector $f_\theta(x)$ has C components, we can assign a one-to-one correspondence between them—denoting $f_\theta^t(x)$ as the component of $f_\theta(x)$ associated with term t. Then, for a target phrase y of length L, define the *hierarchical cross-entropy* (HCE) loss:

$$L_{HCE}(f_\theta(x), y) = -\sum_{\ell=0}^{L} \sum_{t \in \mathcal{S}(y^{1:\ell})} \left[\mathbf{1}\{y^{\ell+1} = t\} \log\left(p_{y^{1:\ell}}(f_\theta(x), t) \right) + \mathbf{1}\{y^{\ell+1} \neq t\} \log\left(1 - p_{y^{1:\ell}}(f_\theta(x), t) \right) \right]$$

where we consider a *branch-specific softmax*,

$$p_{y^{1:\ell}}(f_\theta(x), t) \equiv \frac{\exp\left(f_\theta^t\right)}{\sum_{t' \in \mathcal{S}(y^{1:\ell})} \exp\left(f_\theta^{t'}(x)\right)},$$

and we follow the conventions that

$$\mathbf{1}\left\{ y^{L+1} = t \right\} = \left(1 - \mathbf{1}\left\{ y^{L+1} \neq t \right\} \right) = 0,$$

and that $\mathcal{S}\left(y^{1:0}\right)$ is the empty set.

Given the scores $f_\theta(x)$ and a preset threshold, γ, the *syntax-aware classifier*, $d_\gamma(f_\theta(x))$, is the procedure described in Algorithm 1.

Algorithm 1:

INPUT: $f_\theta(x)$, γ

$\hat{r} = () ; \ell = 0$

WHILE $\left(\mathcal{S}\left(\hat{r}\right) \neq \varnothing \textbf{ AND } \min_{t \in \mathcal{S}(\hat{r})} p_{\hat{r}}\left(f_\theta\left(x\right), t\right) > \gamma\right)$:

$$\hat{r} \leftarrow \left(\hat{r}, \arg\max_{t \in \mathcal{S}(\hat{r})} p_{\hat{r}}\left(f_\theta\left(x\right), t\right)\right)$$

OUTPUT: $\hat{r}$

END

For an arbitrary image, x, the network prediction is, the network prediction is $\hat{y}_\theta\left(x\right) = d_\gamma\left(f_\theta\left(x\right)\right)$.

Section 3. Empirical Results

To demonstrate the advantage of the hierarchical classification framework using the HCE loss and syntax-aware classifier, we show experiments comparing three approaches for classifying the FARL dataset corresponding to different approaches described in Section 2:

1. **(SL):** Single-class classification, where we consider each of the 642 unique FARL headings to be a unique class.
2. **(SML):** Simple-multilabel classification, where any image can be given any subset of the 378 single-term labels that comprise headings in the FARL system.
3. **(HML):** Hierarchical multilabel classification trained with decisions made through the syntax-aware classifier. In this context, the FARL classification contains 642 different phrases built from 378 terms.

In these experiments, we compare the performance of each method on a *test* dataset that contains a *different* set of images than the training data ($\mathcal{D}$). Thus, the training set serves to train the classifiers, while the testing set evaluates on the performance of the trained classifiers on new, previously unseen data.

Section 3.1 Experimental Details

Specifically, we use a dataset of 57,803 images from the American Portraits Collection within the FARL Photoarchive, which have already been labeled by photoarchivists according to the FARL classification system. This dataset is randomly split into a training subset, consisting of 80% of the images (46,242 images), and a testing subset, containing the remainder (11,561 images). Adding to the challenge of the task, different terms and phrases possess varying numbers of training examples. For instance, the most populous term, "portraits," possess 17,075 training examples while the term "with sheep and goats" possesses only one example in the training set. Similarly, the most populous phrase, "genre," possess 1,715 examples

while the "animals: cattle: with figures" possesses only one example. See Section 3.5 on how this class imbalance affects performance.

The images range in size from 55KB (1213x1536 pixels) to 4.4MB (4868x2856 pixels). Pixels are standardized by subtracting the (red, green, or blue) channel mean and dividing by the (red, green, or blue) channel standard deviation. We follow the same data augmentation steps typically performed on the ImageNet dataset: Images are rescaled such that the smaller dimension is 256 pixels, and then a random crop (in training) or a center crop (in testing) of 224x224 pixels is extracted. We do *not* apply random horizontal flips—as is common in deep networks trained on computer vision applications—since the FARL headings distinguish between portraits facing left or right.

In all settings, we train the feature extractor (f_θ) by fine-tuning all the parameters of a ResNet152 architecture pretrained on ImageNet. The pretrained model was downloaded directly from the PyTorch ModelZoo.[26] Following common practice, we minimize the task-specific loss using stochastic gradient descent (SGD), with a momentum of 0.9, and a weight decay of 10^{-4}. Our networks are trained on 4 GPUs, with a total batch size of 32 images, for 100 epochs. The initial learning rate is annealed by a factor of 10 at epoch 34 and 67. We train models with initial learning rates 0.1 and 0.01—picking the one resulting in the best true positive rate in the last epoch. In all three tasks, initial learning rate 0.01 produced the best model.

Section 3.2 Performance Metrics

In order to compare the performance of the three methods on the test set, we gather the following statistics for each of the terms in the FARL classification system (i.e. elements of $\mathcal{T}$ defined in Section 2):

- **Term Count (n_i'):** Number of images such that the i-th term applies.
- **True Positives (TP_i):** Number of images such that the i-th term applies, and the network correctly labeled them with that term.
- **False Negatives (FN_i):** Number of images such that the i-th term applies, but the network didn't label them with that term.
- **True Negatives (TN_i):** Number of images such that the i-th term doesn't apply, and the network correctly didn't label them with that term.
- **False Positives (FP_i):** Number of images such that the i-th term doesn't apply, but the network incorrectly labeled them with that term.

26 PyTorch ModelZoo, URL: https://pytorch.org/docs/stable/model_zoo.html [last accessed: April 2, 2021].

Observe that the following relationships apply for each i:

$$\begin{aligned} TP_i + FN_i + TN_i + FN_i &= n', \\ TP_i + FN_i &= n_i', \\ TN_i + FP_i &= n' - n_i', \end{aligned}$$

where n' is the total number of images in the testing set.

Following common practice, to measure model performance on a particular term, we will use the *precision*, *recall*, and *F1 scores* defined, respectively, as follows:

$$Prec_i \equiv \frac{TP_i}{TP_i + FP_i}, Rec_i \equiv \frac{TP_i}{TP_i + FN_i}, F_1^i \equiv 2\frac{Prec_i \times Rec_i}{Prec_i + Rec_i}.$$

Roughly, precision measures the network's ability to ignore non-examples of a term, while recall (sometimes called true positive rate) measures the network's ability to identify true examples of a term. The F1 score aggregates precision and recall into an overall measure of performance.[27]

To evaluate performance over the *entire testing set*, we use a weighted aggregate of these scores:

$$Rec \equiv \sum_{i=1}^{|\mathcal{T}|} w_i Rec_i, Prec \equiv \sum_{i=1}^{|\mathcal{T}|} w_i Prec_i, F_1 \equiv \sum_{i=1}^{|\mathcal{T}|} w_i F_1^i,$$

where $|\mathcal{T}| = 378$ is the number of terms in the classification system, and the weights,

$$w_i \equiv \frac{n_i'}{\sum_{i=1}^{|\mathcal{T}|} n_i'},$$

capture the relative frequencies of instances of terms in the testing set.

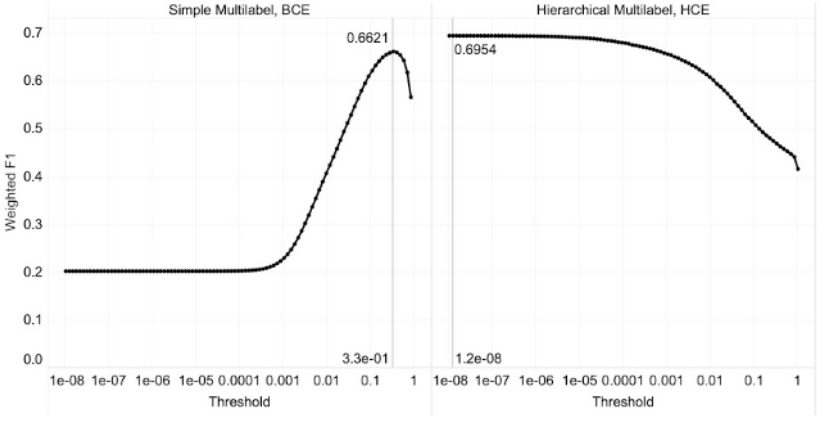
Section 3.3 Determination of Threshold

Recall from Section 2 that both the SML and HML settings require a pre-determined choice for the decision threshold, γ. In practice, we determine this by measuring the final performance on test data for each candidate choice and choosing the best candidate value. In Fig 1.7, we plot an experiment showing the weighted-F1 scores resulting from varying this threshold. The difference in the order of optimal-threshold magnitudes for the SML setting compared to the HML setting reflects the difference between the uniform nature of the BCE loss compared to the adaptive

27 For a more detailed discussion, see: Koo Ping Shung, Accuracy, Precision, Recall or F1? (2018), in: *Towards Data Science. Medium*, URL: https://towardsdatascience.com/accuracy-precision-recall-or-f1-331fb37c5cb9 [last accessed: April 2, 2021].

structure of the HCE loss. Using the best threshold from Fig 1.7 in the remainder of this section, we can then compare the different methods.

Fig 1.7: Weighted-F1 scores versus prediction threshold. Each row corresponds to a different classification approach (described in Section 2) used on images from the Frick Art Reference Library's Photoarchive. In each array cell, we plot (on the logarithmically spaced x-axis) 100 prediction thresholds (γ) and (on the y-axis) their resulting weighted-F1 score on the testing dataset. The largest score and its corresponding threshold are annotated on the top and bottom of each plot, respectively. Training details are given in Section 3.1. Courtesy of the authors.



Section 3.4 Comparison of Performance

In Table 1, for each of the settings described in Section 4.1, we report weighted precision, recall, and F1 scores over the entire testing set. Using the SL setting as a control, we see that changing to the SML setting improves recall (i.e. reduces false negatives) and overall performance but decreases precision (i.e. increases false positives). In comparison, the HML setting improves all three performance metrics compared to both the SL and SML settings.

Table 1. Performance Metrics on Testing Data. Each row corresponds to a different classification approach (described in Section 2) used on images from the Frick Art Reference Library Photoarchive. The first column shows the best threshold chosen according to the weighted-F1 score on the testing data. The last three columns show the weighted precision, recall, and F1 score on the testing data (described in Section 3.2). Training details are given in Section 3.1.

Setting	Best Threshold	Precision	Recall	F1 Score
SL-CE	N/A	0.6538	0.6663	0.6577
SML-BCE	0.33	0.6364	0.7018	0.6621
HML-HCE	1e-08	0.6758	0.7226	0.6954

These results are intuitive. The greater flexibility afforded by predicting single terms out of multiple-term phrases, in the multilabel regime, rather than predicting the complete multi-term phrase, in the single label regime, allows the network to better identify positive instances of labels (better recall) in both the SML and HML case. In the HML case, this performance is enhanced even further by the explicit incorporation of the classification domain-specific language into the training loss—allowing for a more focused and efficient tuning of parameters.

On the other hand, the increased flexibility results in a larger space of possible outputs, which, in turn, increases the possibilities for false positives. For example, in the SML case, all combinations of terms are considered syntactically valid results for the network—even those not in the FARL classification system itself. We see from the decreased precision that the SML case is indeed hurt by failing to use syntax constraints. In contrast, the HML avoids this flaw by reducing the space of possible outputs to the most economical representation: *the only predictions made by the syntax-aware classifier are those in the FARL classification system*. This combined with the advantage of a loss and predictor specialized to the FARL system likely led to HML's increase in precision.

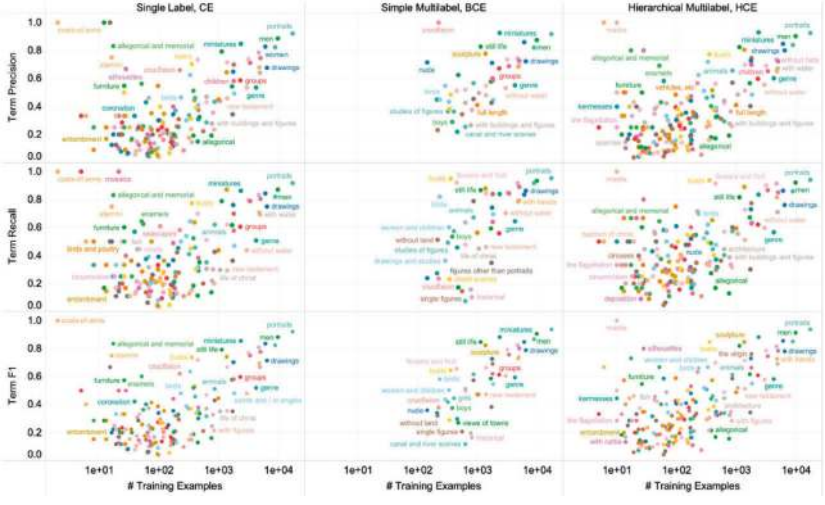
Overall, these results demonstrate the advantage of incorporating the syntax of the classification system into the neural network itself. These performance gains come as a direct consequence of the close collaborative efforts of the AI researchers and art historians where the latter group provided important insights into the structure of the dataset as well as practical aspects of which guided the development of the syntax-aware classifier and HCE loss (Section 2.3).

Section 3.5 Impact of Sample Size

Another prominent factor determining network performance is the number of examples within each class in the training data. For individual terms in the testing

set, Fig 1.8 shows that performance as measured by precision, recall, and F1 score are all approximately proportional to the logarithm of the number of examples of the term in the training dataset. Yet variance is still high even for terms with roughly equal numbers of training examples. Nonetheless, the trends in Fig 1.8 provide an approximate guideline for the necessary amount of prerequisite training data needed to reach a reasonable performance, which is a useful reference for fine art institutions who wish to adopt deep learning methods in the future.

Fig 1.8: Sample size versus precision and recall. Each row corresponds to a different classification approach (described in Section 2) used on images from the Frick Art Reference Library's Photoarchive. The optimal thresholds in the multilabel settings are chosen according to the experiment in Fig 7. Each array row corresponds to a term-wise performance described in Section 3.2. In each array cell, we plot a scatterplot of terms with non-zero true-positive counts: the x-axis (logarithmically spaced) shows the number of examples of a term in the training set, and the y-axis shows the trained network's performance on that term. A subsample of the points is labeled, in the same color, with their corresponding term. Training details are given in Section 3.1. Courtesy of the authors.



Section 4. Related Works

The idea of leveraging computer vision techniques to classify fine art images predates the popularity of deep learning. From the early 2000s to the mid-2010s, researchers tried to automate the annotation of fine art images, usually paintings, by applying regression and classification methods on handcrafted features emerging from the wavelet transform,[28] the scale-invariant feature transform (SIFT),[29] the GIST descriptor,[30] and the histogram of oriented gradients descriptor (HOG).[31] These works predominantly focused on single-label classification tasks, such as the identification of the creator of an artwork. Agarwal et al.—who use five such features for identifying the artist and genre of artworks—provide a comprehensive survey of such works.[32]

The rise of deep learning has led to a paradigm shift. Instead of training a classifier on handcrafted features, deep neural networks are trained *both* to extract the relevant features from an image (in the earlier convolutional layers) and to make the classification decision (in the later linear layers). As a result, around 2015—inspired by the human-level performance of deep convolutional neural networks (CNNs)—researchers shifted their focus to the automatic classification of artworks by fine-tuning CNNs.

The most commonly utilized dataset for this later line of work is WikiArt—a growing online collection of approximately 80,000 digital images of fine art paintings.[33] It includes four tasks: artist identification, style identification, genre identification, and time period identification. For example, Cetinic et al.,[34] Saleh et

28 Jia Li/James Ze Wang, Studying Digital Imagery of Ancient Paintings by Mixtures of Stochastic Models, in: *IEEE Transactions on Image Processing* 13 (3/2004), 340–353.

29 Fahad Shahbaz Khan/Joost Van de Weijer/Maria Vanrell, Who Painted this Painting?, in: *2010 CREATE Conference*, 2010, 329–333.

30 Matthijs Douze et al., Evaluation of Gist Descriptors for Web-Scale Image Search, in: *Proceedings of the ACM International Conference on Image and Video Retrieval*, 2009, 1–8.

31 Xiaoyu Wang/Tony X. Han/Shuicheng Yan, An HOG-LBP Human Detector with Partial Occlusion Handling, in: *2009 IEEE 12th International Conference on Computer Vision*, Kyoto 2009, 32–39.

32 Siddharth Agarwal et al., Genre and Style Based Painting Classification, in: *2015 IEEE Winter Conference on Applications of Computer Vision (WACV)*, Waikoloa, HI, 2015, 588–594.

33 WikiArt, URL: www.wikiart.org [last accessed: April 2, 2021].

34 Eva Cetinic /Tomislav Lipic/Sonja Grgic, Fine-Tuning Convolutional Neural Networks for Fine Art Classification, in: *Expert Systems with Applications* 114 (2018), 107–118.

al.,[35] Tan et al.,[36] Hentschel et al.,[37] and Lecoutre et al.[38] propose different algorithms for tackling the style identification task based upon fine-tuning CNNs and all show how such algorithms achieve state-of-the-art results. The first three of the list of five papers further demonstrate the ability of CNNs to perform well in the artist and genre identification tasks. These advances address various aspects of network architecture design, initialization, and sample size efficiency, but all within the single-label classification regime. Cetinic et al. provides a comprehensive summary and comparison of these works.

Another notable line of research, originating in the mid-2010s, is that inspired by a team from the Rijksmuseum in Amsterdam, the Netherlands. In 2014, Mensink and Gemert released a dataset of 112,039 images of fine art from the collection of the Rijksmuseum.[39] They identified four tasks on this dataset (collectively called the Rijksmuseum Challenge): artist attribution (single-label classification); art-type identification (simple-multilabel classification); materials identification (simple-multilabel classification); and creation year prediction (regression). They then created benchmark performance metrics obtained by encoding the images as Fisher vectors followed by regression or max-margin classification.

In 2017, Strezoski and Worring extended the Rijksmuseum dataset into the OmniArt dataset, which has 432,217 images, adding images and metadata from the Rijksmuseum's collection as well as additional Open Access images from the holdings of the Metropolitan Museum of Art (the Met).[40] They established new benchmark results on the four Rijksmuseum Challenge tasks using fine-tuned CNNs. In 2018, Strezoski and Worring[41] expanded OmniArt to contain more than two

35 Babak Saleh/Ahmed Elgammal, Large-Scale Classification of Fine-Art Paintings: Learning The Right Metric on The Right Feature, in: *International Journal for Digital Art History* (2/2016), 71–93.

36 Wei Ren Tan et al., *Ceci n'est pas une pipe*: A Deep Convolutional Network for Fine-Art Paintings Classification, in: *2016 IEEE International Conference on Image Processing (ICIP)*, Phoenix, AZ, 2016, 3703–3707.

37 Christian Hentschel/Timur Pratama Wiradarma/Harald Sack, Fine Tuning CNNS with Scarce Training Data—Adapting ImageNet to Art Epoch Classification, in: *2016 IEEE International Conference on Image Processing (ICIP)*, Phoenix, AZ, 2016, 3693–3697.

38 Adrian Lecoutre/Benjamin Negrevergne/Florian Yger, Recognizing Art Style Automatically in Painting with Deep Learning, in: Min-Ling Zhang/Yung-Kyun Noh (eds.), *Proceedings of the Ninth Asian Conference on Machine Learning*, *PMLR* 77, 2017, 327–342.

39 Thomas Mensink/Jan Van Gemert, The Rijksmuseum Challenge: Museum-Centered Visual Recognition, in: *ICMR '14: Proceedings of International Conference on Multimedia Retrieval*, Glasgow 2014, 451–454.

40 Gjorgji Strezoski/Marcel Worring, OmniArt: Multi-Task Deep Learning for Artistic Data Analysis (2017), in: *arXiv* [preprint], URL: arXiv:1708.00684 [last accessed: April 2, 2021].

41 Gjorgji Strezoski/Marcel Worring, OmniArt: A Large-Scale Artistic Benchmark, in: *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 14 (4/2018), 1–21.

million images by including Open Access images from additional museums. They added additional metadata that is amenable to the single-label, simple-multilabel, and regression tasks.

In our work with the FARL dataset, we also employed the fine-tuning technique on CNNs. Yet most prior work, such as that by Cetinic et al. and Strezoski and Worring, tackles single-label classification or simple-multilabel classification. In contrast, FARL's in-house classification system is a much better fit to the hierarchical multilabel classification problem. Therefore, our project is distinct from these precedents in two ways. First, unlike in the single-label setting—which requires one to pick only the network prediction with the *highest score*—the multilabel setting requires one to choose all classes whose network-predicted scores exceed some threshold. (See Section 2.) Thus, additional attention must be given to the determination of the threshold. (See Section 3.) Second, fine-tuning a pretrained, single-label network for another single-label or simple-multilabel classification task requires replacing only the last linear layer of the network with another linear layer with dimensions matching the new problem. In contrast, to adapt to FARL's *hierarchical* labeling system, we engineered both a novel *hierarchical cross-entropy loss* as well as a novel *syntax-aware classifier* (described in Section 2), which, as we demonstrated in Section 3, results in a higher performing network.

A related labeling system is Iconclass,[42] which is also a hierarchical classification system for images of fine art but is structurally different from FARL's system. Unlike Iconclass, headings in FARL's system are comprised of *common components*: each multi-term phrase in the hierarchy is composed from a shared set of component terms, whereas the headings in the Iconclass hierarchy are predominantly unique descriptors. (See Section 1.2 and 2.) Additionally, Iconclass, with 28,000 total headings and growing,[43] is much larger than FARL's system. Iconclass may be amenable to the hierarchical multilabel classification framework, but a significant amount of semantic pre-processing of the headings would be required.

Within the hierarchical setting, a prior work by Belhi et al. uses deep learning on the hierarchical multilabel problem of the WikiArt, the Met, and Rijksmuseum datasets for a *two-level* hierarchy.[44] This hierarchy consists of (i) a general asset-type (for example, pottery, paintings, etc.) and (ii) specific characteristics for each asset-type (for example, predicting artist and style for asset-type "paintings"). The

42 Leendert D. Couprie, Iconclass: An Iconographic Classification System, in: *Art Libraries Journal* 8 (2/1983), 32–49.

43 Iconclass, URL: www.iconclass.org [last accessed: April 2, 2021].

44 Abdelhak Belhi/Abdelaziz Bouras/Sebti Foufou, Towards a Hierarchical Multitask Classification Framework for Cultural Heritage, in: *2018 IEEE/ACS 15th International Conference on Computer Systems and Applications (AICCSA)*, Aqaba 2018, 1–7.

authors tackled this problem by training two types of deep networks: one single-label classification network for predicting the asset-type, and another asset-specific multitask CNN[45] for each asset-type.

In contrast to the two-level hierarchy of Belhi et al., FARL's hierarchy could sometimes be five or six levels deep with hundreds of unique non-terminal[46] and terminal[47] branching points—many of which are relevant to only a handful of images. Thus, applying the algorithm of Belhi et al. to the FARL system would involve (i) training new CNNs at each non-terminal branching point and (ii) training a multitask CNN at each terminal point. Due to the depth of the hierarchy, hundreds of models would have to be trained. Not only are the computational costs for training hundreds of networks unattainable for most institutions, but also such an approach would train the parameters of each component network on a *subset* of the data, which would likely cause deterioration in performance. (See Section 3.4.) In contrast, the method described in Section 2 provides a solution that can predict all headings in the FARL hierarchy by training only *one* network on the *entire* prelabeled dataset.

Outside the field of fine art classification, researchers in text and image annotation as well as protein identification have extensively studied the hierarchical classification problem and, for the most part, relied on hand-engineered features that are input into decision-tree or max-margin classifiers. Nakano, Cerri, and Vens provide a thorough survey with insightful discussions.[48] Methods leveraging deep learning appeared only recently. Of particular interest are the works by Wehrmann et al.[49] and by Wehrmann, Cerri, and Barros,[50] which propose various convolutional and recurrent neural networks, respectively, for the hierarchical classification problem. Their approach relies on architectures markedly different than those pretrained for single-label classification. Thus, since most open source pretrained computer vision models were trained for the single-label task—usually on

45 Following a shared series of feature extraction layers, a multitask CNN branches into parallel linear layers that each performs a single-label classification or regression task such as artist or genre prediction.

46 A non-terminal branch point is a location in the hierarchy, associated with some prefix, such that there exists a next-level-down term that, once appended, will create another prefix that is not a full phrase.

47 A terminal branch point is a location in the hierarchy, associated with a prefix, such that appending any next-level-down term to the prefix will create a full syntactically valid phrase.

48 Felipe Kenji Nakano/Ricardo Cerri/Celine Vens, Active Learning for Hierarchical Multi-Label Classification, in: *Data Mining and Knowledge Discovery* 34 (5/2020), 1496–1530.

49 Jônatas Wehrmann et al., Hierarchical Multi-Label Classification with Chained Neural Networks, in: *SAC'17: Proceedings of the Symposium on Applied Computing*, Marrakech 2017, 790–795.

50 Jônatas Wehrmann/Ricardo Cerri/Rodrigo C. Barros, Hierarchical Multi-Label Classification Networks, in: *Proceedings of the 35th International Conference on Machine Learning*, 2018, 5075–5084.

ImageNet—their approach, unlike ours, cannot leverage transfer learning, which improves performance significantly in the low-sample size regime.

During the preparation of this chapter, we learned of a pilot project launched by Lincoln et al. at the Carnegie Mellon University (CMU) Libraries to tackle the tasks of visual similarity search, duplicate and close-match identification, and streamlining image tagging within CMU's General Photograph Collection (GPC).[51] This project resulted in a deep learning pipeline for these tasks that was further refined through testing by and feedback from CMU Libraries staff. In their white paper, Lincoln et al. discuss how computer vision might be used within Photoarchives to streamline cataloging and improve searching. Furthermore, the authors emphasize the importance of using a human-in-the-loop process to tackle these tasks to prevent error and correct any biases promoted by the training set. We strongly agree with these conclusions.

Yet, although the CMU team's tagging task appears similar to that introduced in this chapter, it relies on different methodologies. Our approach consists of training a network (starting from a network pretrained on ImageNet) to *directly predict* the heading associated with an unlabeled image while the CMU team's approach is to identify unlabeled images (by using an ImageNet-pretrained network that is not further trained on the GPC) similar to a particular labeled "seed" image, which the human editors must identify.

Our work and that of the CMU team require human experts to validate the accuracy of the suggestions at the end of the pipeline. As described by Lincoln et al., their pipeline requires that editors locate the seed images and, occasionally, it returns inconveniently large subsets of similar images. In contrast, our Zooniverse app directly presents images—one by one—with a predicted label and the human expert can choose "correct" by swiping right or "incorrect" by swiping left or input an alternative label. Thus, our direct-prediction approach to vetting results is significantly faster than that proposed by the CMU team. The white paper published by Lincoln et al. does not present the performance metrics of their pipeline, so a comparison of accuracy cannot yet be determined.

Section 5. Conclusion

In this chapter, we establish the considerable benefits of collaborations between AI researchers and cultural heritage preservationists. We also demonstrate that deep neural networks can be adapted to classify images according to specialized hierarchical, multilabel classification systems. This adaptation requires only simple

51 Matthew Lincoln et al., CAMPI: Computer-Aided Metadata Generation for Photo archives Initiative (2020), in: Carnegie Mellon University. Preprint, doi:10.1184/R1/12791807.v2.

modifications to the network's loss function and prediction rule, thus leaving the feature extractor unchanged and allowing for the utilization of pretrained models, through fine-tuning, to achieve better performance. Finally, we provide original experiments on digitized images in FARL's Photoarchive, thus indicating the validity and potential of this approach. As discussed above, we achieved a notable improvement in performance by incorporating the FARL classification system into the network through our proposed HCE loss and syntax-aware classifier.

These experiments further indicate that the accuracy for a given label depends logarithmically on the number of training images tagged with that label. For example, results in Section 3.4 suggest that roughly 100 training images are necessary to achieve 30%–50% success in canonical performance metrics, 1,000 images for 50%–80%, and 10,000 images for more than 80%. These experimental insights provide a promising point of reference for future partnerships between art historians and computer scientists.

As noted above, we anticipate that this technology will be ready for deployment in 2022 and once introduced, it is certain to streamline Library staff workflow by relieving photoarchivists of a time-consuming aspect of cataloguing. Thus, for the photoarchivists, the successful partnership as documented in this chapter demonstrates a clear advantage. For the computer vision researchers, the Photoarchive's in-house classification system motivates useful investigations in engineering the syntax of hierarchical, domain-specific languages into neural networks.

Bibliography

AGARWAL, Siddharth/KARNICK, Harish/PANT, Nirmal/PATEL, Urvesh, Genre and Style Based Painting Classification, in: *2015 IEEE Winter Conference on Applications of Computer Vision (WACV)*, Waikoloa, HI, 2015, 588-594.

BALDI, Pierre/SADOWSKI, Peter/WHITESON, Daniel, Searching for Exotic Particles in High-Energy Physics with Deep Learning, in: *Nature Communications* 5 (1/2014), 1–9.

BELHI, Abdelhak/BOURAS, Abdelaziz/FOUFOU, Sebti, Towards a Hierarchical Multitask Classification Framework for Cultural Heritage, in: *2018 IEEE/ACS 15th International Conference on Computer Systems and Applications (AICCSA)*, Aqaba 2018, 1–7.

BOTTOU, Léon/BOUSQUET, Olivier, The Tradeoffs of Large Scale Learning, in: J.C. PLATT et al. (eds.), *Advances in Neural Information Processing Systems 20 (NIPS 2007)*, 2007, 161–168.

BOTTOU, Léon, On-Line Algorithms and Stochastic Approximations, in: D. SAAD (ed.), *Online Learning and Neural Networks*, Cambridge 1998.

Cetinic, Eva/Lipic, Tomislav/Grgic, Sonja, Fine-Tuning Convolutional Neural Networks for Fine Art Classification, in: *Expert Systems with Applications* 114 (2018), 107–118.

Couprie, Leendert D., Iconclass: An Iconographic Classification System, in: *Art Libraries Journal* 8 (2/1983), 32–49.

Deng, Jia/Dong, Wei/Socher, Richard/Li, Li-Jia/Li, Kai/Li, Fei-Fei, Imagenet: A Large-Scale Hierarchical Image Database, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Miami, FL, 2009, 248–255.

Douze, Matthijs/Jégou, Hervé/Sandhawalia, Harsimrat/Amsaleg, Laurent/Schmid, Cordelia, Evaluation of Gist Descriptors for Web-Scale Image Search, in: *Proceedings of the ACM International Conference on Image and Video Retrieval*, 2009, 1–8.

Goodfellow, Ian/Bengio, Yoshua/Courville, Aaron, *Deep Learning*, Cambridge 2016.

He, Kaiming/Zhang, Xiangyu/Ren, Shaoqing/Sun, Jian, Deep Residual Learning for Image Recognition, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas 2016, 770–78.

Hentschel, Christian/Wiradarma, Timur Pratama/Sack, Harald, Fine Tuning CNNS with Scarce Training Data—Adapting ImageNet to Art Epoch Classification, in: *2016 IEEE International Conference on Image Processing (ICIP)*, Phoenix, AZ, 2016, 3693–3697.

Khan, Fahad Shahbaz/Van de Weijer, Joost/Vanrell, Maria, Who Painted this Painting?, in: *2010 CREATE Conference*, 2010, 329–333.

Knox, Katharine McCook, *The Story of the Frick Art Reference Library: The Early Years*, New York 1979.

Lecoutre, Adrian/Negrevergne, Benjamin/Yger, Florian, Recognizing Art Style Automatically in Painting with Deep Learning, in: Min-Ling Zhang/Yung-Kyun Noh (eds.), *Proceedings of the Ninth Asian Conference on Machine Learning, PMLR* 77, 2017, 327–342.

Li, Jia/Wang, James Ze, Studying Digital Imagery of Ancient Paintings by Mixtures of Stochastic Models, in: *IEEE Transactions on Image Processing* 13 (3/2004), 340–353.

Lincoln, Matthew, et al., CAMPI: Computer-Aided Metadata Generation for Photo archives Initiative (2020), in: Carnegie Mellon University. Preprint, doi:10.1184/R1/12791807.v2.

Nakano, Felipe Kenji/Cerri, Ricardo/Vens, Celine, Active Learning for Hierarchical Multi-Label Classification, in: *Data Mining and Knowledge Discovery* 34 (5/2020), 1496–1530.

Mensink, Thomas/Van Gemert, Jan, The Rijksmuseum Challenge: Museum-Centered Visual Recognition, in: *ICMR '14: Proceedings of International Conference on Multimedia Retrieval*, Glasgow 2014, 451–454.

Prokop, Ellen, Digital Art History for the Masses? The Role of the Public Digital Art History Lab, in: *Život umjetnosti: Journal for Modern and Contemporary Art and Architecture* 105 (2/2019), 196–213.

Saleh, Babak/Elgammal, Ahmed, Large-Scale Classification of Fine-Art Paintings: Learning The Right Metric on The Right Feature, in: *International Journal for Digital Art History* (2/2016), 71–93.

Sanger, Martha Frick Symington, *Henry Clay Frick: An Intimate Portrait*, New York 1998.

Shung, Koo Ping, Accuracy, Precision, Recall or F1? (2018), in: *Towards Data Science. Medium*, URL: https://towardsdatascience.com/accuracy-precision-recall-or-f1-331fb37c5cb9 [last accessed: April 2, 2021].

Strezoski, Gjorgji/Worring, Marcel, OmniArt: A Large-Scale Artistic Benchmark, in: *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 14 (4/2018), 1–21.

Strezoski, Gjorgji/Worring, Marcel, OmniArt: Multi-Task Deep Learning for Artistic Data Analysis (2017), in: *arXiv* [preprint], URL: arXiv:1708.00684 [last accessed: April 2, 2021].

Tan, Wei Ren/Chan, Chee Seng/Aguirre, Hernán E./Tanaka, Kiyoshi, *Ceci n'est pas une pipe*: A Deep Convolutional Network for Fine-Art Paintings Classification, in: *2016 IEEE International Conference on Image Processing (ICIP)*, Phoenix, AZ, 2016, 3703–3707.

Yosinski, Jason/Clune, Jeff/Bengio, Yoshua/Lipson, Hod, How Transferable Are Features in Deep Neural Networks?, in: Z. Ghahramani et al. (eds.), *Advances in Neural Information Processing Systems* 27, 2014, 3320–3328.

Wang, Dayong/Khosla, Aditya/ Gargeya, Rishab/Irshad, Humayun/Beck, Andrew H., Deep Learning for Identifying Metastatic Breast Cancer (2016), in: *arXiv* [preprint], URL: arxiv:1606.05718 [last accessed: April 2, 2021].

Wang, Xiaoyu/Han, Tony X./Yan, Shuicheng, An HOG-LBP Human Detector with Partial Occlusion Handling, in: *2009 IEEE 12th International Conference on Computer Vision*, Kyoto 2009, 32–39.

Wehrmann, Jônatas/Cerri, Ricardo/Barros, Rodrigo C., Hierarchical Multi-Label Classification Networks, in: *Proceedings of the 35th International Conference on Machine Learning*, *PMLR* 80, 2018, 5075–5084.

Wehrmann, Jônatas/Barros, Rodrigo C./Dôres, Silvia N. das/Cerri, Ricardo, Hierarchical Multi-Label Classification with Chained Neural Networks, in: *SAC '17: Proceedings of the Symposium on Applied Computing*, Marrakech 2017, 790–795.

Chapter 2: Web Archives and the Problem of Access: Prototyping a Researcher Dashboard for the UK Government Web Archive

Mark Bell, The National Archives, London | Tom Storrar, The National Archives, London | Jane Winters, School of Advanced Study, University of London.

Abstract

There is a burgeoning secondary literature concerning the use of the archived web as a primary source for Humanities research, but it remains centrally concerned with how to work around problems of scale, complexity and access. The manifold barriers encountered include the inability to download and take away data; the opacity of web harvesting processes; the (unknown) scale of content duplication; and the unsuitability of keyword searching as a primary means of exploration. An important record of the recent past remains tantalizingly out of reach for the majority of historians, political scientists, literary scholars and others.

This chapter will explore how a combination of Humanities methodological and research concerns, the expert knowledge of archivists, and machine learning solutions can work together to transform access to the open UK Government Web Archive (UKGWA). We will outline the theory behind and the steps towards building a prototype researcher dashboard for the UKGWA allowing multiple routes into and views of the archives of government online over more than two decades.

1. Introduction

This chapter begins by describing the history and current status of the UK Government Web Archive (UKGWA),[1] which is provided by The National Archives of the UK (TNA), before addressing some of the specific and more general challenges associated with archiving the web. It moves on to discuss the different ways in which researchers in the Humanities and Social Sciences might want to make use of the UKGWA, and outlines some of the factors that currently hinder, or prevent entirely,

1 http://www.nationalarchives.gov.uk/webarchive/ [last accessed: April 2, 2021].

optimal access. Next, it introduces the concept of co-design, involving archivists, researchers and technologists, as a means for developing useful and sustainable tools and modes of access for web archives in general, and the UKGWA in particular, before describing what a co-designed researcher dashboard might include. This section of the chapter also explores the kinds of research that would be enabled by the provision of a suite of tools sitting on top of the web archive, as well as the utility of such a service for the host institution. It concludes by reflecting on the value of, and mechanisms for, collaboration to enhance access to web archives.

1.1 What is the UK Government Web Archive (UKGWA)?

The UKGWA was established in 2003 by The National Archives, the official archive and publisher for the UK government, and for England and Wales. The rapid rise in the use of the web as a platform for disseminating information began in the mid to late 1990s and it had become clear that TNA, as an institution concerned with gathering the evidential record of government, and how the state interacts with the citizen, would need to collect public websites.

The initial archive was formed of a small number of key sites, and some content from the mid-1990s to mid-2000s was added, ingested from the Internet Archive.[2] The scope of the collection was then widened dramatically in 2008 to agencies and "arm's length bodies," and this collecting remit has remained in place ever since. The archive has evolved to archive resources published on other platforms, most notably social media, including Twitter, YouTube and Flickr.[3] This expansion led the latest of TNA's collection policy documents to refer to the collection's scope as the "UK Central Government Web Estate,"[4] broadening it from "traditional" websites.

While the collection remains limited to the UK central government, its departments, agencies, arm's length bodies, the National Health Service (NHS) and public inquiries, the UKGWA, as a web archive,[5] is a complex and varied collection, which is constantly accruing material and adapting to the capturing challenges of the present, while simultaneously accommodating the technology of the past. As

2 The Internet Archive "is a non-profit library of millions of free books, movies, software, music, websites, and more." At the time of writing, its Wayback Machine offers access to more than 486 billion archived web pages from around the world. https://archive.org/ [last accessed: April 2, 2021].

3 https://webarchive.nationalarchives.gov.uk/social/search/ [last accessed: April 2, 2021].

4 https://www.nationalarchives.gov.uk/documents/information-management/osp27.pdf [last accessed: April 2, 2021].

5 https://netpreserve.org/web-archiving/ [last accessed: April 2, 2021].

of autumn 2020, the UKGWA contains approximately 6 billion resources[6] across the 24 years of archives it hosts.

Beyond providing a record for posterity and future research, the UKGWA plays a key role in "Web Continuity."[7] This initiative seeks to reduce the number of broken links on government websites by providing public access to highly complete web archive snapshots, while also redirecting users to the web archive when a resource is no longer available on the original website. Furthermore, the archive has been used as a trusted home for websites closed in the process of consolidating government information onto websites such as gov.uk and previously Directgov and Business Link. It is the combination of these initiatives, and the fact that the UKGWA is open to anyone with an internet connection, that means the archive has many thousands of daily users, according to web server log analysis and Google Analytics.

The archive is updated continually through a number of collection processes, including scheduled captures of websites and exceptional, high priority captures, often in response to events of national significance. The latter collection method is often employed to make time-critical captures, for example the government's response to the COVID-19 pandemic, or the UK's exit from the European Union. This will be described in more detail later, as these factors have an influence on the use of the archive.

The UKGWA, as a well-used and trusted service, meets its core mission of capturing the published government record and providing access to it. As the collection has grown and matured, and the service is often and increasingly the only reliable source of this information, researcher interest in exploiting the collection has increased and there is every indication that that interest will continue to accelerate. While TNA has supported several research events and projects, these are normally large undertakings that require the production of tailored datasets (for example, Computational Archival Science[8] and Alan Turing Institute Data Study Group[9] events held in 2019).

A key element in supporting research is an understanding not only of what the collection contains, but how the collection came to contain it. The original content

---

6 A resource is anything with a Uniform Resource Locator (URL) and includes everything from HTML pages to images to the JavaScript files necessary to reproduce websites via replay software.

7 https://webarchive.nationalarchives.gov.uk/ukgwa/20130102170449/http://nationalarchives.gov.uk/information-management/policies/web-continuity.htm [last accessed: August 31, 2021].

8 https://blog.nationalarchives.gov.uk/network-analysis-of-the-uk-government-web-archive/ [last accessed: April 2, 2021].

9 https://www.turing.ac.uk/events/data-study-group-december-2019 [last accessed: April 2, 2021].

creators were government departments but the act of archiving the information relies on a series of complex interactions between human actors (members of various teams), the technologies used to create and capture the resources (web technologies and capture techniques) and the points in time at which they are captured.

1.2 Structure and Collection Process

In contrast to many other types of archives and collections, web archives do not normally enjoy the same degree of intellectual control in describing their contents. This is necessary because the services often need to prioritize capture at scale above description, in order to minimize the risk of loss.

However, there are sources of contextualizing information relating to provenance or how a resource may relate to others within or outside the collection. The real challenge is to capture, convert and convey this knowledge in a way that can be easily consumed by researchers.

The web itself inherently contains a rich amount of contextual information. These characteristics include URLs, linkages between resources and a wealth of other structures, from unstructured text to highly structured forms, such as XML. The vast majority of this is preserved in the web archive.

Aside from typical structural data that the UKGWA inherits from the resources it captures, the archiving team also has a selection of web archive specific tools available, such as CDX,[10] which presents some of this data in machine-readable formats. The UKGWA has other valuable sources of context too. First, the UKGWA uses a database to manage the archiving process and stores archivist decisions and explanations. These might include, for example, the reason a website was archived on a particular day, which has enormous potential for providing a rich commentary, and may help users to make sense of the shape of the archive.

Second, XML files generated by this database, which act as messenger files between it and the crawler, contain specific technical information for each web crawl, such as "include" and "exclude" rules which often change between crawls. Therefore, associating each XML file with its respective "snapshot" crawl may be desirable. Third, TNA's catalogue service, Discovery,[11] contains a wealth of knowledge relating to the government bodies responsible for each website (i.e. domain) in the collection. Also of significance is that Discovery holds millions of other records at The National Archives, reducing any barrier between them and the web archive. This again conveys essential context that not only explains why a website was se-

10 https://archive.org/web/researcher/cdx_file_format.php [last accessed: April 2, 2021].

11 http://discovery.nationalarchives.gov.uk/, see e.g. http://discovery.nationalarchives.gov.uk/details/r/C16668 for a specific series-level description [last accessed: April 2, 2021].

lected but what happened to it, and where it fits into the wider patchwork of the collection.

Human actions involve decisions on when and how to capture a resource or a website but also why that effort was made. Data on this is kept as part of the archive but most of it is not public, being historically considered purely "administrative" in nature. However, being that web archives are created through actions and decisions, both human and machine, these are rooted in the time and the context in which they are made. It is often necessary for a web archivist to modify rules to include or exclude elements to successfully capture a resource or set of resources. This may be to avoid crawler traps[12] or, more often, to expand the scope of a capture so that it captures a sub-domain, or some externally-hosted content, pertinent to the website. These decisions and rules are easy to implement but can have a significant bearing on the completeness of the archive, the boundary around it, and ultimately on its users' abilities to comprehend it.

As the model trusted to comprehensively capture the published government record, the web archive needs to be of sufficiently high quality. To achieve this, quality assurance is performed by members of the web archiving team. Using a mixed approach of manual and automated methods, tools and experience, web archivists verify the capture of content and its rendering in replay tools. Web archivists do, however, need to prioritize certain aspects of quality assurance, for example capture over some elements of replay. While this is no surprise, as web archiving is a "lossy" process, most current tools and approaches keep it to a minimum in the UKGWA. Decisions relating to this are recorded in some way, be it via checklists, logs, database systems, or archivists' notes. However, these are often only understood by trained web archivists and therefore would not necessarily facilitate greater understanding of the collection.

The issue of scale makes it necessary for high-volume capturing, a process which is not always compatible with producing and disseminating detailed information about the collection process. A good example of this is an average crawl of the gov.uk[13] website, which is archived monthly, and contains over 1.8 million resources. This also largely explains why the UKGWA is only catalogued at TNA at website level. However, a dashboard could still exploit tools intrinsic to the web archive: data from CDX, and potentially from logs generated by the crawler, could

12 "A crawler trap is a set of web pages that create an infinite number of URLs (documents) for [a] crawler to find, meaning that such a crawl could infinitely keep running and finding 'new' URLs." One example of a crawler trap is an online calendar with an almost infinite date range. https://support.archive-it.org/hc/en-us/articles/208332943-Identify-and-avoid-crawler-traps- [last accessed: April 2, 2021].

13 https://webarchive.nationalarchives.gov.uk/ukgwa/20200901093455/https://www.gov.uk/ [last accessed: August 31, 2021].

be presented to show when, how and why a resource was captured on a particular date. Machine learning and AI techniques are likely to be extremely useful in addressing these challenges in the future. However, we already have many promising routes to exposing some of this valuable context, from static dataset files (for example, csv files) to services that support querying to produce machine-readable and "at-scale" data (for example application programming interfaces, or APIs). It is important that these approaches are documented openly, allowing for collaboration between web archiving institutions and a common understanding among researchers of their potential use. Such an approach may lead eventually to forming widely-adopted conventions, or even standards, that will support researchers moving between collections without having to navigate the nuances within each separate collection.

Beyond publishing explanatory metadata, capturing and conveying useful data relating to decision making is challenging in a number of ways. It is therefore likely to be desirable for the web archive tools to do the "heavy lifting" with the dashboard, providing useful functionality to ease digestion of that information. Nevertheless, it is worth briefly discussing some of the challenges that need to be overcome.

In common with all web archives, usability is a difficult challenge and our research supports the notion that new users often need to spend some time with the web archive before becoming confident in using it. The provision of the data and metadata described will be driven by collaboration between the web archiving team and researchers as they use the collection.

The UKGWA is not only well used but also serves a broad user base. A combination of online and "in person" user testing projects have shown users range from members of the public seeking historical reports or data to journalists wishing to see the evolution of policy on a particular topic; from civil servants researching previous policies to solicitors accessing historical guidance.

1.3 What Do Researchers Want to Do with the UKGWA?

The majority of national web archives impose access restrictions, ranging from complete closure to the public (for example, in Sweden) to off-site access for bona fide researchers located in the host country (for example, in Denmark).[14] The most common form of restriction tends to arise from the existence of Legal Deposit Legislation, which allows the harvesting of national web domains at scale but often limits access to browsing only on the premises of the archiving institution. The challenges that this poses for researchers and other users have been well

14 International Internet Preservation Consortium, Legal Deposit, n.d. https://netpreserve.org/web-archiving/legal-deposit/ [last accessed: April 2, 2021].

documented.[15] The UKGWA, however, like the Croatian, Portuguese and Icelandic web archives, permits unrestricted access online to its data, from anywhere in the world. Researchers can readily consult this essential primary source for the history of the late twentieth and early twenty-first century, either through its own search interface or via TNA's Discovery catalogue. As noted above, a full-text search is available, and there is also a browse option for those who are more familiar with the structure of the UK government and its departments and ministries. Given the restrictions that exist for other web archives, the value of this open access, and the broad permission to copy and reproduce that arises from Crown Copyright, should not be underestimated.

Access by means of a public search interface meets many of the needs of users of the UKGWA. It is easy to search for a particular government report if you roughly remember its name (even if you cannot, you might eventually find it); you can carry out a case study of the Department for Culture, Media and Sport by locating it on the site browse list (which handily takes account of the fact that it was renamed the Department of Digital, Culture, Media and Sport in 2017). But the limitations of search for an archive of this size soon become apparent. The UKGWA interface gives "Budget 2010" as an example search. At the time of writing, without placing the phrase in double quotation marks, this generates 100,525,737 results. Even with the quotation marks in place, so that the full phrase is being searched for, 105,052 results are generated. These results, which are not presented in any particular order or ranking, can be browsed and read 25 results at a time, which is not a task that is reasonable for anyone to undertake.[16] Overwhelmed by volume, and with no means to extract and refine the data offline, the researcher effectively hits a dead end.

Quite apart from the challenge posed by scale, and the limitations of in-browser searching, the modes of access currently available to the user fail to provide a sense of the scope of the archive. It is one thing to know that more than 5,000 websites have been archived between 2003 and 2020,[17] but what kind of information does this include, how often has data for particular sites been collected, and how has the UK government web estate changed over time? This is true for many digital

15 See, for example, Ian Milligan, Web Archive Legal Deposit: A Double-edged Sword, 14.07.2015, URL: https://ianmilli.wordpress.com/2015/07/14/web-archive-legal-deposit-a-double-edged-sword/ [last accessed: April 2, 2021]; Jane Winters, Giving with One Click, Taking with the Other: Electronic Legal Deposit, Web Archives and Researcher Access, in: Melissa Terras/Paul Gooding (eds.), *Electronic Legal Deposit: Shaping the Library Collections of the Future*, London 2020, 159-178.

16 At the time of writing, a Google search for "Budget 2010," without the enclosing quotation marks, returns 1.8 billion results (516,000 with the quotation marks), but the top few are highly relevant.

17 The National Archives of the UK, How to Use the Web Archive, n.d. http://www.nationalarchives.gov.uk/webarchive/information/ [last accessed: April 2, 2021].

archives, but in the case of the UKGWA and other web archives the position is further complicated by the vagaries and particularities of the crawling process through which data is harvested. There are multiple levels of complexity here, not least the fact that parts of the UKGWA (pre-2003) are derived from the Internet Archive, which has its own crawling protocols and criteria for collecting websites on the gov.uk domain.[18] This archival and algorithmic context is essential for researchers approaching the UKGWA with anything more than an interest in a single government report or news announcement.

What, then, would help researchers to make the most effective use of the UKGWA? Building on the foundation of open data, we can envisage a researcher dashboard catering for a wide range of use cases and accommodating different levels of technical expertise. A non-exhaustive list of user requirements might include: access to metadata and statistics; the ability to export different kinds of data from the archive (metadata, images, page content stripped of menus, headers and footers); tools for analyzing trends in the data, for example linguistic and cultural change; the option to analyze online networks of government and the flow of information between departments; and visualization tools assisting both navigation and analysis. The context for this data would also be presented to the user, allowing them to explore the ebbs and flows of archiving the gov.uk domain, which is affected by changes in technology and web design as much as by political crises and the transfer of power between administrations.

We would argue that the prototyping of a suite of tools of this kind requires collaboration between archivists, researchers and technologists. The value of collaborative working and co-design for web archives has already been demonstrated by the "Big UK Domain Data for the Arts and Humanities" (BUDDAH) project, whose co-created SHINE interface influenced the development of online access provision for the UK Web Archive at the British Library.[19] The difficulties of contextualizing, accessing and analyzing the archived web are too complex to be addressed by individuals or single institutions, and there is now an opportunity to bring together the three main stakeholder groups to design tools and services that will be robust, flexible, customizable and sustainable. This requires long-term collaboration and engagement: researchers' needs will change over time, as will the challenges of archiving an ever-moving digital target.

---

18 The National Archives of the UK, Information on Web Archiving, n.d. https://webarchive.nationalarchives.gov.uk/ukgwa/20170608213215/https://www.nationalarchives.gov.uk/webarchive/information.htm [last accessed: April 2, 2021].

19 For more information about the BUDDAH project (funded by the Arts and Humanities Research Council, grant reference AH/L009854/1), see Josh Cowls, Cultures of the UK Web, in: Niels Brügger/Ralph Schroeder (eds.), *The Web as History: Using Web Archives to Understand the Past and Present*, London 2017, 220-237.

The suggestions for a dashboard for the UKGWA that follow are the result of a number of meetings and conversations between the authors of this chapter, who bring the perspectives of, respectively, a research software engineer, a web archivist and a Digital Humanities researcher. A different configuration of contributors would no doubt result in other proposals; and not everything that is outlined below would be possible for all web archives. The list does, however, serve as a starting point, highlighting key research themes but adopting a realistic approach to what can be achieved within archiving institutions which have limited resources at their disposal, and multiple competing priorities.

2. Towards a Prototype Dashboard

Our prototype dashboard will allow the researcher to define the scope of their analysis along three dimensions: breadth, depth, and temporality. Breadth is the selection of websites to be included, depth defines the parts of each website to be included, and temporality defines the period of analysis. Brügger proposes five strata to delimit the web as an object of study: web element (for example, a piece of text or an image on a page), web page, website, web sphere (web activity related to an event, concept or theme) and the web itself.[20] The first two map to the depth dimension, selecting the types of content (elements) and setting selection criteria for the pages. For example, the depth could be defined as all hyperlinks appearing on homepages, the text content of accessibility pages, or images extracted from a random sample. Depth could also relate to an entire site, which is the third of the strata and also the minimum value for the breadth dimension. The web sphere, first proposed by Steven Schneider and Kirsten Foot, is defined as activity related to an event, concept or theme, and aligns with the breadth dimension.[21] We would expand the configuration of the breadth dimension to enable the definition of any subset of sites, including random sampling, but the web sphere idea of filtering according to a theme rather than an explicit list of sites is an important feature.

The temporal dimension is arguably the one that sets exploration of the web archive apart from that of the live web. The archive is constructed from multiple snapshots of web pages, the distance between snapshots differs by domain and can be influenced by events, and a snapshot is taken irrespective of whether the page has changed since it was last captured. The depth of crawl can change over time

20 Niels Brügger, Website History and the Website as an Object of Study, in: *New Media & Society* 11 (1-2/2019), 129, doi:10.1177/1461444808099574.

21 Steven M. Schneider/Kirsten A. Foot, Web Sphere Analysis: An Approach to Studying Online Action, in: Christine Hine (ed.), *Virtual Methods: Issues in Social Science Research on the Internet*, Oxford 2005, 157-170.

so a page archived in one crawl may have been missed in the previous one, and may not be crawled again. The researcher should be able to define the temporal aspect of their analysis by a single point of time (the closest snapshot to 1/1/2017), a range (all snapshots in the year 2015) or combinations thereof (closest snapshots to 1 January every year, all snapshots in the first quarter of each year).

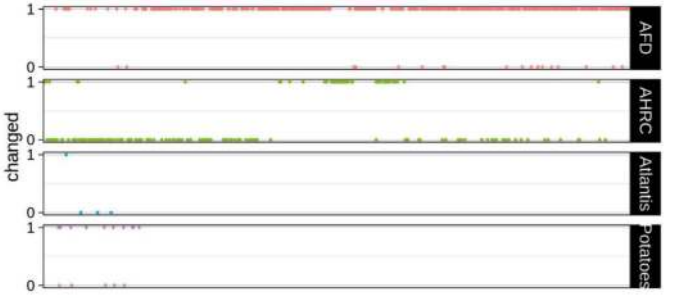
An obvious starting point for a dashboard is summarized aggregations, in tabular or visual form. Using existing data held in the UKGWA in CDX format, it is relatively straightforward to produce summary statistics, or bar and line charts summarizing page/resource captures, optionally by domain, over time. These summaries give the researcher a sense of scale but they are open to misinterpretation without an understanding of the capture process. For example, setting the dimensions to homepages of all websites and all snapshots from 2003 to 2016 produces counts showing a rise from just 80 captures in 2003 to 20,429 by 2012. This value almost halves in 2014 (10,511 captures) and halves again by 2016 (5,300 captures). This can be somewhat explained by understanding frequency of capture during the period, rising from an average of only 1.4 captures per page in 2003 to a peak of 9.57 in 2012, and falling to 3.43 by 2016. The decrease in volume since 2012 was also driven by a trend of centralization of government on the web towards the gov.uk domain, with 2135 unique home pages captured in 2012, falling to 1383 in 2016. During this period the volume of resource captures has increased exponentially.

Returning to the search problem raised earlier, the dashboard could summarize search results rather than presenting a list. Analysis of the first 10,000 results of the "Budget 2010" search found 7,731 unique URLs. The most common one was www.gov.uk/government/publications/budget-2010, which appears 27 times and in this case would be a good result, although it is important not to conflate the number of snapshots with relevance. It first appears on page 4 of the search results when the search is set to return 100 results per page. Remarkably, 5,038 of the results were for URLs in the www.cotswold.gov.uk domain, the website of an English district council. The dashboard could allow the researcher to filter out domains they deem irrelevant, and if this were coupled with returning only one result per URL (with the ability to view all snapshots) the task of sifting through search results could be greatly reduced. A more sophisticated approach would be to use page contents to group them by subject matter and enable a more semantic search.

Rather than analyze overall volumes the researcher may wish to visualize change. CDX files contain checksums, file sizes, and mime types for every captured resource in the archive. Using CDX data we can visualize the capture frequency of a page and derive an indicator showing whether the checksum has changed between snapshots. Fig 2.1 shows four such visualizations, for www.armedforces-

day.org.uk (AFD), www.ahrc.ac.uk (AHRC), www.projectatlantis.net (Atlantis) and potatoesforschools.org.uk (Potatoes).[22]

Fig 2.1: Changes over time derived from CDX files. Crown Copyright, licensed under the Open Government Licence.



The x-axis ranges from 20080604224039 (4 June 2008 at 22:40:39) to 20201006224341 (6 October 2020 at 22:43:41), and the y-axis for each graph is a binary value where 1 indicates that the checksum differed from the previous snapshot. Each page shows a different pattern of activity. Both AFD and AHRC were frequently (but not uniformly) captured throughout the period while the capture of Atlantis and Potatoes ceased in 2010 and 2011 respectively. While AFD appears to be under almost constant change, the AHRC page appears almost static for the first half of the period, is more active in the third quarter, and then returns to being static. The Potatoes site seems to experience intermittent change between less frequent captures, while Atlantis's activity pattern suggests a site which was first captured at the end of its active lifetime. While this is useful to understand archiving activity for a page and gives an idea of how dynamic a page may be, it is misleading in its current form and does not give an idea of what has changed.

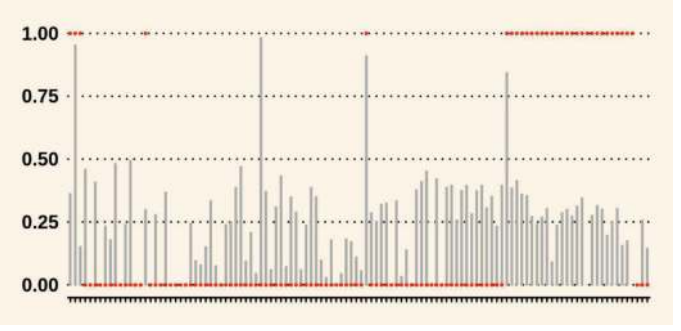
The visualization is misleading because the AHRC website was active throughout the period, and the long run of unchanged checksums is in fact due to the page being redirected for the majority of snapshots. The checksum is therefore of a redirection message and not main site content. This suggests pre-processing is required before creating a visualization of this kind, to include redirections in the

22 Two of the URLs are self-explanatory - Armed Forces Day and Potatoes for Schools. The AHRC is the UK's Arts and Humanities Research Council; the Atlantis Initiative was "a public-sector initiative to understand the underlying issues and agree the standards to collectively provide interoperable base geographic and environmental datasets to better support water management in flooding and water quality for the twenty-first century."

analysis. This is not straightforward, however, since www.ahrc.ac.uk redirects variously to www.ahrc.ac.uk/Pages/default.aspx, www.ahrc.ac.uk/Pages/Home.aspx, and ahrc.ukri.org, and while it would not be unreasonable to disambiguate the first two with the home page, it is not obvious that the third should be treated in the same way, as it is now in the ukri.org domain. Similarly, www.eatwell.gov.uk redirects to http://www.nhs.uk/Livewell/healthy-eating/Pages/Healthyeating.aspx from April 2011. Using checksums to identify change is a blunt tool since the checksum of a file will change if only a single character is amended.

Fig 2.2 shows changes in hyperlinks on the www.ahrc.ac.uk home page over time, with the checksum changes from Fig 2.1 overlaid.

Fig 2.2: Link structure changes over time for www.ahrc.ac.uk. Crown Copyright, licensed under the Open Government Licence.



The x-axis of the graph represents snapshots, as before, while the y-axis measures the Jaccard distance of a page's links versus those on the previous snapshot. The Jaccard similarity score is a ratio of the number of items in common between two sets against the number of distinct items in the sets. The Jaccard distance is one minus the similarity, and so if the hyperlinks are identical between two pages the score is zero, while complete difference results in a score of one. The graph suggests that changes occur frequently and that there have been four occasions (the bars above 0.75) involving a major restructuring of the web page. Hyperlink based analysis can be performed using WAT files which contain metadata extracted from WARC files.[23] To improve these graphs further, they could be annotated with the aforementioned administrative data to provide context for the peaks and troughs of capture activity. The dashboard will need to allow researchers to combine different methods in this way in order to interrogate the archived data effectively.

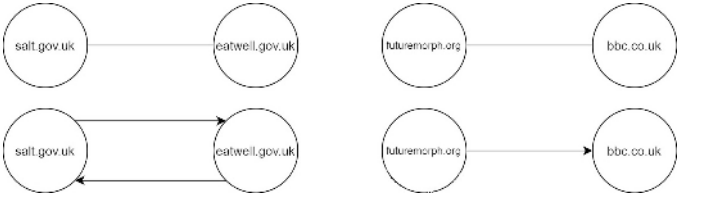
23 https://support.archive-it.org/hc/en-us/articles/360039686611-Web-Archive-Transformation-WAT-files [last accessed: April 2, 2021].

A second type of analysis enabled by WAT files is network analysis. A network graph can be built by representing each web page with a node (visually represented as a circle), and creating an edge (visually a line) between two nodes (or pages) if they are connected by a hyperlink. The graph can be directed or undirected. In a directed graph edges can be thought of as arrows going from the source page of a hyperlink to the page it is linking to, so that a pair of nodes can have up to two edges in opposing directions between them. In the undirected case a link between two nodes would indicate that at least one of the pages has a hyperlink to the other.

It may seem an obvious point but it is important to be aware that in the directed case that there will not be incoming links from web pages which are not in the archive. This is always the case for sites outside the government web domain but can also apply to websites of which the archive was unaware, or pages which no longer exist and have not been archived.

Fig 2.3 illustrates the difference between a directed graph and undirected for two scenarios. The top row shows two pairs of nodes connected in an undirected graph. In the bottom row, the directed case, the pair on the left are connected in each direction because both sites link to each other, while on the right there is an arrow only in one direction because bbc.co.uk is not archived in UKGWA (it may still have linked back to futuremorph.org[24] but we do not have that evidence).

Fig 2.3: Connection of nodes in undirected and directed graphs. Crown Copyright, licensed under the Open Government Licence.



The previously mentioned Computational Archival Science (CAS) workshop explored network analysis of the UKGWA, comparing network structure at different points in time. The dataset consisted of hyperlinks from pages up to a depth of 2, i.e., homepages and the pages linked to by the homepages. Network graphs were generated at the page level and from an aggregated dataset at the domain level.

This experimentation leads to an important question, and a challenge. What is being visualized and summarized in the dashboard? The charts in Fig 2.1 and 2.2 were based on individual web pages, the network graphs from the CAS workshop

24 https://webarchive.nationalarchives.gov.uk/ukgwa/20100730145942/http:/www.futuremorph.org/ [last accessed: April 2, 2021].

were summarized at the domain level. This summarization involved identifying child nodes of the home pages, i.e., pages linked to from the home page which were in the same domain, and linking two home pages if they or at least one of their children were linked. The UKGWA holds over 6 billion archived resources and the ability to aggregate is essential to making clear visualizations, or finding meaningful patterns through analytical means. This is the idea behind the *Historian's Macroscope*, which begins by envisioning a researcher zooming in and out of the archive as they follow different paths of inquiry.[25] The macroscope was first proposed by Joel De Rosnay as a theoretical tool for the study of large complex systems, analogous to the microscope or telescope.[26] Staying with the microscope analogy, we can imagine beginning with blots on a slide, one for each website, visible to the naked eye. At full magnification we see the atomic level, billions of web page elements. What do we see at intermediate magnifications, and how many steps are there between zero and maximum?

Starting at the level of the web page, the object of study consists of multiple elements including structural objects (for example, navigation menus), textual content, and images. These can all be extracted from the HTML and treated as text, in the case of images by extracting a label from the HTML or by using machine learning to generate one. The navigation menus can cause a problem as they are repeated throughout a website, creating unwanted noise and duplication of information. A common approach to this problem is boilerplate removal which strips out the menus leaving only text behind.[27] This can be problematic on home pages, for example, in which almost all of the content consists of hyperlinks, and boilerplate removal can return an empty page. Schneider and Foot suggest that studies of web content that overlook the structuring elements of a page or site are limited.[28] Rather than removing boilerplate we suggest that it should be identified and treated as a contextual object.

The website as an object of study presents a challenge, particularly when the breadth of analysis includes multiple sites. How do we represent a website, should it be treated as the sum total of its content, or is the structure important? The homepage is a high-level summary of a site but it is too distant from the content, while working with every page can provide too much detail. To study it as an object, it needs to be organized or aggregated. We have already considered one form

25 Shawn Graham/Ian Milligan/Scott Weingart, *Exploring Big Historical Data: The Historian's Macroscope*, London 2015, 1-2.

26 Joël De Rosnay, *The Macroscope: A New World Scientific System*, New York 1979, xiii.

27 Marco Baroni et al., Cleaneval: a Competition for Cleaning Web Pages, in: *Proceedings of the Sixth International Conference on Language Resources and Evaluation* 2008, 1.

28 Steven M. Schneider/ Kirsten A. Foot, The Web as an Object of Study, in: *New Media & Society* 6 (1/2004), 114-122, doi:10.1177%2F1461444804039912.

of organization, a network graph, but there are also three sources of hierarchical structure available.

First, the URL provides a natural hierarchy as it maps to a physical folder structure. Second, navigation menus generally form a tree with top-level items and submenus below them. Finally, there are breadcrumb trails placing a web page at the end of a retraceable path through the site.[29] The difficulty that these three sources pose is that they often present a different picture. The URL is influenced by the technical architecture of the website, depending on whether it serves up static or dynamic content, the underlying web framework used, and whether it uses a service-based architecture. As an example, PDF files on gov.uk are found in the sub-domain https://assets.publishing.service.gov.uk/ (not a page in its own right) but they can also be found under https://www.gov.uk/government/publications/ (either a navigational page or redirected to a search page depending on the snapshot). Neither of these provides a meaningful context for analyzing documents. Navigation menus provide more meaning and even though they may differ in form, they tend to be consistently structured across a site. The gov.uk website has presented menus in a number of ways over its eight-year lifespan but generally follows the pattern of providing links to high-level functions (e.g. Business, Driving), or shortcuts to popular services (e.g. Registering a company, driving licence applications) on the home page, and then a more traditional tree-style side menu for the rest of the site. The limitation of the navigation menu is that it does not necessarily lead to every page or resource on the site. So while it may provide a good structure around which to build a macroscope, there is more work to fit all of the pages within that structure. While this could be achieved using the hyperlinks in the WAT file, it becomes difficult when two pages from different branches of the menu link to the same page. Breadcrumb trails, where they exist, could fill this gap since they place a site in context and, if the two align, within the menu structure. Government guidance on applying for leniency for cartel members is found as a page under https://www.gov.uk/guidance, so in this case the URL provides a small amount of meaningful context (it is guidance). The breadcrumb on the 20190102181627 snapshot provides far more context, placing it under "Business and Industry" – "Business Regulation" – "Competition" – "Competition Act and Cartels."[30] Unfortunately, "Business and Industry" is not an option in the main navigation menus elsewhere in the site, which use "Business and self-employed" instead. The page originally belonged under the sub-domain of the Competition and Markets Authority and then moved under a

29 Breadcrumb trails help users to keep track of their position within a website. They typically appear at the top of a web page, and allow users to retrace their steps within the information hierarchy.

30 https://webarchive.nationalarchives.gov.uk/20190102181627tf_/https://www.gov.uk/guidance/cartels-confess-and-apply-for-leniency [last accessed: April 2, 2021].

section called Competition (according to the breadcrumb), which could be found under a "topic" menu (according to the URL) which itself could not be navigated to from the home page.

This single example tells a story of government on the web, which needs to be understood by the researcher and should be conveyed through the dashboard. The government web estate is constantly evolving as sites undergo architectural and structural redesigns, and responsibility for government functions moves within and between departments, which themselves may merge, close down, or be created. It also demonstrates that creating a hierarchical structure which would provide a lens through which to zoom in and out is non-trivial, and any attempt to do so will require assumptions to be made, and pre-processing steps which will change the form of the data, all of which must be transparently presented to the researcher. If such a hierarchy can be created then experimentation, including at a small scale with web data, by The National Archives has shown how the hierarchy can be navigated by summarizing the data upwards.[31] Using this approach, a level in the hierarchy is represented by an aggregation of the levels below it, rather than the text of an individual page. Further experimentation is needed to test the efficacy of this approach at scale.

An alternative is to use clustering techniques to group pages according to the similarity of some attributes. In this approach, each page is converted to a numeric form such as word frequency counts or a more sophisticated vectorized form which places each page in a 300 dimensional space.[32] Another method is to use topic modelling to identify a set of topics within the entire corpus and then classify each page according to its topic composition.[33] Pages are clustered based on some measure of similarity (for example, cosine similarity is common).[34] and a suitable number of clusters is defined either by the user or according to some optimality criteria. Again, the numeric representation selected and the clustering methodology need to be explained in a way that is understandable to the researcher. This does not mean explaining the inner workings of the algorithms and the underlying mathematics, but rather giving an understanding of high level concepts and how choices of representation and clustering technique influence the results.

Previously we suggested the ability to define web spheres around a theme was an important feature of the dashboard. A researcher may wish to define their own

---

31 Mark Bell, From Tree to Network: Reordering an Archival Catalogue, in: *Records Management Journal* 30 (3/2020), 379-394, doi:10.1108/RMJ-09-2019-0051.

32 Tomas Mikolov et al., Efficient Estimation of Word Representations in Vector Space, arXiv:1301.3781 (2013), URL: https://arxiv.org/abs/1301.3781v3 [last accessed: April 2, 2021].

33 David M. Blei/Andrew Y. Ng/Michael I. Jordan, Latent Dirichlet Allocation, in: *Journal of Machine Learning Research* (3/2003), 933-1022.

34 https://programminghistorian.org/en/lessons/common-similarity-measures [last accessed: April 2, 2021].

sphere by curating a set of websites which will be the subject of their analysis. This may be a difficult manual task as individual websites may not be relevant but subsections of those sites may be. For example, a study of health advice around salt consumption would include the whole of www.salt.gov.uk but only a small portion of food.gov.uk and the healthy eating section of nhs.uk. A better method would be to seed an automated approach, by compiling a list of keywords, by selecting a few exemplar websites, or perhaps by using the Wikipedia entry for an event of interest. The web sphere idea could mitigate against the complexity identified in the extraction of hierarchies by curating pages around a concept, which could be a government function without explicitly defining its position in a hierarchy.

Specifying the temporal dimension initially appears easy but its impact on results is influenced by the capture process and must be understood. An earlier example suggested selecting snapshots closest to a specific date. This sounds simple but there is nuance, such as whether there should be a limit to how far from the date a snapshot can be. If the date is 1 January 2016, is a snapshot from 2014 still relevant? Should the October 2015 version of a page take precedence over the February 2016 version considering the latter may not have existed in January? What if October in that example was replaced with July? The ability to set rules that answer these questions must be included in the dashboard. They will be informed by graphs such as Fig 2.1, so that the researcher can understand rates of capture and change for sites in their sphere of interest. An analysis of government activity during a specific month, for example, could neglect many sites which were captured at 3 or 6 month periods. Perhaps more complex rules could be defined to select the "best" snapshot from a time period. The rules can then be tested by visualizing coverage of the corpus against the other two dimensions. When working across periods of time, rules are also required for dealing with duplication, which is prevalent in the web archive. The options include selecting a single (first, last, middle) version of a page, removing duplicate versions, removing near duplicates (according to some threshold of nearness) and removing those where content is unchanged. Pages may be removed from the analysis if they are unchanged for more than some period of time, or based on analysis of hyperlinks not navigable to from other pages. These rules can be summarized as those classifying pages as active, static or dormant.

There is a lot of hidden complexity involved in defining the three dimensions, so the dashboard would include default settings which can all be adjusted. The settings used should also be exportable in an open standard so that researchers can not only easily publish them alongside their analysis but also share them with other researchers. Reproducibility should be at the heart of the design. All visualizations, statistical summaries, and intermediate representations, such as word frequency lists and topic models, should also be exportable, but content in its original form may need to be controlled. The UKGWA, as noted above, is fortunate to be an openly accessible web archive. That said, there are still risks in allowing large exports of

archival material, particularly related to take down requests. There is also the issue of scale, with sites like gov.uk comprising over 2 million pages. This would be a massive download, and it would be unlikely that any archive would sanction a tool that allows their collection to be extracted at such a scale.

This leads to the next question: where is the computation performed? The Hathi Trust has an interesting model which balances their requirement to protect copyright with their aim of opening the data to researchers.[35] They make pre-processed representations of their data, such as word lists, openly available but access to the original material is through a protocol known as non-consumptive research. Instead of the researcher taking the data to their tools, they bring their tools to the data. In both cases our dashboard would be initially used to define the scope of the data, by configuring parameters along the three dimensions. In the first case, the pre-processed data could be filtered and then downloaded for further analysis by the user on their own computer. In the non-consumptive case more complicated analysis can be performed against the archive in its original form (the WARC files themselves). This does, however, mean that the computation is on the archive's infrastructure which raises the question of who pays? Charging models for workflows which will probably involve machine learning are difficult to define. In the case of a web sphere defined around a concept, it will not be possible to calculate the size of the data in advance. A deep neural network model may not converge and therefore produce no results, meaning hours of wasted computation. If that algorithm was built by the archive but the data was defined by the user where does "fault" lie?

The non-consumptive model of a researcher running their own code against the archive is an advanced form of the dashboard. What would a first version look like and how much is already available in the web archiving community? Analysis starts with a definition of scope. Depth could be defined in the same way as a web crawl, following links from the home page, then following those links, and so on. Optionally links outside the website of interest could be followed. The process should be incremental so that the researcher can receive feedback on the number of pages they are including in their analysis, and also an indication of how many links were not followed (either because they were outside the site or not archived). This functionality uses the data in the WAT files and is technically well understood. The researcher will also be able to define the elements of interest, which at first will be hyperlinks or text. Initially the breadth would be restricted to selecting individual websites, or sub-domains within sites. The UKGWA already has an A-Z browsable list which could perform this role.[36] The temporal dimension will be defined by either a single point of time, with thresholds for the allowable time periods either

35 Jacob Jett et al., The HathiTrust Research Center Workset Ontology: A Descriptive Framework for Non-consumptive ResearchCollections, in: *Journal of Open Humanities Data* 2 (2016).

36 https://nationalarchives.gov.uk/webarchive/atoz/ [last accessed: September 1, 2021].

side, or a date range. Having defined the scope the collection will be visualized so that the researcher can understand the relative sizes of each site, and how often each has been collected (in the case of a date range being specified). A network graph will be viewable either at the page level or website level. The text content of each page will be searchable and keyword searching can be used to filter the collection, with visualizations updating accordingly. This prototype version offers some extra functionality that is not already available but it is really the prompt for a conversation with web archive researchers to understand how they would interact with the archive as data, and with computer and data scientists to tackle the complex challenges of enabling macroscopic analysis at scale.

Thankfully, we are not starting from zero if we want to build a dashboard; there are already great tools available to build on. The Archives Unleashed project has built an open source toolkit which enables the large scale processing of WARC files.[37] The toolkit is aimed at advanced users who are comfortable working with the command line and either of the programming languages Scala and Python. Recognizing that most researchers will not have the necessary programming skills or computing power to handle large scale collections, they also have a cloud service. This service can generate network graphs and summary statistics, but it is currently only available to Archive-It subscribers. The GLAM workbench project has developed a number of Python notebooks for extracting and visualizing data from four web archives.[38] Rather than working with WARC files, they instead use the APIs of the archives to access CDX data, and analyze changes over time using the Memento protocol.[39] While not an integrated tool, the notebooks provide much of the functionality that would form the basis of a dashboard in terms of selecting snapshots, extracting text and visualizing changes over time. They are intended to encourage researchers to explore the possibilities of web archives and to understand the data they contain. Although the developers claim some level of scalability, using an institution's API may have limits and the big data tools of the Archives Unleashed toolkit may be more appropriate for large scale analysis. What the developers have also highlighted is that not every notebook works for each of the four archives, and the implementations of the APIs mean that the data that comes back from a particular query may differ for each archive, so some adaptation may be needed to apply them to the UKGWA. What is key to all of these initiatives, and to the UKGWA's approach to tool design, is the open sharing of code so that the

37 Nick Ruest et al., The Archives Unleashed Project: Technology, Process, and Community to Improve Scholarly Access to Web Archives, in: *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020 (JCDL '20)*, New York 2020, 157-166.

38 Tim Sherratt/Andrew Jackson, GLAM-Workbench/web-archives (Version 0.1.1), Zenodo - CERN 2020, doi:10.5281/zenodo.3894079.

39 https://mementoweb.org/guide/quick-intro/ [last accessed: April 2, 2021].

web archiving field as a whole is able to advance, to the benefit of archivists and researchers internationally.

Conclusion

In this chapter, we have tried to address the challenges of making data available to researchers from multiple disciplines who have different levels of exposure to web archives and their complex, multi-layered contexts. Some of the individual tools and methods described above have already been trialled within the UKGWA, while others remain ideas on a whiteboard. Whether its individual components have been realized or not, our prototype dashboard is the result of a process of collaboration and co-design. The stakeholders in web archiving and web archive studies have varied disciplinary interests, work in different sectors (with different cultures and imperatives) and bring different knowledge and expertise. An open exchange of knowledge, leading to the establishment of a common language and shared assumptions, will help to engender trust in web archives, to conceive of tools that are both feasible to develop and of immediate use to researchers, to embed archival expertise in new modes of access, and to plan services that will be sustainable and extensible in the long term. We hope that the way of working we have outlined, and the prototype dashboard that we have begun to specify here, will be the beginning rather than the end of a conversation.

Bibliography

BARONI, Marco, et al., Cleaneval: a Competition for Cleaning Web Pages, in: Proceedings of the Sixth International Conference on Language Resources and Evaluation 2008, 1.

BELL, Mark, From Tree to Network: Reordering an Archival Catalogue, in: Records Management Journal 30 (3/2020), 379-394, doi:10.1108/RMJ-09-2019-0051.

BLEI, David M./NG, Andrew Y./JORDAN, Michael I., Latent Dirichlet Allocation, in: Journal of Machine Learning Research (3/2003), 933-1022.

BRÜGGER, Niels, Website History and the Website as an Object of Study, in: New Media & Society 11 (1-2/2019), 115-132, doi:10.1177/1461444808099574.

COWLS, Josh, Cultures of the UK Web, in: Niels Brügger/Ralph Schroeder (eds.), The Web as History: Using Web Archives to Understand the Past and Present, London 2017, 220-237.

DE ROSNAY, Joël, The Macroscope: A New World Scientific System, New York 1979.

GRAHAM, Shawn/MILLIGAN, Ian/WEINGART, Scott, Exploring Big Historical Data: The Historian's Macroscope, London 2015.

JETT, Jacob, et al., The HathiTrust Research Center Workset Ontology: A Descriptive Framework for Non-consumptive ResearchCollections, in: Journal of Open Humanities Data 2 (2016).

MIKOLOV, Tomas, et al., Efficient Estimation of Word Representations in Vector Space, arXiv:1301.3781 (2013), URL: https://arxiv.org/abs/1301.3781v3 [last accessed: April 2, 2021].

MILLIGAN, Ian, Web Archive Legal Deposit: A Double-edged Sword, 14.07.2015, URL: https://ianmilli.wordpress.com/2015/07/14/web-archive-legal-deposit-a-double-edged-sword/[last accessed: April 2, 2021].

RUEST, Nick, et al., The Archives Unleashed Project: Technology, Process, and Community to Improve Scholarly Access to Web Archives, in: Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020 (JCDL '20), New York 2020, 157-166.

SCHNEIDER, Steven M./FOOT, Kirsten A., The Web as an Object of Study, in: New Media & Society 6 (1/2004), 114-122, doi:10.1177%2F1461444804039912.

SCHNEIDER, Steven M./FOOT, Kirsten A., Web Sphere Analysis: An Approach to Studying Online Action, in: Christine Hine (ed.), Virtual Methods: Issues in Social Science Research on the Internet, Oxford 2005, 157-170.

SHERRATT, Tim/JACKSON, Andrew, GLAM-Workbench/web-archives (Version 0.1.1), Zenodo - CERN 2020, doi:10.5281/zenodo.3894079.

WINTERS, Jane, Giving with One Click, Taking with the Other: Electronic Legal Deposit, Web Archives and Researcher Access, in: Melissa Terras/Paul Gooding (eds.), Electronic Legal Deposit: Shaping the Library Collections of the Future, London 2020, 159-178.

Chapter 3: Design Thinking, UX and Born-digital Archives: Solving the Problem of Dark Archives Closed to Users[1]

Lise Jaillant, Loughborough University, UK

Abstract

As a human-centered method to solve business and social problems, design thinking has been applied to "wicked problems" in a wide range of sectors. However, the archival sector has rarely engaged with this methodology. This chapter argues that design thinking is a productive way to solve the problems of access and use of archival collections in the digital age. Indeed, the vast majority of born-digital archives are not available to users due to data protection, copyright and other issues. Drawing on the author's experience as a researcher who has had access to "dark" archives normally closed to the public, the chapter presents examples of research that can be done using born-digital records. It demonstrates the importance of seeking early feedback from researchers via design thinking workshops, and of designing and improving access procedures through an iterative process. Researchers have too often played the role of passive users of archival collections. They now need to work closely with archivists to shape access policies that will facilitate the use of innovative methodologies such as Artificial Intelligence.

Introduction

Popularized by Tim Brown and his IDEO team at Stanford, design thinking is a human-centered method to solve business and social problems. This creative problem-solving approach reaches beyond professionally trained designers. Interdisciplinary teams can use design thinking to tackle even society's most intractable "wicked problems," argues Brown in the updated 2019 edition of his bestseller *Change by Design*.[2] A key message is for organizations to focus on the people they

1 This research was funded by the Arts and Humanities Research Council (grant ref AH/R00773X/1). See www.poetrysurvival.com for more information about this project.

2 Tim Brown, *Change by Design: How Design Thinking Transforms Organizations and Inspires Innovation*, New York 2019, see 250-256.

are serving. The first question should always be: what is the human need behind the problem we are trying to solve?

Design thinking has been extremely influential in the business world, leading to the multiplication of "Innovation Labs" and "Experience Centers." Yet, it is not a term that Digital Archivists and Digital Humanists frequently use. Although there is a growing body of work linking design thinking to the development of services within libraries, museums and exhibitions spaces, the archive sector has only recently started paying more attention to this field.[3] "User-centered design thinking as a driver for innovation" was the theme of a panel at the 2019 "Designing the Archive" conference in Adelaide (Australia).[4] As part of this panel, the National Archives of Norway presented their work on digital records, and the need to maintain reliability and trustworthiness of these records at all stages of the process (from creation to transfer to the archival institution). In the past few years, they have applied design thinking to develop creative solutions to the issues of digital archives. In June 2019, the Norwegian National Archives also presented this work to the European Archives Group, which is part of the European Commission expert groups. Following this meeting, "Archiving by design" has been identified as a priority for EU-wide collaboration.

While this work claims a user-centered approach, it is led by archivists in partnership with record creators. Too often, end-users[5] do not fully participate in debates on archives, including born-digital archives.[6] "Citizens" remain abstract fig-

3 For examples of design thinking applied to museums and other cultural organisations, see: Lucy Larson, Engaging Families in the Galleries Using Design Thinking, in: *Journal of Museum Education* 42 (4/2017), 376-384, doi:10.1080/10598650.2017.1379294; Suzanne MacLeod et al., New Museum Design Cultures: Harnessing the Potential of Design and "Design Thinking" in Museums, in: *Museum Management and Curatorship* 30 (4/2015), 314-341, doi:10.1080/09647775.2015.1042513; and Mahendra Mahey, *Open a GLAM Lab. Digital Cultural Heritage Innovation Labs*, Doha, Qatar, 2019.

4 See https://www.archivists.org.au/conference/program-and-abstracts/abstracts [last accessed: Mar. 31, 2021].

5 In this essay, users are defined as end-users, i.e. academic researchers and members of the public who are using archival collections for professional or personal research. The point of this distinction is to clarify the archival circuit, from creators of records and archivists, to end users. For more on users of archives, see Elizabeth Yakel/Deborah A. Torres, AI: Archival Intelligence and User Expertise, in: *The American Archivist* 66 (1/2003), 51-78, http://www.jstor.org/stable/40294217 [last accessed: April 5, 2021].

6 In 2010, OCLC listed various kinds of born-digital resources, broadly defined as "items created and managed in digital form" (Ricky Erway, Defining "Born Digital," Dublin, OH, 2010, URL: https://www.oclc.org/content/dam/research/activities/hiddencollections/borndigital.pdf [last accessed: April 5, 2021], see 1). This includes digital photographs; digital documents such as PDFs; harvested web content; the digital manuscripts of noteworthy individuals; the electronic records of institutions; static data sets generated by researchers; dynamic data such as Facebook and Twitter accounts; digital art; and digital media publications – music

ures that have a right to access in theory, but rarely do in practice. This chapter draws on my experience as a researcher who has had access to "dark" archives (defined as archives normally closed to the public).[7] In particular, it presents the work done during my project funded by the Arts and Humanities Research Council (2018-2020), focusing on the Carcanet Press archive at the John Rylands Library in Manchester. As a Humanities scholar, my main methodology has always been archival work. In this essay, however, I also use autoethnography – reflecting on my own experience and connecting this experience with a wider context.[8]

My central argument is that design thinking is a productive way to solve the problems of *access* and *use* of archival collections in the digital age. A key message for archivists will be to seek early feedback from researchers via design thinking workshops, and to design and improve access procedures through an iterative process (Fig 3.1). Researchers have too often played the role of passive users of archival collections.[9] They now need to work closely with archivists to shape access policies that will facilitate the use of innovative methodologies such as Artificial Intelligence.

---

and movies, for example. This chapter focuses mostly on materials immediately relevant to historians, literary scholars, and other digital humanists – including web archives, authors' personal archives, and government records.

7 Although this chapter focuses on born-digital archives, it relies on the extensive literature on confidentiality and access to archives, which applies both to physical and digital archives. For example, authors' physical archives are also subject to highly restrictive access frameworks. See, for instance, Mark Greene/Dennis Meissner, More Product, Less Process: Revamping Traditional Archival Processing, in: *The American Archivist* 68 (2/2005), 208-263, doi:10.17723/aarc.68.2.c741823776k65863; Ben Goldman/Timothy D. Pyatt, Security Without Obscurity: Managing Personally Identifiable Information in Born-Digital Archives, in: *Library & Archival Security* 26 (1–2/2013), 37–55, doi:10.1080/01960075.2014.913966; Valerie Harris/Kathryn Stine, Politically Charged Records: A Case Study with Recommendations for Providing Access to a Challenging Collection, in: *The American Archivist* 74 (2/2011), 633–51, doi:10.17723/aarc.74.2.f252r28174251525; Julia Kastenhofer/Shadrack Katuu, Declassification: A Clouded Environment, in: *Archives and Records* 37 (2/2016), 198–224, doi:10.1080/23257962.2016.1194814.

8 According to Carolyn Ellis, autoethnography can be defined as "research, writing, story, and method that connect the autobiographical and personal to the cultural, social, and political." Carolyn Ellis, *The Ethnographic I: A Methodological Novel about Autoethnography*, Walnut Creek, CA, 2004, see xix.

9 See Lise Jaillant, After the Digital Revolution: Working with Emails and Born-Digital Records in Literary and Publishers' Archives, in: *Archives and Manuscripts* 47 (3/2019), 285-304, doi:10.1080/01576895.2019.1640555.

Fig 3.1: Involving end-users of born-digital archives. Courtesy of the author.

The first section examines design thinking as a methodology that can be applied to archival collections. I then turn to my own experience of gaining access and using "dark" archives, focusing on the Carcanet Press collection. The third section offers practical ways of involving users more closely, moving from user-testing to design thinking in action. Finally, the conclusion articulates what research users' priorities are when it comes to dark archive accessibility, and how design thinking helps design solutions to meet these.

1. Design Thinking as a Methodology

In January 2017, the UK Cabinet Office, working in collaboration with The National Archives (TNA), identified key issues with the management of born-digital records within government, resulting from the implementation of digital technologies and the decentralization of record-keeping processes. It has created an environment where data and records are scattered across multiple platforms and not organized in a way to facilitate timely retrieval.[10] The digital revolution has challenged established processes: "Much of what has accumulated over the past fifteen to twenty years is poorly organized, scattered across different systems and almost impossible to search effectively."[11] This inefficient organization of records is a pressing issue as government records now need to be transferred to TNA after twenty years, rather than the previous thirty-year rule.

10 Cabinet Office (UK), *Better Information for Better Government*, 18.01.2017, https://www.gov.uk/government/publications/better-information-for-better-government [last accessed: Mar. 30, 2021], see 3.

11 Cabinet Office (UK), *Better Information for Better Government*, 3.

The report made several suggestions to improve the management of digital records. First, departments with large – and potentially very sensitive – unstructured data were encouraged to use e-discovery tools (which are often used by legal professionals in their investigations). In 2015, TNA carried out a study on the applicability of e-discovery for appraisal, selection and sensitivity to determine the strengths and weaknesses of these technologies and published their findings in 2016.[12] A second recommendation from the Better Information for Better Government report encouraged civil servants to take care of their records by making record management "Easy, Attractive, Social and Timely." The third recommendation in the report was to include information management compliance as part of performance review, as an additional incentive to make progress. This can also be a lever to change perceptions on the usefulness of record-keeping. These measures would "unlock the value of legacy collections for the benefit of civil servants and citizens alike."[13]

While citizens are mentioned a few times in the report, there are no details about specific ways to engage with these end users. How will government born-digital records be made available? Will users have to travel to The National Archives to access these records? Is it feasible and desirable to make born-digital archives available to anyone with an internet connection? TNA are doing important work on the issue of access, in particular via initiatives led by their Head of Digital Access.[14]

The problems of access and use of archival collections in the digital age are multi-faceted and complex. Although there is no quick fix, we can make progress by turning to design thinking – a set of methods and skills to solve challenging problems. It is not a new concept. John E. Arnold, a professor of mechanical engineering and business administration at Stanford, was one of the first to use the term "design thinking." Arnold was convinced that innovators should start with human needs, rather than develop a technical product first and then see if it is of interest. In his lectures, compiled under the title *Creative Engineering* (1959), Arnold described the "creative engineer" as a professional who combined technical skills with a human-centered approach more comprehensive than industrial design.[15]

---

12 https://www.nationalarchives.gov.uk/documents/technology-assisted-review-to-born-digital-records-transfer.pdf[last accessed: April 5, 2021].

13 Cabinet Office (UK), *Better Information for Better Government*, 9.

14 TNA's Head of Digital Access, Catherine Elliott, has been doing a lot of work on revamping the presentation of the catalogue and access to born digital information. See the presentation she gave at the ICA 2019 Adelaide conference: https://www.ica.org/sites/default/files/session_lt1_a-project-alpha_by_catherine_elliott_0.pdf and https://www.youtube.com/watch?v=bsiWfnUGLZk&feature=emb_logo [last accessed: April 5, 2021].

15 John E. Arnold, *Creative Engineering*, https://stacks.stanford.edu/file/druid:jb100vs5745/Creative%20Engineering%20-%20John%20E.%20Arnold.pdf [last accessed: Mar. 30, 2021].

Over the next fifty years, the term "design thinking" broadened its reach well beyond engineering and its applications in the business world.

"Design thinking" did not emerge out of nowhere. Systems thinking – theorized and practiced by Russell Ackoff, C. West Churchman, Peter Checkland and others – developed ideas that are also found in design thinking. This includes engaging with a broad range of stakeholders, considering and exploring idealized options, reframing problems, using iteration, diagrams and pictures, and tirelessly searching for better alternatives. Design ideas were then applied to various organizations and to society as a whole. Likewise, in *The Design Way: Intentional Change in an Unpredictable World*, Harold G. Nelson and Erik Stolterman moved beyond design theory and practice, to formulate design culture's fundamental core of ideas. The "design way" was applicable not only to traditional fields such as architecture and graphic design, but also to other organizations – including education and health care.

Previously confined to professional and academic circles, the concept of "design thinking" became mainstream due largely to the work of IDEO, a California-based design firm that specializes in innovation and strategy. In *Change by Design* (2009, revised edition 2019), Tim Brown offered an easy-to-read guide to design thinking, presented as a way to transform organizations and inspire innovation. He described the design thinking process as a system of three overlapping areas: *inspiration*, *ideation*, and *implementation*.[16] In the case of institutions that do not engage directly with the market, this third phase is replaced with *iteration*, i.e. continual experimentation based on user feedback.

Let's start with *inspiration*, which is about framing a challenge and discovering new perspectives. As described in the 2014 IDEO toolkit *Design Thinking for Libraries*,[17] the response to a challenge should be: "How do I approach it?" The mindset should be optimistic and forward-looking. Problems need to be re-framed as opportunities in disguise. To improve their understanding of the challenge, design thinkers need to gather insights through observations and interviews of target users and experts. They need to emphasize with the audience and learn more about their beliefs and behaviors.

The *ideation* phase transforms this research into actionable insights. It aims to produce a wide range of ideas (divergent thinking) before selecting the best ones (convergent thinking). "To have a good idea, you must first have lots of ideas," said the Nobel Prize winner Linus Pauling.[18] To visualize and evaluate ideas, design thinkers create prototypes. These sketches and models make ideas tangible. The objective is not to produce something perfect – prototypes are "quick and dirty." They allow designers to share their ideas with others and to obtain rapid feedback.

16 Brown, *Change by Design*, 22.

17 Available at: http://designthinkingforlibraries.com/ [last accessed: April 1, 2021].

18 Cited in Brown, *Change by Design*, 73.

After prototyping, the next phase is to test the model with users and refine it. During this *iteration* phase, design thinkers continue to build on the original prototype thanks to feedback. They go through multiple rounds of iteration of their concept before they are ready to launch their new ideas in the world. Design thinkers seek a perfect balance of *desirability*, *feasibility*, and *viability*. A new idea should make sense to people and for people, it should be possible within the foreseeable future, and it should take part in a sustainable business model.

The popularity of design thinking in the business world has led to attacks – including from a young sociologist called Tim Seitz. Drawing heavily on Luc Boltanski and Eve Chiapello's *The New Spirit of Capitalism*, Seitz sees criticism as a catalyst for changes in the spirit of capitalism.[19] In the 1990s onwards, the rise of a digitally dominated world with fewer human contacts could have threatened capitalism, accused of lacking authenticity. Instead, capitalism has harnessed these criticisms, using design thinking and its focus on humans and empathy as a catch-all solution to all "wicked problems."

While I find the broad claims of design thinking gurus problematic, I believe that we should not throw the baby out with the bath water. Design thinking is not a magic solution to all our problems, but it can be a useful method – particularly when various groups do not communicate well. In the case of born-digital records produced by government agencies, end-users are held at a distance. This lack of communication has an impact on producers of records, who lose touch with the finality of information management compliance. Why spend time on administering records? With closer interactions with end-users, producers of records could see that filing digital records (either manually or automatically) is not a waste of time, but a service to their fellow citizens and to future generations. Design thinking is a way to bring people together and start important conversations.

How can the three phases of the design thinking process (inspiration, ideation, iteration) be applied to libraries and archives? This is a disparate group, and the 2014 IDEO report focuses mostly on public libraries with a general audience of readers – for example elderly people who don't know how to use computers, or parents with their babies.[20] These groups have specific needs – which can of course be very different from the needs of people who engage with national archives or Special Collections libraries.

Design thinking differs from user-testing of interfaces. In a good design thinking workshop, library and archive professionals listen to users and observe their behavior so they can empathize with their needs and identify the problems that

19 Luc Boltanski and Eve Chiapello, *The New Spirit of Capitalism*, London 2007; Tim Seitz, *Design Thinking and the New Spirit of Capitalism: Sociological Reflections on Innovation Culture*, London 2020, doi:10.1007/978-3-030-31715-7.

20 Available at: http://designthinkingforlibraries.com/ [last accessed: Mar. 31, 2021].

require solving. Whereas user-testing of interfaces attempts to solve predefined problems, design thinking uncovers real problems that need solving. Design thinking is also different from iterative methods such as Agile. The Agile Manifesto, released in 2001, outlined a way for project managers to make software design more responsive, and less burdened by paperwork and predefined specifications.[21] As a project management method, Agile relies on gathering fast feedback, producing iterative releases, and rapidly adapting the design plan to best meet the needs of the users. While Agile is a method used to build better software, design thinking can be used by anyone to solve any "wicked problem" that has no clear solution. In the case of access frameworks in archives, the main objective is not to test a software or an interface. It is to understand the needs of the users and respond creatively to the problems they face.

In their 2019 conference presentation, the National Archives of Norway explained how they used design thinking to re-envision the concepts of records and archive in the digital world.[22] With the rise of digital records, the challenge was to modernize inefficient processes born in the paper age. In March 2017, a report to the Norwegian Parliament gave a bleak picture of the state of digital records in government. With around half a million documents produced a year, the problem is not the lack of records, but the poor management of these records. Buried under a flood of digital materials, the public sector finds it difficult to actively manage its archives.

The Norwegian National Archives identified a "toxic cocktail" of four main factors. First, the volume and range of data, which comes in many formats and through various channels, make the preservation and management of digital data complicated. Second, nobody is expected to spend any time on this, and government officials often prioritize their daily job. Third, potentially important records are reviewed individually, which can be inefficient and time consuming. And fourth, the complexity and low level of integration of IT systems also contribute to the problem. As a consequence of these inefficient practices, no or poor records exist for important areas such as Norwegian bilateral relations with China 2010 – 2017 (after the Nobel peace prize to Chinese political opponent Liu Xiabao); the decision of Norwegian leaders not to meet the Dalai Lama in 2014; and the investigation of international law as a basis for Norway to take part in the action against ISIS.

To start addressing this complex problem, the National Archives of Norway adopted a design thinking approach. Teams were encouraged to challenge existing truths and ways of doing things. Their process was iterative, starting with the

21 https://agilemanifesto.org/[last accessed: April 5, 2021].

22 Espen Sjøvoll, *Design Thinking & Innovation: Norwegian Approach*, https://www.archivists.org.au/documents/item/1640 [last accessed: Mar. 30, 2021].

learning process: What have we learned? What do we need to learn now? The next step was to share this knowledge, adjusting insights along the way. This collective discovery process led to a prototype that allowed people to test the new idea. The learning process could then start again, building on the previous prototype to improve it.

This approach yielded significant results, including new approaches to appraisal and digitization. The National Archives of Norway now digitize 90% of modern paper archives, reducing long-term costs. The number of records transferred to the National Archives also increased – with fewer staff members needed to transfer these records. By 2025, the vision of the Norwegian archives is that government employees will no longer need to spend time on archiving records. While this approach claims to be user-centered, the focus is mostly on record creators (i.e. civil servants) rather than end-users of digital archives.

The activities of the National Archives of Norway influenced the creation of the EU "Archives by Design" working group in 2019. Making born-digital records more accessible is a key priority. Indeed, the ambition of the group is to "get new insights on opportunities and experiences with early intervention with ICT [Information and Communications Technology] systems development to achieve sustainable accessibility of data, persistent/authoritative data and ensuring data protection across European countries – also known as Archiving by design."[23] For Erik Saaman, Strategic Advisor at the Dutch National Archives, "archiving by design means designing information systems to support the work process in such a way that the long-term accessibility of that information is taken into account from the outset."[24] Despite the proclaimed focus on accessibility, this goal remains a largely abstract principle. We need to move away from accessibility as an abstract principle, towards actual end-end-users who want (and often fail) to access digital records (Fig 3.2).

23 *Draft Mandate: Archives by Design Working Group*, https://ec.europa.eu/transparency/regexpert/index.cfm?do=groupDetail.groupMeetingDoc&docid=34956 [last accessed: Mar. 30, 2021]. My emphasis.

24 https://www.nationaalarchief.nl/en/archive/knowledge-base/archiving-by-design [last accessed: Mar. 31, 2021].

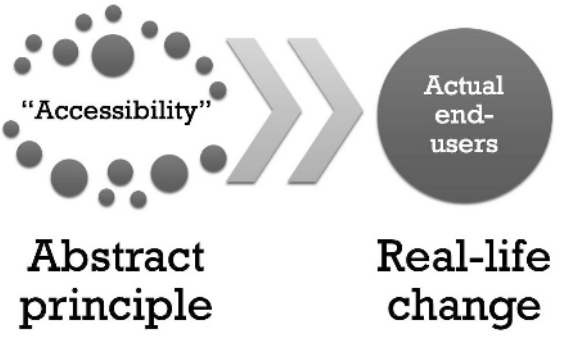


Fig 3.2: Moving away from "accessibility" as an abstract principle. Courtesy of the author.

2. Gaining access and using "dark" archives

I will now turn to my own experience as a publishing historian who has gained access to "dark" digital archives normally closed to the public. Publishers often treat their archives as rubbish. Fortunately, we still have a large number of records from firms such as Random House or Chatto & Windus due to two main reasons: the determination of enterprising archivists to preserve materials they thought valuable; and the incentivization of publishers who were promised prestige and financial rewards if they transferred their collections to university libraries. The pressure to preserve publishers' records rarely came from scholars and other users. As a publishing historian, I stand on the shoulders of giants in my field: but these great scholars did little to gather the collections that made possible their own scholarship. With archivists in the driving seat, the question of access was often relegated to the "desirable" rather than "essential" criteria. At the University of Reading, for example, users need to ask Random House UK for permissions to consult archival documents, which severely restricts access. At the John Rylands Library in Manchester, large sections of the archive of Carcanet, a leading poetry publisher in the UK, are closed to the public.

The relationship between access issues and the underlying copyright, confidentiality, and privacy requirements of archives is a complicated one. In addition

to copyright restrictions,[25] there are at least two issues at the core of the "dark" archive situation: first, the lack of a technical infrastructure to make born-digital records available; and second, issues relating to the confidentiality or sensitivity of these documents. There are also collecting policy questions involved here, not just access and interface issues. Could researchers work with archivists and authors to define and clarify the expectations when born-digital archives are being acquired? Do researchers have a right to expect these kinds of archives to be as open as possible? What is the point of acquiring these archives if copyright and confidentiality severely restrict access or close them entirely? Even if we put aside these collecting policy questions and focus on archives already acquired, we still need more collaboration between archivists and researchers. After several years of discussions and collaborative work with archivists, I am convinced that we should move fast (and avoid breaking things). Open data respectful of privacy is possible, and the first step is to quickly build prototypes to give access to archival records.

Since the late 1970s, the John Rylands Library in Manchester has acquired the Carcanet Press archive on a yearly basis. Founded in 1969 by Michael Schmidt and Peter Jones, Carcanet moved from Oxford to Manchester in 1972. The press went on to build a diverse list, including poetry in translation and by neglected women poets. Among the distinguished writers associated with Carcanet are Elizabeth Jennings, Ted Hughes and many others. In the past three decades, the Carcanet Press archive has become hybrid: it is now composed of paper records but also emails and other born-digital documents. The vast majority of the paper archive is uncatalogued and closed to researchers; and the digital part of the collection is a "dark" archive, open only to a handful of staff.

From 2012 to 2014, Fran Baker led the Carcanet Press E-mail Preservation Project at the John Rylands Library.[26] The project resulted in the successful rescue and preservation of 215,000 e-mails and 65,000 attachments generated by Carcanet Press, as well as comprehensive metadata. Towards the end of the project, Fran Baker experimented with network graphs to analyze this data. She used metadata such as the person who sends the email; the person who receives the message; the time and date of the message. The first graph represents the correspondence of Michael Schmidt (the founder of Carcanet Press) "with two writers," but we do not know who these writers are. The second graph represents "Schmidt's network of correspondents."[27] But again, we do not know the names

25 For example, copyright is at the centre of the requirement that mandates on-premises access, e.g. web archives at the British Library and other institutions.

26 Fran Baker, E-Mails to an Editor: Safeguarding the Literary Correspondence of the Twenty-First Century at The University of Manchester Library, in: *New Review of Academic Librarianship* 21 (2/2015), 216-224, doi:10.1080/13614533.2015.1040925.

27 Baker, E-Mails to an Editor, 222.

of these correspondents. There is something else we do not know: the corpus that Fran Baker used to create these graphs. Did she use the entire Carcanet Press archive? Or a selection of it? There are a lot of missing gaps in this story, which shows that even metadata can be seen as potentially confidential.

What happens if people want access to the full text of the emails rather than just metadata? The archive is normally closed to the public. However, my AHRC grant allowed me to employ a Project Archivist for a couple of months in 2019. She was based at the Rylands Library and had access to the entire collection. In Summer 2019, she prepared a selection of 200 emails that she thought would be interesting for me to see. She then submitted the selection to Michael Schmidt, the founder of Carcanet Press, for approval. Schmidt requested that some materials be closed or redacted for confidentiality reasons. The redacted selection of emails was then sent to me as a PDF, with email attachments in a separate ZIP folder.

For archivists, only basic technical skills are necessary to provide access to emails and other born-digital archives. There is no need to build a complicated system, or to buy expensive tools. Creating a PDF is enough to allow users to see content that will be useful for their research. This is of course not perfect: as a researcher, I wish I could download thousands of emails and do some data analysis. But even a small selection of data is better than no data at all. Issues with technical infrastructure, at the core of the "dark" archive problem can be easily resolved if archivists and researchers embrace imperfection. As we have seen, a prototype inspired by design thinking is "quick and dirty" and can be improved over time, whereas a closed archive remains static and inaccessible.

Not everyone will agree that faster is always better for users. Archivists often look at the longer perspective, and will happily sacrifice instant gratification (access now) for a more sustainable and considered enjoyment in future (preservation for access by later generations). Since the 1980s, the slow movement has been advocating a cultural change toward slowing down life's pace, and favouring quality over quantity in everything from work, food and parenting, to archival processing and research.[28] For advocates of this movement, it is important to do everything at the right speed. Yet, a key issue is to define what kind of speed is "right." Even the participants of the Slow philosophy recognise that doing everything at a snail's pace is not beneficial. Accelerating the opening up of dark archives would benefit

28 See for example, Carl Honoré, *In Praise of Slow: How a Worldwide Movement Is Challenging the Cult of Speed*, London 2010; Maggie Berg/Barbara K. Seeber, *The Slow Professor: Challenging the Culture of Speed in the Academy*, Toronto 2018; Kimberly Christen/Jane Anderson, Toward Slow Archives, in: *Archival Science* 19 (2/2019), 87–116, doi:10.1007/s10502-019-09307-x.

the research community, and reinforce the legitimacy of archival collections at a time when libraries and cultural organisations are under attack.[29]

The second issue (the confidentiality and sensitivity of some born-digital documents) can also be addressed with a change of mindset. This applies to archivists, but also to researchers, who need to critically evaluate their own expectations too. Archivists were understandably nervous when they gave me access to the selected Carcanet emails. Even after redaction, emails often contain information that the sender had not intended for public release. This is particularly problematic in light of the GDPR (General Data Protection Regulation) that applies to UK and European collections. But it is essential to embrace risk and trust that researchers will make good use of the data they access. And for users, it is important to respect privacy. Each time I saw CLOSED or [.....REDACTED.....] on the PDF, I wished I could see the entire message. It reminded me of the asterisks used for censored passages in early-twentieth-century books. Yet, I also realize that I would not want people to access all of my emails. Closure and redaction are reasonable measures, as long as the user is informed of the withdrawal of information.

Many libraries and archival collections are now experimenting with new systems to make their digital collections more accessible. For example, ePADD (an open-source software developed at Stanford University) is a valuable tool to discover born-digital materials, but researchers still need to travel to Special Collections to consult relevant records. For archival repositories with limited staff time and funding, one solution is to create PDFs based on certain themes and to make them available to users after obtaining permissions. This is a low-tech solution that nearly all institutions could implement rapidly to respond to user needs. "Our users are crying out for faster access," argued Mark Greene and Dennis Meissner in their influential article "More Product, Less Process."[30] To resolve the "dark" archive problem, archivists need to start with the users and quickly work backwards. But unlocking born-digital data is not a one-way process. We need more collaboration between archivists and users. We also need more empathy: the ability to understand the concerns of archivists, and the needs of users of born-digital collections.

Design thinking is at its core a user-centered approach. Gaining access to "dark" archives is one thing, producing new knowledge is another. The selection of 200 emails that the Project Archivist prepared were generated by Carcanet during a single year (2010). Why did I choose this particular year? In 2010, the Arts sector faced major cuts following the economic crisis. For Michael Schmidt, this period brought anxiety, but also new opportunities. In October 2010, the report "Mapping Contemporary Poetry" offered an overview of the poetry sector from the

29 See Richard Ovenden, *Burning the Books: A History of Knowledge under Attack*, Cambridge, MA, 2020.

30 Mark Greene/Dennis Meissner, More Product, Less Process, 235.

perspective of poetry publishers (including Carcanet and Bloodaxe). In the past, there had been many tensions between Michael Schmidt of Carcanet and Neil Astley of Bloodaxe. But the two men had decided to put aside their differences and work together. I was therefore interested in this key moment: the moment when poetry publishers united in order to survive a tough funding landscape.

The Project Archivist divided the corpus of emails into 11 sections, focusing on the Publishing Landscape; on Arts Council England; on poetry publishers (Peepal Tree; Bloodaxe); on women poets and editors (such Helen Tookey; Alison Brackenbury; Sujata Bhatt; Mimi Khalvati; Elaine Feinstein); and on Poetry in translation. It was sometimes very difficult to understand the context. For example, this is what EMAIL#20 looks like: *"[…….REDACTED……..] I agree with Andrew. But I am glad the Seraglio is rewarding. I hope you wear your curly-toed slippers and as many of your veils as possible. I will be at the Book Mess on Wednesday wouldn'tyaknowit so I will miss you. We will palaver soon I hope… All best! T B."*

When I read the other emails, I understood that "TB" refers to "Teddy Bear." EMAIL#25 was sent by the editor Helen Tookey to Michael Schmidt. She wrote: "*you will be intrigued to hear that you were starring the other day as one of the characters in a strange wedding ceremony that Patrick and Rowan created in our living room between a large teddy bear and a pink panther, named for the occasion Michael Schmidt and Julie Vanberger respectively. (I have no idea where ' Julie Vanberger' came from.)*"

Now, if you go back to EMAIL#24, you will find Michael Schmidt's response to Tookey: "Very pleased to be a large teddy bear! If only I'd married a pink panther!" At this point, I realized that "Teddy Bear" or "T.B." was actually Michael Schmidt. Remember that this email correspondence is between Tookey (an editor and published poet) and Schmidt (a publisher). It tells us a lot about the tone of emails – which is of course much more informal than business letters.

What kind of methodologies did I use to analyze my corpus of Carcanet emails? The first methodology is of course close reading – because I had a limited number of emails, I could read everything and make notes on the main topics. Unsurprisingly, a key topic is the changing funding landscape in 2010. Just after the announcement for a new program on 1,000 poets working in school, Simon Thirsk of Bloodaxe wrote to Jeremy Poynting and Michael Schmidt:

> "Were you surprised by the amount of money going to writer development organisations, Apples and Snakes, Etc? I think it's important for us to realise that **Literary Merit** is only one way of viewing the world. There are several other. Some philistines, bigots and Tories take a purely **Financial Merit** view: there is only merit in what sells, (I just read bestsellers, like my friends.) Others take a **Social Merit** view: the importance of poetry in education (as a useful way to test intelligence but not important in itself), in community work (self-expression, releasing neurotic thoughts, recollection)."

Here, Simon Thirsk identifies three ways to define the value of poetry: it could be based on Literary Merit, Financial Merit or Social Merit. Among these publishers, there is the impression that funding is now allocated on social merit rather than on literary merit. If you do social good, you get the funding, but if you only publish old white males, you will not get the funding, even if they are very gifted. This email is very detailed, very elaborate and is closer to a letter, and offers precious information for researchers.

In addition to close reading, there are many other things we can do with email records. First, I learned to use Gephi, using the tutorials on Martin Grandjean's website.[31] I prepared a spreadsheet with the list of correspondents, and their gender. I then prepared another spreadsheet with the relationships between the correspondents. Once you combine the two spreadsheets on Gephi, you can get this kind of visualizations (Fig 3.3).

Michael Schmidt is of course at the center of the network – and there are two interesting things here. First, his main correspondents are men, often older men in the publishing industry. You see a lot of green on the graph – I chose green for men, and pink for women. Among the women who are important correspondents, you see Judith Willson, who is an editor and also a published poet. Helen Tookey is not a major correspondent. Visualizations are a great way to see connections that were not obvious at first sight. But of course, it all depends on the data that you have. I had little control over the selection of Carcanet emails, and this selection of course had an impact on the visualizations.

31 http://www.martingrandjean.ch/[last accessed: Mar. 31, 2021].

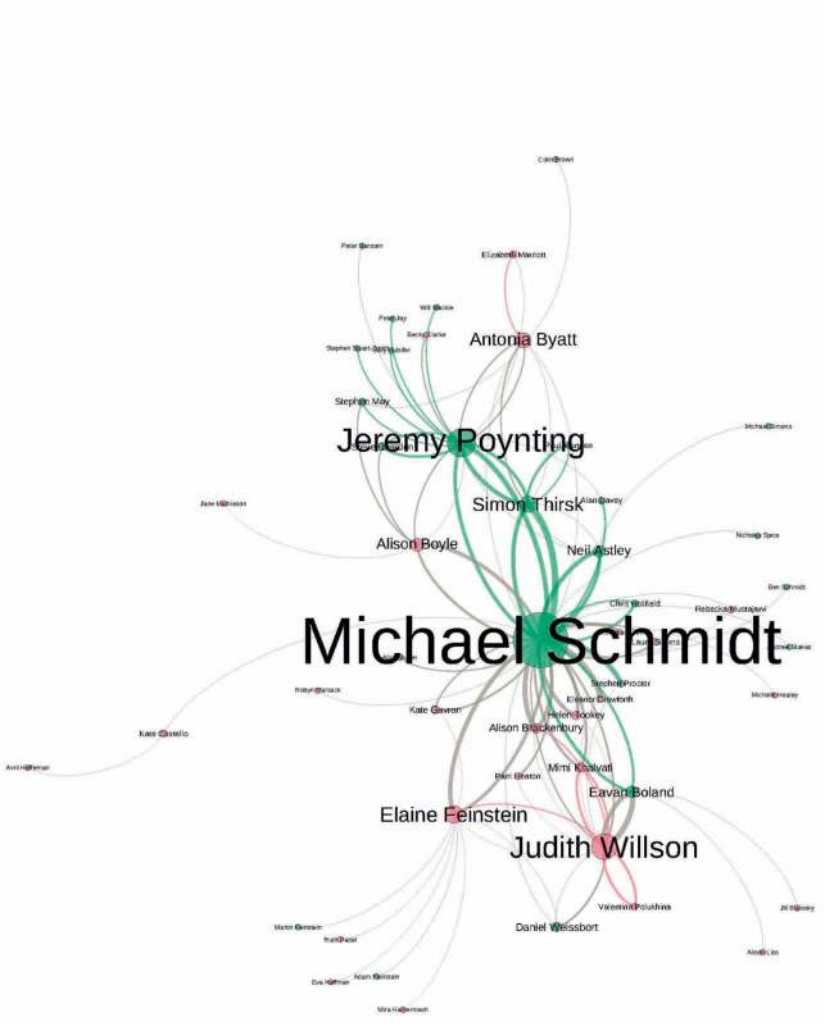


Fig 3.3: Michael Schmidt's network of correspondents (based on a selection of 200 emails exchanged in 2010). Courtesy of the author.

I also experimented with Voyant to analyze my corpus, and I discovered that frequent words include: "funding," "council" including Arts Council, and "love" (Fig 3.4).

Fig 3.4: Frequent words in Michael Schmidt's selection of emails. Courtesy of the author.

Emails that mention funding do not mention "love" and vice versa. Of course, "love" is a way to conclude a personal email. Michael Schmidt and his correspondents almost never use the term "funding" in their personal emails. The corpus I have is a mix of personal and professional emails. It shows that like most of us, Michael Schmidt uses email for all aspects of his life. It would be very difficult to write the biography of a contemporary figure without access to their emails.

In the future, I would like to experiment with other methodologies, including sentiment analysis and artificial intelligence. Sentiment analysis has been used for Germaine Greer's archive.[32] Germaine Greer is of course one of the leading figures of second-wave feminism. She was born in Australia and has lived in Britain since 1964. Her archive was purchased by the University of Melbourne in 2013. It is a huge collection with more than 500-plus boxes documenting all aspects of her personal and professional life. The University of Melbourne decided to catalogue the collection at file level – which is unusual for a collection of this size. They wanted to manage risk and improve the discoverability of the collection. The entire archive has detailed file-level metadata and subject indexing. Researchers can use keywords to search the collection, but it is not the most effective method: it does not offer a sense of the archive as a whole. Keyword search tells us nothing about the trends and themes that characterize the archive. Sentiment analysis can be a very useful approach to find information with a strong affective content. Text is classified as positive or negative based on the strength of sentiment that it expresses. In the case of the Germaine Greer's archive, researchers used a tool called SentiStrength, which has an inbuilt lexicon.

Machine learning is increasingly used to identify sensitive materials in born-digital archives.[33] Removing sensitive information is the first step, and the next

32 Millicent Weber/Rachel Buchanan, Metadata as a Machine for Feeling in Germaine Greer's Archive, *Archives and Manuscripts* 47 (2/2019), 230-241, doi:10.1080/01576895.2019.1568266.

33 See the presentation slides at the "Archives, Access and AI" conference (London, January 2020): https://www.poetrysurvival.com/presentation-slides-archives-access-and-ai-conference/[last accessed: Mar. 31, 2021].

step is to give access to this data. Scholars can then use machine learning to gain more information about the context of a collection. Instead of using keyword search to discover materials, we could train machines to tell us more about the context of a collection. For example, we could learn about patterns in the correspondence of two writers, Ian McEwan[34] and Kazuo Ishiguro: perhaps their emails often mention science, or religion, or the university they went to: UEA. These themes could be recommended to scholars – on the model of the Amazon recommendation engine, "customers who bought also bought."

3. Involving users more closely: From user-testing to design thinking in action

For researchers, collaborations with archivists are absolutely key. So far, the priority has been to secure the materials and preserve them first, before liaising with users and researchers. We now need to move the focus towards access. There are ways to do that. For example, the Wellcome Collection in London organized a workshop on born-digital archives in 2017.[35] And in September 2019, the British Library brought together curators, PhD students and academics (including myself) to discuss and test born-digital and "hybrid" archives. The examples of the Wellcome Collection and the British Library offer a model for other organizations interested in involving users. Although more could be done to fully apply design thinking to the case to archival collections, these two institutions are moving in the right direction.

Since 2015, the Contemporary Archives and Manuscripts division of the British Library has worked mostly with the personal born-digital archives of writers and scientists, which present multiple challenges.[36] Most of these born-digital files are text based, and British Library curators acquire and process them via a six-stage workflow: acquisition; capture; extracted capture; metadata extraction; migration as PDF/As; and access.

How do these workflows apply to email in particular? For example, the British Library holds the email archive of the poet Wendy Cope. This collection of around 25,000 emails is not easy to process and make accessible for several reasons. First, email is a highly distributed service: several actors are involved to accomplish end-

34 Lise Jaillant, From Letters to Emails: Reading Ian McEwan's Correspondence, *TLS Online*, 21.11.2017, https://www.the-tls.co.uk/articles/public/ian-mcewans-emails-letters/ [last accessed: Mar. 30, 2021].

35 Victoria Sloyan, *Overview of a Born-Digital Archives Access Workshop Held at Wellcome Collection*, London 2018, doi:10.6084/m9.figshare.6087194.v1.

36 See Josh Schneider et al., Appraising, Processing, and Providing Access to Email in Contemporary Literary Archives, in: *Archives and Manuscripts* 47 (3/2019), 305-326, doi:10.1080/01576895.2019.1622138, which includes a section on the British Library.

to-end mail exchange. Second, it is impossible to verify the authenticity of senders without universally adopted forms of digital signature verification, such as DKIM (DomainKeys Identified Mail). Third, the logic of "deliverable units" that applies to other materials does not work well with email threads.

Processing the Wendy Cope email archive involved several steps. The emails came to the Library as a .PST file stored on a USB Flash Drive. This drive was captured forensically using FTK Imager. The PST file was then converted to an .mbox file using Aid4Mail. The last step was to load the .mbox file into ePADD, an application intended to guide email archives from ingestion to access. What might the future look like? For the British Library, ePADD (or a similar tool) could be used in combination with bespoke access to metadata and raw files following request and clearance. Another option is to use emulation, which allows users to use computers similar to the donors' machines and to interrogate email files.

The Library aims to offer full access to born-digital archives in its Manuscripts reading room, via computer terminals (not from laptops within the reading room). During the workshop's user-testing session, participants were able to use the terminals to view and interrogate born-digital content as a pilot project. Users then answered questions on several collections, starting with the born-digital files of the writer Ronald Harwood. Due to issues with processing some of the files, not all of them were available. Is it better to include these files in the directory so that users have the complete set of digital objects available to them? Or does this inclusion detract from the experience of using the catalogue? My own recommendation was to make a note of this in the catalogue, explaining that the Library has not been able to process these files to a satisfactory level.

The next step was to turn to the OHS (Oral History Society) archive. The born-digital material in this collection was extracted from thirteen 3.5-inch floppy disks and totals 20 MB of data over 118 digital objects. Much of this born-digital material relates to the OHS Regional Network but also includes items connected with the journal, *Oral History*. Digital content from the archive can be viewed in the Manuscripts reading room, in PDF and JPEG formats. During the workshop, participants were asked to find digital objects with the same name and file extension. Why do some objects with the same name appear in different locations in the catalogue? By looking at the "Size" and "Last Modified" information, users could see that files with the same name were not necessarily identical. Indeed, the dates when the born-digital objects were "Last Modified" by the creator have been specified in the "Scope and Content" field as this may differ from the creation date of the object.

Turning to the archive of the writer Hanif Kureishi, which arrived at the British Library in 2014, users were asked to compare electronic drafts of his work in the form of Word files. The process involved switching between tabs in the browser, which seemed time-consuming and inefficient considering the fact that Word has a "Compare documents" tool. To a large extent, it replicated the experience of con-

sulting paper drafts and comparing them manually – instead of harnessing the power of digital technologies. Embracing imperfection does not mean settling for less than commonly used tools offered by MS Word and other software.

Organizing a similar workshop does not involve a lot of time or resources, and it is an excellent way to gather feedback from users. Drawing on the British Library's questionnaire, the following five sets of questions could be a starting point for other institutions interested in user-testing sessions. First, users would be asked about their experience with descriptive metadata. Is it necessary to describe the digital object in detail? Or is it feasible for researchers to use only technical metadata? The second set of questions would focus on digital surrogates (for instance, PDF/A files): does it present difficulties regarding the authenticity of the information? Third, for hybrid collections, users could be asked if they need to consult physical papers alongside born-digital files (normally, only one collection at a time can be consulted). The fourth set of questions would relate to emulation as a model of access: is it better for users to access born-digital material through an emulated model? If so, what does emulation offer that other access models do not offer? Finally, users could be asked about their preferences for visualizations rather than data sets: what would be helpful for their research?

Comparing the British Library workshop with the user-testing session organized by the Wellcome Collection two years before brings interesting insights. A key finding of the Wellcome workshop was that "many researchers have limited experience of using born-digital archives."[37] Participants did not find the information easily, and they expected more guidance from archivists – particularly in terms of overview and context. This level of curation would of course be time-consuming and costly due to the huge volume of born-digital data.

At the British Library, participants (myself included) seemed equally lost. The experience of using a PC in the British Library reading room included many pain points. Although I have experience of using born-digital archives, I struggled to find the files listed on the questionnaire. The fact that I could not use my own laptop to consult the digital files made things worse. As I see it, having a detailed level of description would not necessarily make the user experience better (after all, researchers would still need to travel to the reading room and lose time trying to understand a complicated interface). What would make the user experience better would be to improve the interface giving access to files. Partnerships with UI/UX designers are urgently needed to remove pain points and lead to greater user satisfaction.

Following the 2017 workshop, the Wellcome Collection pointed out that users rarely know that some born-digital records are available, which of course results in low engagement with these materials. The Collection decided to change access

37 Sloyan, *Overview of a Born-Digital Archives Access Workshop*, 7.

conditions. While born-digital records were previously listed as unavailable to researchers, the catalogue now states that users can contact the Collections Information Team to request access to these records. Let's take the example of the record entitled "Birmingham Children's hospital/1996 Liverpool," a 3.5-inch floppy disk containing correspondence, designs and reports. Regarding access conditions, the catalogue now states that "digital records cannot be ordered or viewed online. Requests to view digital records onsite are considered on a case-by-case basis." This wording seems designed to discourage users: not only do they need to travel onsite to view materials, they also need to convince the Collections Information Team that they have a good reason to do this research. This is reminiscent of the policy of the Bibliothèque Nationale de France, where users of web archives are asked to first come to the library for an interview.[38] These obstacles risk discouraging many users, leaving only a minority of determined academics and journalists who have the confidence to push for access (and the funding to travel onsite). My suggestion is to change the wording of the catalogue to make clear that users are *encouraged* to request access, and that the library welcomes such requests since an archive is meant to be used – not locked away. Collaborations between archivists and academics who study conversation analysis would lead to more inclusive language on the catalogue and finding aids, broadening the range of users of born-digital records.

Like the Wellcome Collection, the British Library does not make it easy for users to access born-digital collections. The main problem is that born-digital records are not always listed on the catalogue and finding aids. In the case of the Will Self archive, the catalogue lists the collection as "Will Self: Personal and Literary Papers." Users who download the finding aid will find no mention of any born-digital records. The collection seems to be a traditional, paper-based archive. In fact, shortly after acquiring the collection, the British Library declared:

> Self's archive, like most of the contemporary archives we acquire, is a hybrid archive containing both paper and born-digital material. The collection includes his computer hard drive which holds a wealth of electronic manuscript drafts and approximately 100,000 emails along with a huge number of other files yet to be mined and identified (including downloads of his i-Tunes, which offer an intriguing line of investigation for future users of the archive).[39]

38 Sara Aubry, Introducing Web Archives as a New Library Service: The Experience of the National Library of France, in: *LIBER Quarterly* 20 (2/2010), 179-199, doi:10.18352/lq.7987.

39 Will Self's Archive Acquired by the British Library - *English and Drama Blog*, 21.12.2016, https://blogs.bl.uk/english-and-drama/2016/12/will-selfs-archive-acquired-by-the-british-library.html [last accessed: Mar. 30, 2021].

My suggestion is to be transparent and include details about born-digital collections holdings on the catalogue and finding aids. This would achieve two main purposes: first, inform users that these materials exist; and second, make the library accountable. Since "dark" archives cannot be dark forever, the library would need to provide an approximate date when these materials could be made available (Fig 3.5).

Fig 3.5: Three recommendations to improve access to born-digital records. Courtesy of the author.

Conclusion

The common purpose of these workshops was to place users at the center of the development of access to born-digital archives. As a researcher who has had access to archives that are normally closed to the public, I am convinced that this user-friendly approach is the way forward. Since 2017, I have led several international projects to bring together archivists and researchers (including Computer Scientists who specialize in AI).[40] A leitmotiv of the events organized as part of these projects has been the need for researchers to gain easier access to dark archives. Reflecting the priorities of the research community, the UKRI (UK Research and Innovation) Infrastructure Roadmap Progress Report (2019) states that making born-

40 "After the Digital Revolution" 2017-2018 (www.afterthedigitalrevolution.com) funded by a British Academy Rising Star award; AURA – Archives in the UK/ Republic of Ireland and AI 2020-2021 (www.aura-network.net) funded by the AHRC and the Irish Research Council; AEOLIAN – Artificial Intelligence for Cultural Organisations (www.aeolian-network.net) funded by the AHRC and the National Endowment for the Humanities.

digital archives "discoverable and accessible in a coherent fashion, in perpetuity" is a key objective.[41] The UK National Data Strategy (2020) also stresses the need for data to be "appropriately accessible, mobile and re-usable"[42] – which has an impact on researchers and other users. There is little doubt that more needs to be done to unlock archives that are currently closed to the public, and design thinking is part of the solution.

Indeed, design thinking is a productive method to solve the problems of *access* and *use* of born-digital archives. Academic and non-academic researchers should no longer take the back seat and wait for archivists to offer access. They should co-design access policies alongside archives professionals. The archival community must have a clear understanding of the needs of end users, before designing for machine-to-machine interoperability. For archivists, working closely with users is important to make a case for the relevance of their collections. It is also crucial for the archive sector to facilitate the use of innovative methodologies such as Artificial Intelligence. Not all users want to download data on their laptops and apply computer methods to produce new knowledge. Some users (myself included) favor hybrid approaches – for example combining close reading with network graphs, as explained in this chapter. But the increasing availability of AI platforms, tools and services (such as Microsoft Azure AI) will encourage a growing number of users to request access to large datasets rather than single items. Staying in touch with human users will allow archival institutions to remain relevant in an AI-dominated world.

Bibliography

Ackoff, Russell L./Emery, Fred E., On Purposeful Systems: An Interdisciplinary Analysis of Individual And Social Behavior As a System of Purposeful Events, New Brunswick, NJ, 1972.

Arnold, John E. Creative Engineering, https://stacks.stanford.edu/file/druid:jb100vs5745/Creative%20Engineering%20-%20John%20E.%20Arnold.pdf [last accessed: Mar. 30, 2021].

Aubry, Sara, Introducing Web Archives as a New Library Service: The Experience of the National Library of France, in: LIBER Quarterly 20 (2/2010), 179-199, doi:10.18352/lq.7987.

Baker, Fran, E-Mails to an Editor: Safeguarding the Literary Correspondence of the Twenty-First Century at The University of Manchester Library, in: New Re-

41 UKRI Infrastructure Roadmap Progress Report, 2019, see 59.

42 https://www.gov.uk/government/publications/uk-national-data-strategy[last accessed: April 5, 2021].

view of Academic Librarianship 21 (2/2015), 216-224, doi:10.1080/13614533.2015.1040925.

Berg, Maggie/Seeber, Barbara K., The Slow Professor: Challenging the Culture of Speed in the Academy, Toronto 2018.

Boltanski, Luc/Chiapello, Eve, The New Spirit of Capitalism, London 2007.

Brown, Tim, Change by Design: How Design Thinking Transforms Organizations and Inspires Innovation, New York 2019.

Cabinet Office (UK), Better Information for Better Government, 18.01.2017, https://www.gov.uk/government/publications/better-information-for-better-government [last accessed: Mar. 30, 2021].

Checkland, Peter, Systems Thinking, in: Wendy Currie/ Bob Galliers (eds.), Rethinking Management Information Systems: An Interdisciplinary Perspective, Oxford 1999, 45–56.

Christen, Kimberly/Anderson, Jane, Toward Slow Archives, in: Archival Science 19 (2/2019), 87–116, doi:10.1007/s10502-019-09307-x.

Churchman, C. West, Guest Editorial: Wicked Problems, in: Management Science 14 (4/1967), B141–42.

Draft Mandate: Archives by Design Working Group,
https://ec.europa.eu/transparency/regexpert/index.cfm?do=groupDetail.groupMeetingDoc&docid=34956 [last accessed: Mar. 30, 2021].

Ellis, Carolyn, The Ethnographic I: A Methodological Novel about Autoethnography, Walnut Creek, CA, 2004.

Erway, Ricky, Defining "Born Digital," Dublin, OH, 2010, URL: https://www.oclc.org/content/dam/research/activities/hiddencollections/borndigital.pdf [last accessed: April 5, 2021].

Goldman, Ben/Pyatt, Timothy D., Security Without Obscurity: Managing Personally Identifiable Information in Born-Digital Archives, in: Library & Archival Security 26 (1–2/2013), 37–55, doi:10.1080/01960075.2014.913966.

Greene, Mark/Meissner, Dennis, More Product, Less Process: Revamping Traditional Archival Processing, in: The American Archivist 68 (2/2005), 208-263, doi:10.17723/aarc.68.2.c741823776k65863.

Harris, Valerie/Stine, Kathryn, Politically Charged Records: A Case Study with Recommendations for Providing Access to a Challenging Collection, in: The American Archivist 74 (2/2011), 633–51, doi:10.17723/aarc.74.2.f252r28174251525.

Honoré, Carl, In Praise of Slow: How a Worldwide Movement Is Challenging the Cult of Speed, London 2010.

Jaillant, Lise, After the Digital Revolution: Working with Emails and Born-Digital Records in Literary and Publishers' Archives, in: Archives and Manuscripts 47 (3/2019), 285–304, doi:10.1080/01576895.2019.1640555.

Jaillant, Lise, From Letters to Emails: Reading Ian McEwan's Correspondence, TLS Online, 21.11.2017, https://www.the-tls.co.uk/articles/public/ian-mcewans-emails-letters/ [last accessed: Mar. 30, 2021].

Kastenhofer, Julia/Katuu, Shadrack, Declassification: A Clouded Environment, in: Archives and Records 37 (2/2016), 198–224, doi:10.1080/23257962.2016.1194814.

Larson, Lucy, Engaging Families in the Galleries Using Design Thinking, in: Journal of Museum Education 42 (4/2017), 376-384, doi:10.1080/10598650.2017.1379294.

MacLeod, Suzanne, et al, New Museum Design Cultures: Harnessing the Potential of Design and "Design Thinking" in Museums, in: Museum Management and Curatorship 30 (4/2015), 314-341, doi:10.1080/09647775.2015.1042513.

Mahey, Mahendra, Open a GLAM Lab. Digital Cultural Heritage Innovation Labs, Doha, Qatar, 2019.

Nelson, Harold G./Stolterman, Erik, The Design Way: Intentional Change in an Unpredictable World, Cambridge, MA, 2014.

Ovenden, Richard, Burning the Books: A History of Knowledge under Attack, Cambridge, MA, 2020.

Schneider, Josh, et al., Appraising, Processing, and Providing Access to Email in Contemporary Literary Archives, in: Archives and Manuscripts 47 (3/2019), 305-326, doi:10.1080/01576895.2019.1622138.

Seitz, Tim, Design Thinking and the New Spirit of Capitalism: Sociological Reflections on Innovation Culture, London 2020, doi:10.1007/978-3-030-31715-7.

Sjøvoll, Espen, Design Thinking & Innovation: Norwegian Approach, https://www.archivists.org.au/documents/item/1640 [last accessed: Mar. 30, 2021].

Sloyan, Victoria, Overview of a Born-Digital Archives Access Workshop Held at Wellcome Collection, London 2018, doi:10.6084/m9.figshare.6087194.v1.

Weber, Millicent/Buchanan, Rachel, Metadata as a Machine for Feeling in Germaine Greer's Archive, Archives and Manuscripts 47 (2/2019), 230-241, doi:10.1080/01576895.2019.1568266.

Will Self's Archive Acquired by the British Library - English and Drama Blog, 21.12.2016, https://blogs.bl.uk/english-and-drama/2016/12/will-selfs-archive-acquired-by-the-british-library.html [last accessed: Mar. 30, 2021].

Yakel, Elizabeth/Torres, Deborah A., AI: Archival Intelligence and User Expertise, in: The American Archivist 66 (1/2003), 51-78, http://www.jstor.org/stable/40294217 [last accessed: April 5, 2021].

Chapter 4: Towards Critically Addressable Data for Digital Library User Studies

Paul Gooding, University of Glasgow

Abstract

This chapter addresses two key questions: what can the concept of the black box add to our understanding of library catalogues as data in digital library user studies? And how might data-driven approaches help us to increase the transparency of these black boxes and render them critically addressable? Libraries are complex systems comprising a complex interrelationship of staff, space, users and technical infrastructure. However, digital library user studies have not applied the same attention to the creation of large-scale datasets as they have to the ethical and methodological implications of reporting on them. This chapter positions the study of library catalogue data in relation to black box theory and the collections as data imperative. It argues that collaborations between data science, the critical digital humanities, and library and information science can help us to be more transparent in how we reuse catalogue data, and to redefine how this data is created, processed, and documented in the first place.

Introduction

Libraries are information organisations consisting of a complex interrelationship of collections, staff, spaces, users, and technical infrastructures. They have long been conceived of as 'systems', a useful metaphor when viewed in light of Donella H. Meadow's definition: "a system isn't just any old collection of things. A **system** [author's emphasis] is an interconnected set of elements that is coherently organized in a way that achieves something."[1] Huge amounts of administrative data are produced across this system to describe resources and record actions, including: bibliographic metadata about information resources; (meta)data describing actions performed upon those information resources; user data including search queries, borrower records, access requests, and information seeking behaviour; and much more. These datasets have the potential to illuminate our understanding

1 Donella H. Meadows, *Thinking in Systems: A Primer*, ed. by Diana Wright London 2008, see 11.

of user behaviour online, and so can be of great value to researchers undertaking digital library user studies.

In general, user studies of this kind do not address the library as a complex system, and instead address a specific resource, user group, or content type. Many such studies utilize catalogue and user data to investigate information-seeking behaviour within a specific digital resource, using aggregated library patron data relating to user interactions with the online interface alongside administrative and bibliographic metadata relating to accessed information resources. Adam Chandler and Melissa Wallace draw attention to several projects that utilize Google Analytics for this purpose,[2] while other studies draw upon a wider range of methodological tools including the analysis of server weblogs, interviews, and surveys.[3] Such studies share common methodological approaches that derive from the multi- and inter-disciplinary nature of library and information studies (LIS), drawn from computational science, the social sciences, and the arts and humanities. In this chapter, I intend to focus on the latter, and specifically on how the epistemologies of the digital humanities (DH) can help us to interrogate the library catalogue as data. In this account, the library catalogue acts as the central point of interaction between digital library users and the resources, standards, technologies, categories and phenomenology that Bowker and Star argue converge in large-scale information systems.[4] To date, there remains a gap in our understanding of the link between library systems, library resources, and user behaviour, and particularly little attention to how the catalogue data that provides these insights is created, processed and received by researchers. In short, the data we rely upon to understand digital library usage too often represents a black box.

The name of this chapter takes inspiration from Bruno Latour's 1987 study of knowledge creation in the physical sciences, where he describes black boxes as devices that take some form of input, and provide some form of output, but which require no knowledge of their internal workings. I will establish the opening of these black boxes in digital library user studies as a humanistic problem, and ar-

2 Adam Chandler/Melissa Wallace, Using Piwik Instead of Google Analytics at the Cornell University Library, in: *The Serials Librarian* 71 (3-4/2016), 173-179, see 173-174, doi:10.1080/0361526X.2016.1245645.

3 For examples, Eric T. Meyer/Kathryn Eccles, *The Impacts of Digital Collections: Early English Books Online & House of Commons Parliamentary Papers*, London 2016, URL: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2740299 [last accessed: April 2, 2021]; Claire Warwick et al., If You Build It Will They Come? The LAIRAH Study: Quantifying the Use of Online Resources in the Arts and Humanities, in: *Literary and Linguistic Computing*, 23 (1/2008), 85-102, doi:10.1093/llc/fqm045.

4 Geoffrey C. Bowker/Susan Leigh Star, *Sorting Things Out: Classification and Its Consequences*, Boston, MA, 1999.

gue that the "collections as data" framework[5] provides a rich basis upon which to transform our understanding of library catalogue data as a humanities dataset. I will address two key questions: what can the concept of the black box add to our understanding of library catalogues as data in digital library user studies? And how might data-driven approaches help us to increase the transparency of these black boxes and render them critically addressable? My response is aimed at two audiences: those who wish to be more transparent in their own adoption of existing library catalogue data, and those who might seek to define how this data is created, gathered and processed in the first place. I will therefore finish by suggesting a set of guidelines by which each group might embed a more transparent approach to the adoption of library catalogue data in digital library user studies. My hope is that by making explicit the link between collections as data and catalogues as data, that the black boxes of library catalogue data may be more effectively refined and critiqued in future.

Black Boxes all the Way Down: The Research Context

The observations that inform this chapter were inspired by my work on *Digital Library Futures* (DLF), funded by the Arts and Humanities Research Council in the UK, which ran between 2017 and 2019. DLF set out to investigate the impact of Non-Print Legal Deposit (NPLD) upon UK academic deposit libraries and their users. Legal deposit ensures the systematic preservation of a nation's published output. It has existed in English law since 1662, and in British law since 1710.[6] Until recently the legal deposit regulations gave the libraries the right only to receive print materials including books, periodical, music and maps, but in 2013 the UK extended this right to incorporate digital materials.[7] My interest in this topic arose because the regulations emphasize the bequest value of legal deposit over contemporary usage, creating a tension between posterity-driven preservation and the theoretical accessibility benefits of digital materials. These tensions are represented in access protocols that can be described as "e-reading,"[8] the reading of single items on a screen, and within the reading rooms of the six legal deposit libraries. These

5 Thomas Padilla et al., *Santa Barbara Statement on Collections as Data*, 2017, URL: https://collectionsasdata.github.io/statement/ [last accessed: April 2, 2021].

6 Paul Gooding/Melissa Terras/Linda Berube, *Towards User-Centric Evaluation of Non-Print Legal Deposit: A Digital Library Futures White Paper*, URL: http://elegaldeposit.org/resources [last accessed: April 2, 2021].

7 *The Legal Deposit Libraries (Non-Print Works) Regulations 2013*, URL: http://www.legislation.gov.uk/uksi/2013/777/contents/made [last accessed: April 2, 2021].

8 Georgi Alexandrov, Transformation of Digital Legal Deposit in Terms of Legislation and Public Access, in: *Knygotyra* 70 (2018), 136-153, doi:10.15388/Knygotyra.2018.70.11812.

protocols contrast with libraries' efforts to widen online participation, and with shifting perspectives on the material forms of textual publications, exemplified by N. Katherine Hayles' argument that "the advent of electronic textuality presents us with an unparalleled opportunity to reformulate fundamental ideas about texts."[9] We were motivated to investigate these frictions, and set out to investigate how the new collections were being used by researchers within the UK academic deposit libraries.

My work sits within an established tradition of user studies that span LIS and the digital humanities (DH), and which are united by an explicitly humanistic perspective. Both fields share common methodological characteristics in how they address usage of digital library resources. They aim to inform working practices around digital library collections, and to develop theoretical and practical models of online information behaviour. Previous studies have evaluated the impact of online resources upon users, or developed theoretical and methodological frameworks for impact evaluation.[10] More broadly, DH has much to offer user studies in libraries through a shared interest in problems that are central to both fields, such as "information organization, information behaviour, information retrieval, sociotechnical systems, human-computer interaction, computer supported co-operative work, and information systems."[11] LIS as a field of study has long incorporated multi- and inter-disciplinary perspectives upon the central topic of research into human-recorded information,[12] a tradition that has been enriched by the explicitly humanistic and data-literate epistemology of the digital humanities. In doing so, these authors place digital collections into conversation with conceptions of library usage, and move us towards a rich set of questions around how online delivery of library resources impacts upon access and usage.

To address these questions in DLF, we adopted a mixed methods case study approach, combining expert interview, surveys, and webometric approaches including web analytics and a subject-based analysis of user requests to access NPLD resources. It is these last two methods that occupy me here, as they illuminate how library usage data comes from diverse sources and can resemble black boxes which render their internal workings invisible. Yet this data is imbued with biases

9 N. Katherine Hayles, *My Mother Was a Computer: Digital Subjects and Literary Texts*, Chicago 2005.

10 Lorna M. Hughes, ed., *Evaluating and Measuring the Value, Use and Impact of Digital Collections*, London 2012.

11 Tanya Clement/Daniel Carter, Connecting Theory and Practice in Digital Humanities Information Work, in: *Journal of the Association for Information Science and Technology* 68 (6/2017), 1385-1396, doi:10.1002/asi.23732.

12 David Bawden, Organised Complexity, Meaning and Understanding: An Approach for a Unified View of Information for Information Science, in: *ASLIB Proceedings* 59 (4/5/2007), 307-327, doi:10.1108/00012530710817546.

and assumptions that are embedded in their creation, processing and reception by scholars.

This chapter will outline a systemic view of digital libraries that embraces both the complexity and scale of system-level user analysis. To date, studies have focused on individual resources, or particular subsets of users, but addressing libraries as complex systems will increasingly require us to harness machine learning techniques to cope with data that is typologically diverse and large in scale. As I will indicate, the challenges associated with this are shared across many fields wrestling with new forms of data, and emerging forms of computational research. For instance, the debates around the topic of Explainable Artificial Intelligence (XAI) are highly relevant to this analysis. As Matt Turek has noted, work on the concept of XAI aims to create machine learning techniques that produce more explainable models, increase the transparency of machine learning models, and enable human users to understand, trust and manage artificial intelligence operations in their own work.[13] XAI requires the production of an explainable model, and some form of explanation interface that allows a user to understand the decisions that were made. For this reason, the metaphor of the black box provides a common ground with other disciplines, including that on XAI. Adadi and Berrada, for instance, explicitly refer to the black box in their survey of writing on XAI, and note it as a common term within the literature on AI and machine learning.[14]

As Computer Science has developed a theoretical basis for XAI, theorists within DH have developed varied critical theoretical lenses by which to address the study of humanities data. Both disciplines offer models for ethical, transparent utilisation of data, and of machine learning models for data analysis. A similar perspective on the question of library catalogues as data would, I will argue, greatly enrich digital library user studies. It would help to centre questions of data provenance and bias at the heart of our research. In order to establish the rationale for this more humanistic approach to catalogue data, I will explore how library user studies have engaged with the notion of libraries as complex systems before situating this history in relation to black box theory.

13 Matt Turek, Explainable Artificial Intelligence, URL: https://www.darpa.mil/program/explainable-artificial-intelligence [last accessed: April 2, 2021].

14 Amina Adabi/Mohammed Berrada, Peeking inside the Black-Box: A Survey on Explainable Artificial Intelligence, in: *IEEE Access*, 6 (2018), 52138-52160, see 52146, doi: 10.1109/ACCESS.2018.2870052.

Library User Studies and the Library as System

While user studies within LIS have a long history, T. D. Wilson's seminal paper from 1981[15] has been credited as one of the first works to establish a clear conceptual and methodological framework for library user studies. Wilson described his own paper as a "way of thinking about the field" of user studies.[16] The paper sought to differentiate between information needs as a specific need perceived by the user, and the behaviours that were taken to meet that need. Wilson argues that the study of information behaviour on its own does not necessarily address the information needs of users, suggesting that information-seeking behaviour should be placed within a much broader context. The library is positioned as an intermediary between the user's "life world" on the one hand, and the "embodiments of knowledge" that they need in order to meet an information need on the other.[17] In Wilson's model, the information system mediates between the user and the embodiments of knowledge, thus hinting at its overlapping and interlinked components:

> The user will be in contact with a variety of 'information systems', only one of which is shown in the diagram, hence the indicated overlap with the user and his life-world. Within the information system two subsystems are shown: the 'mediator' (generally a living system, i.e., a human being) and the 'technology', used here in the general sense of whatever combination of techniques, tools and machines constitute the information-searching subsystem.[18]

In LIS, a large amount of research exists that focuses on components of this system, primarily in relation to information behaviour rather than the broader contexts relating to user needs. Many of these studies incorporate data-driven methods derived from webometrics and informetrics. However, a gap remains in how we approach the data that underpins these studies, which are often received by researchers as figurative black boxes. This is not to say that there has been no criticism of webometric techniques within the LIS literature. For instance, Michael Thelwall has argued that the insights derived from webometric data are limited in the absence of other contextual factors.[19] This echoes Wilson's warning that early user studies failed to identify the wider context for information-seeking behaviour.[20]

---

15 T. D. Wilson, On User Studies and Information Needs, in: *Journal of Documentation*, 37 (1/1981), 3-15, doi:10.1108/eb026702.

16 Wilson, On User Studies and Information Needs, 4.

17 Wilson, On User Studies and Information Needs, 5-6.

18 Wilson, On User Studies and Information Needs, 6.

19 Mike Thelwall, *Introduction to Webometrics: Quantitative Research for the Social Sciences*, San Rafael, CA, 2009.

20 Wilson, On User Studies and Information Needs, 5.

Researchers have engaged with several aspects of this broader context. Mixed methods are extremely common in digital resource impact evaluation, while a large body of work exists to address the ethical concerns around using library catalogue data within research. These studies have addressed the reuse of library user data, addressing difficult questions around surveillance,[21] the role of research libraries in university-wide learning analytics programs,[22] student privacy,[23] and ethical data practices for librarians.[24] However, such debates focus largely upon the end use of this data, leaving a gap in our understanding of the contexts around the creation, processing, and management of library catalogue and user data. This chapter aims to address that context, the information systems that sit at the centre of Wilson's information-seeking model. In doing so, I hope to support critique of the black boxes that sit within these systems, and which constitute major data sources for LIS scholars. I will draw on existing concepts of the library as a system, in order to understand how the problem of black boxes arises in digital library user studies.

The Digital Library as a Complex System

In the introduction, I proposed that we might understand the library as a coherently organized system of interconnected elements. The idea that libraries can be understood as interoperable systems is by no means new, and indeed significantly predates Wilson's model. Indeed, Merrill M. Flood argued in 1964 that the services and collections offered by a single library are dependent upon a complex infrastructure of linked collections. These collections can be hosted locally, or be physically distributed across a geographic area; Flood, for instance, refers to the University of California, whose collections are dispersed across an entire state.[25] While Flood considers the collections as the key component of the library system, he does note that collections alone are not likely to be considered a library and that the system

21 Martin Patrick, Patron Data and the Fear of Surveillance: Some Thoughts, in: *Medium* 28.03.2016, URL: https://medium.com/@martinpatrick/patron-privacy-and-freedom-b6ebc625021a [last accessed: April 2, 2021].

22 Kyle M. L. Jones/Dorothea Salo, Learning Analytics and the Academic Library: Professional Ethics Commitments at a Crossroads, in: *College & Research Libraries* 79 (3/2018), 304-323, doi:10.5860/crl.79.3.304.

23 Alan Rubel/Kyle M. L. Jones, Student Privacy in Learning Analytics: An Information Ethics Perspective, in: *The Information Society*, 32 (2/2016), 143-159, doi:10.1080/01972243.2016.1130502.

24 Andrew Asher, Risk, Benefits, and User Privacy: Evaluating the Ethics of Library Data, in: Bobbi Newman/Bonnie Tijerina (eds.), *Protecting Patron Privacy: A LITA Guide*, Lanham, MD, 2017, 43-56.

25 Merrill M. Flood, The Systems Approach to Library Planning, in: *The Library Quarterly: Information, Community, Policy* 34 (4/1964), 326-338, doi:10.1086/619267.

also needs to incorporate discovery mechanisms and collections management processes: "in a sentence, a library is a collection of records together with a retrieval procedure. The total library system includes the libraries and the persons responsible for them, along with the concepts and procedures that guide the planning and operation of the entire system."[26]

The scope of this library system must inevitably be expanded somewhat to include the technical and physical infrastructures of digital libraries. The digital shift has seen the library system atomizes even further than in Flood's account; libraries now have physical and digital collections, with information resources and related data spread across multiple platforms. A large research library is likely to hold hundreds or thousands of subscriptions to online resources, spread across several proprietary data platforms provided by multiple vendors, each with their own data structures, usage metrics, and licensing conditions. A brief glance at OCLC, an American non-profit organisation that develops services to support access to information resources, shows the level of dispersion in digital library systems. OCLC works with thousands of partners including software vendors, library service providers, publishers, and major consumer services such as Google Books and Google Scholar. These partnerships allow them to provide access to millions of pages of digitized content, to hold descriptive metadata and cover art for over 1.2 billion resources, and make it accessible via a simple library discovery platform.[27] OCLC's services depend upon a myriad of dependencies, and yet as with all library service providers still do not encompass anywhere near the full range of resources available to libraries. The proliferation of platforms, stakeholders, and data creators means that many unseen assumptions are embedded into both library catalogues and the administrative data that structures and records user interactions.

The study of user data is also the study of interaction between users, information resources, and the library systems that act as intermediaries. Bowker and Star explore the role of classifications, in the broadest sense, in shaping our lives. They propose a helpful definition of infrastructure as an embedded structure "sunk into, inside of, other structures, social arrangements and technologies"[28] that has a temporal and physical reach. They identify as a key characteristic the way that infrastructure acts as an embodiment of standard that is plugged into other infrastructures and tools in a standardized way. This helps us to understand library systems as infrastructural representations of the standards, classifications and working practices that shape knowledge within that system. As Wilson argued, in order to fully understand usage of these systems we must also engage with the contexts,

26 Flood, The Systems Approach to Library Planning, 327.

27 OCLC, Examples of OCLC Partnerships, 2021, URL: https://www.oclc.org/en/partners-for-libraries/partnerships_examples.html [last accessed: April 2, 2021].

28 Bowker and Star, *Sorting Things Out: Classification and Its Consequences*, 35.

practices and assumptions that inform them. The process of opening these contexts up to scrutiny, of making visible the hidden, is therefore an important tool in the study of the interactions between complex systems and their users:

> Standards and classifications, however imbricated in our lives, are ordinarily invisible. The formal, bureaucratic ones trail behind them the entourage of permits, forms, numerals, and the sometimes-visible work of people who adjust them to make organizations run smoothly. In that sense, they become more visible, especially when they break down or become objects of contention.[29]

As this broad overview of library systems and user studies demonstrates, the context in which information seeking occurs in libraries is dependent upon an interwoven system of linked resources, datasets, standards, classifications and working practices. However, we often receive these datasets as if they were unmediated, raw data. Rosenberg, Jackson and Gitelman argue that this is rarely the case:

> At first glance, data are apparently before the fact: they are the starting point for what we know, who we are, and how we communicate. This shared sense of starting with data often led to an unnoticed assumption that data are transparent, that information is self-evident, the fundamental stuff of truth itself.[30]

If we treat library catalogues as data without understanding the context of how that data was actually created, then we risk misunderstanding the biases, assumptions and practices that inform its creation. As a result, user studies are at risk of treating the datasets that form the basis of quantitative research into library users as "black boxes" rather than embodiments of intricate human and technological infrastructures.

The Black Box in Science

The black box, as Philip von Hilgers and William Rauscher observe in their history of the concept, is both "word and thing."[31] The black box as a physical object became an idea of great interest to the scientific community in the 1950s and 1960s, leading to theorisation of its relevance in scientific work.[32] In the cybernetics community,

---

29 Bowker and Star, *Sorting Things Out: Classification and Its Consequences*, 2-3.

30 Dan Rosenberg/Virginia Jackson/Lisa Gitelman, Introduction, in: Lisa Gitelman (ed.),"*Raw Data" Is an Oxymoron*, Cambridge, MA, 2013.

31 Phillip Von Hilgers/William Rauscher, The History of the Black Box: The Clash of a Thing and Its Concept, in: *Cultural Politics*, 7 (1/2011), 41-58, see 45, URL: https://muse.jhu.edu/article/584290 [last accessed: April 2, 2021].

32 Von Hilgers/Rauscher, The History of the Black Box, 46-51.

researchers developed Black Box Theory as a means for understanding and manipulating complex systems that were too large to understand in other ways. In her thorough account of the history of Black Box Theory, Elizabeth Petrick argues that cyberneticians modelled the concept of black boxes to address two related desires: first, they wanted to be able to model complex systems using other systems, such as modeling the human brain using an electronic computer; and second, they wanted a way to understand inputs and outputs for modeling systems that were otherwise closed from inquiry.[33] To this end, the black box was theorized as a tool or system where only the inputs or outputs are known, and the inner workings obscured.

Since this foundational work, black box theory has been adopted broadly across the computer sciences, social sciences, and humanities disciplines such as philosophy. The characteristics of the black box remain relatively stable across this disciplinary spectrum: it continues to represent an object, method, or tool for which only the inputs and outputs are known. The questions posed by William Ashby in his *Introduction to Cybernetics* therefore continue to hold great resonance for digital library user studies:

> How should an experimenter proceed when faced with a Black Box?
> What properties of the Box's contents are discoverable, and what are fundamentally not discoverable?
> What methods should be used if the Box is to be investigated efficiently?[34]

Ashby introduces the possibility that some elements of the black box are fundamentally undiscoverable, a largely conceptual point that is nevertheless worth considering in a more literal sense in relation to questions of research integrity, data privacy, and commercially sensitive data sources.

Bruno Latour's investigation into scientific knowledge construction establishes black boxes as a metaphor for the progress of scientific research. He refers to black boxes as a way of compartmentalising knowledge that is no longer open to questioning, that is made more solid by using it without further questioning. In Latour's account, black boxes serve a particular purpose, acting as a form of shorthand that allows researchers to modularize particular tools, or knowledge, so that new ideas can be more efficiently developed.[35] Latour views the black box as a sign of success, an indication that a particular module of scientific knowledge is working ef-

---

33 Elizabeth R. Petrick, Building the Black Box: Cyberneticians and Complex Systems, in: *Science, Technology, & Human Value*, 45 (4/2019), 575-595, doi:10.1177%2F0162243919881212 [last accessed: April 2, 2021].

34 William R. Ashby, *An Introduction to Cybernetics*, London 1956, see 87.

35 Victoria Stodden, The Scientific Method in Practice: Reproducibility in the Computational Sciences, in: *MIT Sloan Research Papers*, 477 (3-10/2010), doi:10.2139/ssrn.1550193.

ficiently to the point that we only need to focus upon its inputs and outputs.[36] The construction of a black box in a certain community becomes a collective process, and as it becomes more integrated in the practices of that community it becomes increasingly difficult to revise and discourages questioning of results.[37] The black box therefore serves a particular purpose, allowing new ideas to be more efficiently built upon old, and saving the need to rehearse existing arguments around it. As a result, the black box becomes a shorthand for a complex set of "commands, machinery, or a methodology underlying a result."[38]

However, as I have argued above, the library catalogue as data is often received by researchers without this process of conceptual stabilisation within the community, and thereby we require additional labour to make the data transparent and open to critique. Critical addressability refers to the notion that one should be able to evaluate the technical and social forces that shape data, through data documentation and transparent workflows. Thomas Padilla describes the minimum requirements for data to be critically addressable:

> A researcher should be able to understand why certain data were included and excluded, why certain transformations were made, who made those transformations, and at the same time a researcher should have access to the code and tools that were used to effect those transformations.[39]

In the following section, I will address the concept of "collections as data",[40] which has emerged in response to a particular moment when cultural heritage organisations are wrestling with how to operationalize their collections for digital scholarship in the humanities. I will explore the humanistic roots of the collections as data imperative, before turning to consider how collaborations between LIS and DH researchers and data scientist might enhance the field of digital library user studies.

The Collections as Data Imperative

The Santa Barbara Statement on Collections as Data sets out ten principles for thinking of cultural heritage collections as data, imagining it as an ongoing process of making collections more accessible, transparent, interoperable and, ultimately,

36 Bruno Latour, *Science in Action: How to Follow Scientists and Engineers through Society*, Cambridge, MA, 1987, see 304.

37 Stodden, The Scientific Method in Practice, 5.

38 Stodden, The Scientific Method in Practice, 5.

39 Thomas Padilla, Humanities Data in the Library: Integrity, Form, Access, in: *D-Lib Magazine* 22 (3-4/2016), doi:10.1045/march2016-padilla.

40 Padilla et al., *Santa Barbara Statement on Collections as Data*.

readable as humanities datasets.[41] It takes as a starting point the idea that the digital shift requires us to reframe all digital objects as data, and to thereby address the shift in collections management workflows that must occur in response. The principles have helped to develop a clear community of practice spanning LIS and DH, with a common goal of increasing the usability of library collections within digital scholarship. The digital humanities contribute a strong critique of the notion of so-called "raw data"[42] through critical theoretical lenses from fields as diverse as feminist studies, critical race studies, sexuality studies, queer theory, and class studies. These scholars share a desire to make legible the assumptions that underpin the creation, manipulation and analysis of humanities data. Roopika Risam applies an intersectional frame, addressing the ways in which race, class, gender and other aspects of identify overlap with each other, to the practices of producing digital humanities data. Risam notes that:

> Existing digital humanities projects provide examples of how, in small and large ways, theory and method can be combined to address recurring questions of the role of race, class, gender, ability, sexuality, nationality, and other categories of difference within the field. These phenomena subtend the development and production of digital humanities projects but they may not be evident. Therefore, it is incumbent on us, as digital humanities practitioners, to make them legible, to move them beyond the margins.[43]

Risam's work, along with that of Miriam Posner[44] and Victoria Stodden,[45] has inspired Padilla to propose a simple rubric for evaluating the readiness of humanities collections for digital forms of scholarship:

- **Posner:** to what extent is information about Humanities data collection provenance, processing, and method of presentation available to the user?
- **Stodden:** to what extent are data and the code that generates data available to the user?
- **Risam:** to what extent are the motivations driving all of the above available to the user?[46]

---

41 Padilla et al., *Santa Barbara Statement on Collections as Data*.

42 Rosenberg, Jackson, and Gitelman, Introduction.

43 Roopika Risam, Beyond the Margins: Intersectionality and the Digital Humanities, in: *Digital Humanities Quarterly*, 9 (2/2015), URL: http://www.digitalhumanities.org/dhq/vol/9/2/000208/000208.html [last accessed: April 2, 2021].

44 Miriam Posner, *How Did They Make That?*, 29.08.2013, URL: http://miriamposner.com/blog/how-did-they-make-that/ [last accessed: April 2, 2021].

45 Stodden, The Scientific Method in Practice.

46 Padilla, Humanities Data in the Library: Integrity, Form, Access.

In this respect, collections as data link to wider imperatives for research data to be open, accessible and interoperable. For instance, the FAIR Data Principles provide a general set of guidance for making scientific data "Findable, Accessible, Interoperable, and Reusable."[47] The FAIR Data Principles have been influential in the library sector, and encompass specific principles aimed at ensuring richly described, transparent, and interoperable data that is suited to both human-driven and machine-driven activities. For humans, the principles focus upon the semantics, or contexts, of data and digital objects, while for machines the focus is upon developing "steps along a path" towards data that is more easily machine-actionable.[48] Collections as data, with its foundations in the critical theory of the digital humanities, develops principles that are more explicitly humanistic in nature.

Padilla notes that the whole library sector is already rising to the challenge of reframing humanities information resources as data, and what I propose here is closely aligned but subtly different. Whereas collections as data reimagines existing humanities information resources as data, emphasising considerations of form, integrity, and access,[49] here I propose that the first step in imagining library catalogues as data is to address them as a humanities information resource. Johanna Drucker usefully distinguishes between data as a "given" that is able to be recorded and observed, and "capta" that is actively taken.[50] By viewing data as something that is created, and constituted, she identifies several humanistic principles for creating and analysing data:

> Humanistic Inquiry acknowledges the situated, partial and constitutive character of knowledge production, the recognition that knowledge is constructed, *taken*, not simply given as a natural representation of pre-existing fact.[51]

Based on these foundations, I will argue that it is precisely this conceptual shift that is required in relation to library catalogues as data. While humanistic perspectives embed new forms of interpretation, it is their combination with data science approaches that offers us the opportunity to fully address the mode of capture and analysis of catalogue data.

---

47 Mark D. Wilkinson et al., The FAIR Guiding Principles for Scientific Data Management and Stewardship, in: *Scientific Data*, 3 (1/2016), doi:10.1038/sdata.2016.18.

48 Wilkinson et al., The FAIR Guiding Principles, 3.

49 Padilla, Humanities Data in the Library: Integrity, Form, Access.

50 Johanna Drucker, Humanities Approaches to Graphical Display, in: *Digital Humanities Quarterly* (2011) URL: http://www.digitalhumanities.org/dhq/vol/5/1/000091/000091.html [last accessed: April 2, 2021].

51 Drucker, Humanities Approaches to Graphical Display.

Interdisciplinary Collaborations: Towards Humanistic Catalogue Datasets

The previous section has dealt with the humanistic foundations of the collections as data imperative, and how they overlap with the objectives of the FAIR Data Principles to emphasize the core values that would help us to break out of catalogue datasets as black boxes. The overall aim is for research into library catalogues as data that aspire to greater transparency, and that account for the way that broader professional, technical, social and user contexts embed particular forms of meaning within administrative and bibliographic datasets relating to information resources and their usage. It is certainly the case that data within the library catalogue system represents a black box, in many cases one that has never truly been opened up to scrutiny. Proprietary data analytics platforms present one particular example: web analytics have become an almost ubiquitous data source in digital library user studies. The field of web analytics aims to collect, analyse and report data relating to web traffic, with a particular focus on improving website effectiveness. However, these platforms are often proprietary in nature and do not provide researchers with unprocessed data. In many cases, the researcher will also be unable to access the tools and code that were used to effect transformations and present results to users. Google Analytics (GA), due to its extremely high profile, has been the brunt of just such criticism; it is very obviously a black box, with its raw data inaccessible and hidden from users for reasons including data privacy. Adam Chandler and Melissa Wallace provide a useful introduction to the adoption of GA in library studies, addressing its limitations in relation to user privacy. They criticize the inability to undertake local data collection, and the willingness of libraries to accept a loss of control of user privacy by allowing patron data to be owned by Google and stored on its US-based servers.[52] This is highly problematic for researchers in light of the UK and EU General Data Protection Regulations, which remain largely aligned in 2021 and mandate that data should not be transferred to legislatures without similar data protection requirements. Google has taken several actions to comply with GDPR, but the situation around data residency is evolving and fluid, and largely outside the control of libraries. While organisations can take actions to ensure their implementation of GA is GDPR compliant, then, users face the problem that data processing is offsite and opaque. The inaccessibility of the underlying data renders GA a black box, with only the inputs and outputs known.

Platforms such as GA fall well short of the FAIR Data Principles, as well as the test of their critical addressability, even if their speed and convenience makes them an invaluable tool for user studies. However, other less obvious forms of black box

52 Chandler and Wallace, Using Piwik Instead of Google Analytics, 173-179.

also exist. This was a problem we faced in the DLF project, where we chose to utilize a subject-based approach to the analysis of users of NPLD. We were inspired by Marcia Bates' observation that scholarly communication practices function differently across domains, and set out to see the extent to which subject-based practices were visible in access requests for NPLD materials. We therefore developed a methodology that applied Dewey Decimal Classifications (DDC) to access logs which recorded all requests for NPLD eBooks and eJournals in the legal deposit library reading rooms. This made it possible to infer information about users at an aggregate level, and to explore the frequency with which information resources from each discipline were accessed. This method gave us clear and valuable insights to compare usage of NPLD resources against established information-seeking behaviour in various academic disciplines, and further details were reported in our white paper.[53] Here, though, I will focus briefly upon the problems of adopting a particular classification scheme for this purpose. While the insights that we gained were highly relevant to NPLD in the United Kingdom, the wider contexts informing the DDC classification scheme suggest limits to this method's broader applicability.

The data we used, of course, was not neutral. This is a general feature of library classifications that attempt to provide a universal view of the world's knowledge, which is a subjective process that is undertaken by humans and reflects existing biases.[54] The Dewey Decimal Classification which we used was conceived by Melvil Dewey in 1873 and published in 1876 as a "general organizational tool that is continuously revised to keep pace with knowledge."[55] It provides notations in Arabic Numerals from 000 to 999, and utilizes a hierarchy based upon ten top level disciplinary classes. According to OCLC, who have administrative responsibility for publishing and maintaining DDC, it is the world's most widely used library classification, and is used by libraries in at least 138 countries.[56] Despite its widespread adoption, several accounts exist that address the biased perspective that arises from its historic origins. For instance, Kua has criticized its poor representation of non-Western languages and literatures. It provides categories for "literature, rhetoric and criticism" that use national boundaries informed by a nineteenth century North American perspective of the most important schools of literature.[57] As a result, the rest of the world is lumped under a single class of "other literatures,"

53 Gooding/Terras/Berube, *Towards User-Centric Evaluation of Non-Print Legal Deposit.*

54 Jens-Erik Mai, Classification in a Social World: Bias and Trust, in: *Journal of Documentation*, 66 (5/2010), 627-642, doi:10.1108/00220411011066763.

55 OCLC, Introduction to the Dewey Decimal Classification, 2019, URL: https://www.oclc.org/content/dam/oclc/dewey/versions/print/intro.pdf [last accessed: April 2, 2021].

56 OCLC, Introduction to the Dewey Decimal Classification, 2.

57 Eunice Kua, Non-Western Languages and Literatures in the Dewey Decimal Classification Scheme, in: *Libri* 54 (4/2008), 256-265, doi:10.1515/LIBR.2004.256.

occluding much of the granularity that is afforded classical and canonical literary traditions. Further criticisms have included its marginalization of certain sexual orientations and sexual preferences as forms of perversion,[58] and its lack of flexibility that amplifies existing inaccuracies and limitations.

These examples both support Latour's assertion that it is far simpler to understand how a black box works if we are able to witness, or to be involved in, its creation: "instead of black boxing the technical aspects of science and *then* looking for social influences and biases, we realised… how much simpler it was to be there *before* the box closes."[59] How, then, is library classification a black box? It is much less opaque than something like Google Analytics, after all. Yet it still shares features that indicate a lack of transparency. We used an automated tool to apply classmarks to the dataset, due to its size. As such, it shares several features that indicate its status as a black box: it still takes an input, in the form of an information resource, and provides an output, in the form of a library classmark. The process is only transparent insofar as we are open to engaging with the biases of the classification, and automated tools, used in our research. Indeed, Mai usefully distinguishes between "administrative authority" – those individuals trusted to design and edit classifications as a technical task – and "cognitive authority" – those who can be trusted to make ontological statements about the relationships between entities:

> Cognitive authorities are those people we turn to for knowledge, insight, and advice on particular matters. These people have gained their authority not by being chosen by some body but because we let them influence our thinking. We recognize that there are certain people whom we trust on particular matters and other people we trust on other matters.[60]

Mai argues that transparency is a key criterion for granting cognitive authority to particular classifications.[61] This does not necessarily entail agreeing with the structure of the data to find it trustworthy: rather, closed systems that rely upon their administrative authority appear less trustworthy than those that acknowledge their partiality and bias in a transparent, documentable manner.

This final point is an important one, because it hints at the role of data science in collaborations around library catalogues as data. Analytics platforms simply do not meet most of the FAIR Data Principles, or the rubric for humanities data proposed by Padilla, and I would argue that the user studies community has never

58 Melissa Adler, *Cruising the Library: Perversities in the Organization of Knowledge*, New York 2017.

59 Latour, *Science in Action: How to Follow Scientists and Engineers through Society*, 21.

60 Mai, Classification in a Social World: Bias and Trust, 635-636.

61 Mai, Classification in a Social World: Bias and Trust, 639.

opened these black boxes up to full scrutiny. However, their ubiquity within library systems and digital library user systems suggests they fill a clear and obvious need within the research community. What I therefore propose is that we approach transparency around library catalogue data in a layered and reflective manner, with data literacy at one end and the definition and creation of new data structures at the other. As a starting point, I envisage three layers: first, datasets that are unable to be critically addressed due to the inaccessibility of unprocessed data; second, datasets that are "received" but allows a degree of critical addressability due to documentation, code, accessible data, or other indicators of transparency; and third, datasets that are "captured" by researchers to *be* critically addressable. The latter type represent an ontological argument that attempts to embed a particular critical theoretical perspective into the act of data creation, processing, and analysis, and is the category where I foresee a major role for data scientists and digital humanists. I would therefore propose differing approaches, based upon the degree of critical addressability of the available data:

1. For data that is not critically addressable (e.g., Google Analytics, where "raw" data and code are both hidden), researchers should:
    - Be open in their methodology about what aspects of the dataset are not critically addressable.
    - Investigate whether there are alternative data sources that would provide a greater degree of transparency.
    - Transparently record the extent to which their dataset corresponds to the FAIR data principles.
2. For data that is received, and relies on existing classifications and infrastructures (e.g., subject-based analysis that utilizes existing library classifications), researchers should:
    - Engage with the broader literature to identify the broader epistemological and methodological contexts that inform the creation of the data.
    - Consider the extent to which those contexts shape their results.
    - Transparently record both the actions that they take in their own data analysis, and the assumptions and processes behind its creation.
3. For data that is captured by the researcher, in a manner that defines how it is created, organized and processed (e.g., subject-based analysis with categories defined by the researcher), researchers should:
    - Engage in interdisciplinary collaboration to ensure that data is critically addressable.
    - Transparently record the actions that they take in the entire data lifecycle.

Data scientists have begun to intervene already, with proposals for bias-aware methodologies in natural language processing research,[62] and interventions that seek to apply data visualisation approaches to library catalogue data. However there remains a great opportunity for work that combines these approaches to address the library catalogue as a humanistic data source. In particular, there is a rich opportunity to investigate the application of NLP methodologies to the study of bibliographic metadata, and to contextualize it in relation to datasets representing the wider library system. Furthermore, informed by Victoria Stodden's argument of how black boxes discourage dissent, we should apply two generalized criteria: first, that the community should have collectively, and adequately, interrogated a tool before it becomes a black box; and second, that researchers should be transparent about the biases and assumptions that go into the creation and analysis of their own data, to the extent that this is practicable.

Conclusion

I have argued in this chapter that researchers engaged in digital library user studies must take further steps to ensure the critical addressability of library catalogue and user data. Library practitioners and researchers have defined good practice in the ethical reuse of catalogue and user data, but I believe we still have more work to do to understand the data itself. Many datasets are received as black boxes, with only the input and output known, and therefore fall short of the requirements for transparency and addressability of the FAIR Data Principles and the rubrics required to ensure critical addressability. The library catalogue, and the data on collections, library work, infrastructural contexts, and users, are where I believe that interdisciplinary collaboration can support us to begin to open these black boxes: both to more transparently interrogate the data that we receive, and to look towards DH and data science to actually redefine how it is created and processed. Furthermore, as the complexity of library data increases, it is likely that the field will move towards forms of analysis that incorporate Machine Learning to support the analysis of linked, typologically diverse datasets. The field therefore has much to learn from how Computer Science has theorized Explainable Artificial Intelligence, and how Data Science and the Digital Humanities have addressed the notion of critically addressable data. I propose that this work should proceed on a humanistic basis, and have laid out a broad critical context for doing so based on the complementary theoretical lenses of black boxes and collections as data which speak to this broad

62 Lucy Havens et al., Situated Data, Situated Systems: A Methodology to Engage with Power Relations in Natural Language Processing Research, in: *The 28th International Conference on Computational Linguistics*, URL: https://arxiv.org/abs/2011.05911 [last accessed: April 2, 2021].

interdisciplinary framework. The key intervention of this chapter is to position catalogue data within an explicitly humanistic framework, thus opening it up to the same forms of critique as collections data.

I have proposed some preliminary ideas for how we might go about critically addressing catalogues as data, but intend this chapter to be a starting point for further discussion rather than a definitive perspective on the subject. The proposals recognize the complexity of the digital library "system" that sits at the interface between users and information resources, and that this will influence the adoption of opaque datasets by individual research practitioners. In all cases, my suggestions represent an attempt not only to think of the contexts that surround information seeking behaviour, but the data that we use to study this behaviour. Black box theory helps to underline suitable approaches to the development of knowledge on catalogues as data, while the collections as data principles can aid us to establish a humanistic perspective to the resultant interrogation of data creation, management, and analysis. The challenge that remains unanswered by this chapter, and the focus of potential collaborative activity, is what might it look like in practice to investigate library catalogues as data? And just as importantly, how might data-driven approaches help us to understand, and even to address, the contexts for data capture in digital library systems?

A focus for future work should be on developing prototypical work for putting these principles into practice. These interventions must develop workflows that address large-scale, fragmented, library ecosystems. Data-driven approaches, informed by cutting edge critical humanistic theory, allow researchers to embrace pluralism in the ontological statements that become embedded in our library datasets. By adopting a position of critical addressability, we can account for the assumptions, decisions, and labour that underpin the order and usage of library collections as data. In doing so, I believe we can begin to imagine a more explicitly humanist basis by which to analyse the creation and capture of library user data that has traditionally been treated as largely administrative.

Bibliography

Adabi, Amina/Berrada, Mohammed, Peeking inside the Black-Box: A Survey on Explainable Artificial Intelligence, in: *IEEE Access*, 6 (2018), 52138-52160, doi: 10.1109/ACCESS.2018.2870052.

Adler, Melissa, *Cruising the Library: Perversities in the Organization of Knowledge*, New York 2017.

Alexandrov, Georgi, Transformation of Digital Legal Deposit in Terms of Legislation and Public Access, in: *Knygotyra* 70 (2018), 136-153, doi:10.15388/Knygotyra.2018.70.11812.

Ashby, William R., *An Introduction to Cybernetics*, London 1956.

Asher, Andrew, Risk, Benefits, and User Privacy: Evaluating the Ethics of Library Data, in: Bobbi Newman/Bonnie Tijerina (eds.), *Protecting Patron Privacy: A LITA Guide*, Lanham, MD, 2017, 43-56.

Bawden, David, Organised Complexity, Meaning and Understanding: An Approach for a Unified View of Information for Information Science, in: *ASLIB Proceedings* 59 (4/5 /2007), 307-327, doi:10.1108/00012530710817546.

Bowker, Geoffrey C./Star, Susan Leigh, *Sorting Things Out: Classification and Its Consequences*, Boston, MA, 1999.

Chandler, Adam/Wallace, Melissa, Using Piwik Instead of Google Analytics at the Cornell University Library, in: *The Serials Librarian* 71 (3-4/2016), 173-179, doi:10.1080/0361526X.2016.1245645.

Clement, Tanya/Carter, Daniel, Connecting Theory and Practice in Digital Humanities Information Work, in: *Journal of the Association for Information Science and Technology* 68 (6/2017), 1385-1396, doi:10.1002/asi.23732.

Drucker, Johanna, Humanities Approaches to Graphical Display, in: *Digital Humanities Quarterly* (2011) URL: http://www.digitalhumanities.org/dhq/vol/5/1/000091/000091.html [last accessed: April 2, 2021].

Flood, Merrill M., The Systems Approach to Library Planning, in: *The Library Quarterly: Information, Community, Policy* 34 (4/1964), 326-338, doi:10.1086/619267.

Gooding, Paul/Terras, Melissa/Berube, Linda, *Towards User-Centric Evaluation of Non-Print Legal Deposit: A Digital Library Futures White Paper*, URL: http://elegaldeposit.org/resources [last accessed: April 2, 2021].

Havens, Lucy, et al., Situated Data, Situated Systems: A Methodology to Engage with Power Relations in Natural Language Processing Research, in: *The 28th International Conference on Computational Linguistics*, URL: https://arxiv.org/abs/2011.05911 [last accessed: April 2, 2021].

Hayles, N. Katherine, *My Mother Was a Computer: Digital Subjects and Literary Texts*, Chicago 2005.

Hughes, Lorna M., ed., *Evaluating and Measuring the Value, Use and Impact of Digital Collections*, London 2012.

Jones, Kyle M. L./Salo, Dorothea, Learning Analytics and the Academic Library: Professional Ethics Commitments at a Crossroads, in: *College & Research Libraries* 79 (3/2018), 304-323, doi:10.5860/crl.79.3.304.

Kua, Eunice, Non-Western Languages and Literatures in the Dewey Decimal Classification Scheme, in: *Libri* 54 (4/2008), 256-265, doi:10.1515/LIBR.2004.256.

Latour, Bruno, *Science in Action: How to Follow Scientists and Engineers through Society*, Cambridge, MA, 1987.

Mai, Jens-Erik, Classification in a Social World: Bias and Trust, in: *Journal of Documentation*, 66 (5/2010), 627-642, doi:10.1108/00220411011066763.

Meadows, Donella H., *Thinking in Systems: A Primer*, ed. by Diana Wright, London 2008.

Meyer, Eric T./Eccles, Kathryn, *The Impacts of Digital Collections: Early English Books Online & House of Commons Parliamentary Papers*, London 2016, URL: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2740299 [last accessed: April 2, 2021].

OCLC, Examples of OCLC Partnerships, 2021, URL: https://www.oclc.org/en/partners-for-libraries/partnerships_examples.html [last accessed: April 2, 2021].

OCLC, Introduction to the Dewey Decimal Classification, 2019, URL: https://www.oclc.org/content/dam/oclc/dewey/versions/print/intro.pdf [last accessed: April 2, 2021].

Padilla, Thomas, Humanities Data in the Library: Integrity, Form, Access, in: *D-Lib Magazine* 22 (3-4/2016), doi:10.1045/march2016-padilla.

Padilla, Thomas, et al., Santa Barbara Statement on Collections as Data, 2017, URL: https://collectionsasdata.github.io/statement/ [last accessed: April 2, 2021].

Patrick, Martin, Patron Data and the Fear of Surveillance: Some Thoughts, in: *Medium* 28.03.2016, URL: https://medium.com/@martinpatrick/patron-privacy-and-freedom-b6ebc625021a [last accessed: April 2, 2021].

Petrick, Elizabeth R., Building the Black Box: Cyberneticians and Complex Systems, in: *Science, Technology, & Human Value*, 45 (4/2019), 575-595, doi:10.1177%2F0162243919881212 [last accessed: April 2, 2021].

Posner, Miriam, *How Did They Make That?*, 29.08.2013, URL: http://miriamposner.com/blog/how-did-they-make-that/ [last accessed: April 2, 2021].

Risam, Roopika, Beyond the Margins: Intersectionality and the Digital Humanities, in: *Digital Humanities Quarterly*, 9 (2/2015), URL: http://www.digitalhumanities.org/dhq/vol/9/2/000208/000208.html [last accessed: April 2, 2021].

Rosenberg, Dan/Jackson, Virginia/Gitelman, Lisa, Introduction, in: Lisa Gitelman (ed.), *"Raw Data" Is an Oxymoron*, Cambridge, MA, 2013.

Rubel, Alan/Jones, Kyle M. L., Student Privacy in Learning Analytics: An Information Ethics Perspective, in: *The Information Society*, 32 (2/2016), 143-159, doi:10.1080/01972243.2016.1130502.

Stodden, Victoria, The Scientific Method in Practice: Reproducibility in the Computational Sciences, in: *MIT Sloan Research Papers*, 477 (3-10/2010), doi:10.2139/ssrn.1550193.

The Legal Deposit Libraries (Non-Print Works) Regulations 2013, URL: http://www.legislation.gov.uk/uksi/2013/777/contents/made [last accessed: April 2, 2021].

Thelwall, Mike, *Introduction to Webometrics: Quantitative Research for the Social Sciences*, San Rafael, CA, 2009.

Turek, Matt, Explainable Artificial Intelligence, URL: https://www.darpa.mil/program/explainable-artificial-intelligence [last accessed: April 2, 2021].

Von Hilgers, Phillip/Rauscher, William, The History of the Black Box: The Clash of a Thing and Its Concept, in: *Cultural Politics*, 7 (1/2011), 41-58, URL: https://muse.jhu.edu/article/584290 [last accessed: April 2, 2021].

Warwick, Claire, et al., If You Build It Will They Come? The LAIRAH Study: Quantifying the Use of Online Resources in the Arts and Humanities, in: *Literary and Linguistic Computing*, 23 (1/2008), 85-102, doi:10.1093/llc/fqm045.

Wilkinson, Mark D., et al., The FAIR Guiding Principles for Scientific Data Management and Stewardship, in: *Scientific Data*, 3 (1/2016), doi:10.1038/sdata.2016.18.

Wilson, T. D., On User Studies and Information Needs, in: *Journal of Documentation*, 37 (1/1981), 3-15, doi:10.1108/eb026702.

Chapter 5: Reviewing the Reviewers: Training Neural Networks to Read Peer Review Reports

Martin Paul Eve, Birkbeck, University of London | Robert Gadie, University of the Arts, London | Victoria Odeniyi, University of the Arts, London | Shahina Parvin, Brandon University, Canada and Jahangirnagar University, Bangladesh.

Abstract

The study of academic peer review is often difficult owing to the confidentiality of reports. As an occluded genre of writing that nonetheless underpins scientific publication, relatively little is known about the ways that academics write and behave, at scale, in their reviewing practices. In this chapter, we describe for the first time the database of peer review reports at PLOS ONE, the largest scientific journal in the world, to which we had unique access. Specifically, we detail the approach that we took to training a multi-label, multi-class text classifier using the TenCent NeuralClassifier toolkit to examine the peer review reports. Although this resulted in a predictable failure to produce accurate levels of recall and precision, we argue that as these technologies further develop there are a range of uses – for both good and ill – that could be used to machine-read these archives.

1. Introduction - Reading Peer Review

Peer review is the system by which manuscripts are vetted for validity, appraised for originality, and selected for publication as articles in academic journals (serials) or as academic books (monographs).[1] Since an editor of an academic title cannot be expected to be an expert in every single area covered by a publication and since it appears undesirable to have a single person controlling the publication's flow of scientific and humanistic knowledge, there is a need for input from more people. Manuscripts submitted for consideration are shown to external expert advisers (peers) who deliver verdicts on the novelty of the work, criticisms or praise of the piece, and a judgement of whether to proceed to publication. A network of experts

1 Portions of this chapter are adapted from the openly licensed Martin Paul Eve et al, *Peer Review and Institutional Change in Academia*, Cambridge 2021.

with appropriate degrees of knowledge and experience within a field are coordinated to yield a set of checks and balances for the scientific and broader research landscapes. Editors are then bound, with some caveats and to some extent, to respect these external judgements in their own decisions, regardless of how harsh the mythical "reviewer 2" may be.[2]

The premise behind peer review may appear sound or even incontrovertible. Who could object to the best in the world appraising one another, nobly ensuring the integrity of the world's official research record? Yet, considering the system for even a few moments leads to several questions. What is a "peer" and who decides? What does it mean when a "peer" approves somebody else's work? How many "peers" are required before a manuscript can be properly vetted? What happens if "peers" disagree with each other? Does (or should) peer review operate in exactly the same fashion in disciplines as distinct as Neuroscience and Sculpture? Particle Physics and Social Geography? Math and Literary Criticism? When academics rely on publications for their job appointments and promotions, how does peer review interact with other power structures in universities? Do reviewers act with honour and integrity in their judgements within this system?

It is abundantly clear that the peer-review process is far from infallible. Every year, thousands of articles are retracted (withdrawn) for containing inaccuracies, for conducting unethical research practices, and for many other reasons.[3] On occasion, this has had devastating consequences in spaces such as public health. The

2 The age-old academic joke is that if the first reviewer loves an article, the second reviewer will hate it and be ultra-harsh in his or her judgement. Von Bakanic/Clark McPhail/Rita J. Simon, Mixed Messages: Referees' Comments on the Manuscripts They Review, in: *The Sociological Quarterly*, 30 (4/1989), 639-654, URL: http://www.jstor.org/stable/4121469; Lutz Bornmann/Hans-Dieter Daniel, The Effectiveness of the Peer Review Process: Inter-Referee Agreement and Predictive Validity of Manuscript Refereeing at *Angewandte Chemie*, in: *Angewandte Chemie International Edition*, 47 (38/2008), 7173-7178, doi:10.1002/anie.200800513; Louis Fogg/Donald W. Fiske, Foretelling the Judgments of Reviewers and Editors, in: *American Psychologist*, 48 (3/1993), 293-294, doi:10.1037/0003-066X.48.3.293; Stephen Lock, *A Difficult Balance: Editorial Peer Review in Medicine*, Philadelphia, PA, 1986; Richard E. Petty/Monique A. Fleming/Leandre R. Fabrigar, The Review Process at PSPB: Correlates of Interreviewer Agreement and Manuscript Acceptance, in: *Personality and Social Psychology Bulletin*, 25 (2/1999), 188-203, doi:10.1177/0146167299025002005; Robert J. Sternberg et al., Getting in: Criteria for Acceptance of Manuscripts in Psychological Bulletin, 1993-1996, in: *Psychological Bulletin*, 121 (2/1997), 321-323, doi:10.1037/0033-2909.121.2.321; Harriet Zuckerman/Robert K. Merton, Patterns of Evaluation in Science: Institutionalisation, Structure and Functions of the Referee System, in: *Minerva*, 9 (1/1971), 66-100, doi:10.1007/BF01553188.

3 Jeffrey Brainard/Jia You, What a Massive Database of Retracted Papers Reveals about Science Publishing's "Death Penalty," in: *Science | AAAS*, 25.10.2018, URL: https://www.sciencemag.org/news/2018/10/what-massive-database-retracted-papers-reveals-about-science-publishing-s-death-penalty [last accessed: Mar. 31, 2021]; Retraction Watch, URL: https://retractionwatch.com/ [last accessed: Mar. 31, 2021]. Björn Brembs/Katherine Button/Marcus Munafò,

at-the-time respected researcher Andrew Wakefield's notorious 1998 retracted paper claiming a link between the mumps, measles, and rubella (MMR) vaccine and the development of autism in children was published in perhaps the most prestigious medical journal in the world, *The Lancet*.[4] The work was undoubtedly subject to stringent single-blind pre-publication review and was cleared for publication. Yet the article was later retracted and branded fraudulent, having caused immense and ongoing damage to public health.[5] It is, alas, always easier to make an initial statement than subsequently to retract or to correct it. As a result, a worldwide anti-vaccination movement has seized upon this circumstance as evidence of a conspiracy. The logic uses the supposed initial validation of peer review and the prestige of *The Lancet* as evidence that Wakefield was correct and that he is the victim of a conspiratorial plot to suppress his findings. Hence, when peer review goes wrong, the general belief in its efficacy, coupled with the prestige of journals founded on the supposed expertise of peer review, has damaging real-world effects. Indeed, there are longstanding criticisms of the validity of peer review, exemplified in Franz J. Ingelfinger's notorious statement that the process is "only moderately better than chance" and Drummond Rennie's (the then deputy editor of the Journal of the American Medical Association) "if peer review was a drug it would never be allowed onto the market."[6]

Despite the aforementioned challenges, the role of peer review in improving the quality of academic publications and in predicting the impact of manuscripts through criteria of "excellence" is widely seen as essential to the research endeavour. As a term that first entered critical and popular discourse around 1960 but also as a practice that only became commonplace far later than most suspect, peer review is sometimes described as the "gold standard" of quality control and the majority of researchers consider it crucial to contemporary science.[7] Indeed, peer review is

---

Deep Impact: Unintended Consequences of Journal Rank, in: *Frontiers in Human Neuroscience*, 7 (2013), 1-12, doi:10.3389/fnhum.2013.00291.

4 A. J. Wakefield et al., RETRACTED: Ileal-Lymphoid-Nodular Hyperplasia, Non-Specific Colitis, and Pervasive Developmental Disorder in Children, in: *The Lancet*, 351 (9103/1998), 637-641, doi:10.1016/S0140-6736(97)11096-0.

5 Fiona Godlee/Jane Smith/Harvey Marcovitch, Wakefield's Article Linking MMR Vaccine and Autism Was Fraudulent, in: *BMJ*, 342 (2011), c7452, doi:10.1136/bmj.c7452.

6 Franz J. Ingelfinger, Peer Review in Biomedical Publication, in: *The American Journal of Medicine*, 56 (1974), 686-692; Hans-Dieter Daniel, *Guardians of Science: Fairness and Reliability of Peer Review*, Weinheim 1993, see 4; Peter M. Rothwell/Christopher N. Martyn, Reproducibility of Peer Review in Clinical Neuroscience, in: *Brain*, 123 (9/2000), 1964-1969, doi:10.1093/brain/123.9.1964; Richard Smith, Peer Review: A Flawed Process at the Heart of Science and Journals, in: *Journal of the Royal Society of Medicine*, 99 (4/2006), 178-182; Richard Smith, Classical Peer Review: An Empty Gun, in: *Breast Cancer Research*, 12 (4/2010), S13, doi:10.1186/bcr2742.

7 Melinda Baldwin, In Referees We Trust?, in: *Physics Today*, 70 (2/2017), 44-49, doi:10.1063/PT.3.3463; Melinda Baldwin, Scientific Autonomy, Public Accountability, and the Rise of "Peer Re-

much younger than many suspect. In 1936, for instance, Albert Einstein was outraged to learn that his unpublished submission to Physical Review had been sent out for review.[8] Yet, despite its relative youth, peer review has nonetheless become a fixture of academic publication. This raises the question, though, of why this might be the case. For surprisingly little evidence exists to support the claim that peer review is the best way to pre-audit work, leading Michelle Lamont and others to note the importance of ensuring that "peer review processes [… are] themselves subject to further evaluation."[9]

Research into peer review processes, however, can be difficult to conduct. Nevertheless, this has not prevented a burgeoning field from emerging around the topic.[10] Certainly, following the influential work of John Swales, there has been an

---

view" in the Cold War United States, in: *Isis*, 109 (3/2018), 538-558, doi:10.1086/700070; Irene Hames, *Peer Review and Manuscript Management of Scientific Journals Guidelines for Good Practice*, Malden, MA, 2007, see 2; Bruce Alberts/Brooks Hanson/Katrina L. Kelner, Reviewing Peer Review, in: *Science*, 321 (5885/2008), 15, doi:10.1126/science.1162115; Samuel Moore et al., Excellence R Us: University Research and the Fetishisation of Excellence, in: *Palgrave Communications*, 3 (2017), doi:10.1057/palcomms.2016.105; Adrian Mulligan/Louise Hall/Ellen Raphael, Peer Review in a Changing World: An International Study Measuring the Attitudes of Researchers, in: *Journal of the American Society for Information Science and Technology*, 64 (1/2013), 132-161, doi:10.1002/asi.22798; Aileen Fyfe et al., Managing the Growth of Peer Review at the Royal Society Journals, 1865-1965, in: *Science, Technology, & Human Values* 45 (3/2019), 405-429, doi:10.1177/0162243919862868; David Shatz, *Peer Review: A Critical Inquiry*, Issues in Academic Ethics, Lanham, MD, 2004, see 1.

8 Baldwin, Scientific Autonomy, 542.

9 Michèle Lamont, *How Professors Think: Inside the Curious World of Academic Judgment*, Cambridge, MA, 2009, see 247. See also Marcel C. LaFollette, *Stealing into Print: Fraud, Plagiarism, and Misconduct in Scientific Publishing*, Berkeley, CA, 1992.

10 For just a selection, see Vladimir Batagelj/Anuška Ferligoj/Flaminio Squazzoni, The Emergence of a Field: A Network Analysis of Research on Peer Review, in: *Scientometrics*, 113 (1/2017), 503-532, doi:10.1007/s11192-017-2522-8; Jonathan Tennant/Tony Ross-Hellauer, The Limitations to Our Understanding of Peer Review, in: *SocArXiv*, 2019, doi:10.31235/osf.io/jq623; Erwin O. Smigel/H. Laurence Ross, Factors in the Editorial Decision, in: *The American Sociologist*, 5 (1/1970), 19-21; Charles M Bonjean/Jan Hullum, Reasons for Journal Rejection: An Analysis of 600 Manuscripts, in: *PS*, 11 (1978), 480-483; Elizabeth Ehrhardt Mustaine/Richard Tewksbury, Reviewers' Views on Reviewing: An Examination of the Peer Review Process in Criminal Justice, in: *Journal of Criminal Justice Education*, 19 (3/2008), 351-365, doi:10.1080/10511250802476178; Richard Tewksbury/ Elizabeth Ehrhardt Mustaine, Cracking Open the Black Box of the Manuscript Review Process: A Look Inside, in: *Justice Quarterly*, *Journal of Criminal Justice Education*, 23 (4/2012), 399-422, doi:10.1080/10511253.2011.653650; Omar Sabaj Meruane/Carlos González Vergara/Álvaro Pina-Stranger, What We Still Don't Know About Peer Review, in: *Journal of Scholarly Publishing*, 47 (2/2016), 180-212, doi:10.3138/jsp.47.2.180; Francisco Grimaldo/Ana Marušić/Flaminio Squazzoni, Fragments of Peer Review: A Quantitative Analysis of the Literature (1969-2015), in: *PLOS ONE*, 13 (2/2018), e0193148, doi:10.1371/journal.pone.0193148; Francisco Grimaldo/Mario Paolucci/Jordi Sabater-Mir, Reputation or Peer

ever-increasing number of studies that examine the language and mood of published academic articles, grant proposals, and editorials.[11] This is not surprising.

Review? The Role of Outliers, in: *Scientometrics*, 116 (3/2018), 1421-1438, doi:10.1007/s11192-018-2826-3; Ann C. Weller, Editorial Peer Review: Its Strengths and Weaknesses, in: *Journal of the Medical Library Association*, 90 (1/2002); Lutz Bornmann, Peer Review and Bibliometrics: Potentials and Problems, in: Jung Cheol Shin/Robert K. Toutkoushian/Ulrich Teichler (eds.), *University Rankings: Theoretical Basis, Methodology and Impacts on Global Higher Education*, Dordrecht, 2011, 145-164, doi:10.1007/978-94-007-1116-7; Nyssa J. Silbiger/Amber D. Stubler, Unprofessional Peer Reviews Disproportionately Harm Underrepresented Groups in STEM, in: *PeerJ*, 7 (2019), e8247, doi:10.7717/peerj.8247; Flaminio Squazzoni, Peering Into Peer Review, in: *Sociologica*, 3 (2010), 1-27, doi:10.2383/33640; Cassidy R. Sugimoto and Blaise Cronin, Citation Gamesmanship: Testing for Evidence of Ego Bias in Peer Review, in: *Scientometrics*, 95 (3/2013), 851-862, doi:10.1007/s11192-012-0845-z; Steven N. Goodman, Manuscript Quality before and after Peer Review and Editing at Annals of Internal Medicine, in: *Annals of Internal Medicine*, 121 (1/1994), 11, doi:10.7326/0003-4819-121-1-199407010-00003; Jean-Pierre E.N. Pierie/Henk C. Walvoort/A. John P.M. Overbeke, Readers' Evaluation of Effect of Peer Review and Editing on Quality of Articles in the Nederlands Tijdschrift Voor Geneeskunde, in: *The Lancet*, 348 (9040/1996), 1480-1483, doi:10.1016/S0140-6736(96)05016-7; Michael J. Mahoney, Publication Prejudices: An Experimental Study of Confirmatory Bias in the Peer Review System, in: *Cognitive Therapy and Research*, 1 (2/1977), 161-175, doi:10.1007/BF01173636; Richard L. Kravitz et al., Editorial Peer Reviewers' Recommendations at a General Medical Journal: Are They Reliable and Do Editors Care?, in: *PLOS ONE*, 5 (4/2010), e10072, doi:10.1371/journal.pone.0010072; Daniel M. Herron, Is Expert Peer Review Obsolete? A Model Suggests That Post-Publication Reader Review May Exceed the Accuracy of Traditional Peer Review, in: *Surgical Endoscopy*, 26 (8/2012), 2275-2280, doi:10.1007/s00464-012-2171-1; Ferric C. Fang/Anthony Bowen/Arturo Casadevall, NIH Peer Review Percentile Scores Are Poorly Predictive of Grant Productivity, in: *ELife*, 5 (2016), e13323, doi:10.7554/eLife.13323; Amber E. Budden et al., Double-Blind Review Favours Increased Representation of Female Authors, in: *Trends in Ecology & Evolution*, 23 (1/2008), 4-6, doi:10.1016/j.tree.2007.07.008; Margaret E. Lloyd, Gender Factors in Reviewer Recommendations for Manuscript Publication, in: *Journal of Applied Behavior Analysis*, 23 (4/1990), 539-543, doi:10.1901/jaba.1990.23-539; Tom Tregenza, Gender Bias in the Refereeing Process?, in: *Trends in Ecology & Evolution*, 17 (8/2002), 349-350, doi:10.1016/S0169-5347(02)02545-4; E. Ernst/T. Kienbacher, Chauvinism, in: *Nature*, 15.08.1991, 560, doi:10.1038/352560b0; Ann M. Link, US and Non-US Submissions: An Analysis of Reviewer Bias, in: *JAMA*, 280 (3/1998), 246-247, doi:10.1001/jama.280.3.246; Paolo Dall'Aglio, Peer Review and Journal Models, in: *ArXiv*, 2006, URL: http://arxiv.org/abs/physics/0608307 [last accessed: Mar. 31, 2021]; Gilbert W. Gillespie/Daryl E. Chubin/George M. Kurzon, Experience with NIH Peer Review: Researchers' Cynicism and Desire for Change, in: *Science, Technology, & Human Values*, 10 (3/1985), 44-54, doi:10.1177/016224398501000306; Stephen J. Ceci/Douglas P. Peters, Peer Review: A Study of Reliability, in: *Change*, 14 (6/1982), 44-48; Douglas P. Peters/Stephen J. Ceci, Peer-Review Practices of Psychological Journals: The Fate of Published Articles, Submitted Again, in: *Behavioral and Brain Sciences*, 5 (2/1982), 187-195, doi:10.1017/S0140525X00011183; Blaise Cronin, Vernacular and Vehicular Language, in: *Journal of the American Society for Information Science and Technology*, 60 (3/2009), 433-433, doi:10.1002/asi.21010; Joseph S. Ross et al., Effect of Blinded Peer Review on Abstract Acceptance, in: *JAMA*, 295 (14/2006),

After all, as Peter van den Besselaar, Hélène Schiffbaenker, Ulf Sandström, and Charlie Mom note, "[l]anguage embodies normative views about who/where we communicate about, and stereotypes about others are embedded and reproduced

1675-1680, doi:10.1001/jama.295.14.1675; G. D. L. Travis/H. M. Collins, New Light on Old Boys: Cognitive and Institutional Particularism in the Peer Review System, in: *Science, Technology, & Human Values*, 16 (3/1991), 322-341, doi:10.1177/016224399101600303; Daryl E. Chubin/Edward J. Hackett, *Peerless Science: Peer Review and U.S. Science Policy*, Albany, NY, 1990; Mahoney, Publication Prejudices; A. H. Bardy, Bias in Reporting Clinical Trials, in: *British Journal of Clinical Pharmacology*, 46 (2/1998), 147-150, doi:10.1046/j.1365-2125.1998.00759.x; Kay Dickersin et al., Publication Bias and Clinical Trials, in: *Controlled Clinical Trials*, 8 (4/1987), 343-353, doi:10.1016/0197-2456(87)90155-3; Kay Dickersin/Yuan-I. Min/Curtis L. Meinert, Factors Influencing Publication of Research Results: Follow-up of Applications Submitted to Two Institutional Review Boards, in: *JAMA*, 267 (3/1992), 374-378, doi:10.1001/jama.1992.03480030052036; Daniele Fanelli, Do Pressures to Publish Increase Scientists' Bias? An Empirical Support from US States Data, in: *PLoS ONE*, 5 (4/2010), doi:10.1371/journal.pone.0010271; John P. A. Ioannidis, Effect of the Statistical Significance of Results on the Time to Completion and Publication of Randomized Efficacy Trials, in: *JAMA*, 279 (4/1998), 281-286, doi:10.1001/jama.279.4.281; Dalmeet Singh Chawla, Thousands of Grant Peer Reviewers Share Concerns in Global Survey, in: *Nature*, 15.10.2019, doi:10.1038/d41586-019-03105-2; Tony Ross-Hellauer, What Is Open Peer Review? A Systematic Review, in: *F1000Research*, 6 (2017), 588, doi:10.12688/f1000research.11369.2; Kanu Okike et al., Single-Blind vs Double-Blind Peer Review in the Setting of Author Prestige, in: *JAMA*, 31 (12/2016), 1315-1316, doi:10.1001/jama.2016.11014; Stanley Fish, No Bias, No Merit: The Case against Blind Submission, in: *PMLA*, 103 (5/1988), 739-748; Dakota Murray et al., Author-Reviewer Homophily in Peer Review, in: *BioRxiv* 400515 (2019), doi:10.1101/400515.

11 John Swales, *Genre Analysis: English in Academic and Research Settings*, Cambridge 1990; Rahime Nur Aktas/Viviana Cortes, Shell Nouns as Cohesive Devices in Published and ESL Student Writing, in: *Journal of English for Academic Purposes*, 7 (1/2008), 3-14, doi:10.1016/j.jeap.2008.02.002; Nigel Harwood, "I Hoped to Counteract the Memory Problem, but I Made No Impact Whatsoever": Discussing Methods in Computing Science Using I, in: *English for Specific Purposes*, 24 (3/2005), 243-267, doi:10.1016/j.esp.2004.10.002; Nigel Harwood, "Nowhere Has Anyone Attempted … In This Article I Aim to Do Just That": A Corpus-Based Study of Self-Promotional I and We in Academic Writing across Four Disciplines, in: *Journal of Pragmatics*, Focus-on Issue: Marking Discourse, 37 (8/2005), 1207-1231, doi:10.1016/j.pragma.2005.01.012; W. Shehzad, How to End an Introduction in a Computer Science Article: A Corpus-Based Approach, in E. Fitzpatrick (ed.): *Corpus Linguistics Beyond the Word: Corpus Research from Phrase to Discourse*, Amsterdam 2015, 227-241, URL: https://brill.com/view/title/30110 [last accessed: Mar. 31, 2021]; Ulla Connor/Anna Mauranen, Linguistic Analysis of Grant Proposals: European Union Research Grants, in: *English for Specific Purposes*, 18 (1/1999), 47-62, doi:10.1016/S0889-4906(97)00026-4; Davide Simone Giannoni, Medical Writing at the Periphery: The Case of Italian Journal Editorials, in: *Journal of English for Academic Purposes*, 7 (2/2008), 97-107, doi:10.1016/j.jeap.2008.03.003. These examples are drawn from Theresa Lillis/Mary Jane Curry, The Politics of English, Language and Uptake: The Case of International Academic Journal Article Reviews, in: *AILA Review*, 28 (2015), 127-150, doi:10.1075/aila.28.06lil.

in language."[12] Indeed, a number of existing studies have examined the linguistic properties of peer review reports written by the authors themselves.[13]

2. Scaling our Understanding of Peer Review Using Neural Networks

As part of our Andrew W. Mellon Foundation-funded project, "Reading Peer Review," we were granted access to the archive of peer review reports at the world's largest, trans-disciplinary scientific journal, *PLOS ONE*.[14] This title has a radical policy on peer review that is different to many other journals. As Catriona J. MacCallum put it, "[t]he basis for […] decisions" in conventional pre-publication peer review "is inevitably subjective. The higher-profile science journals are consequently often accused of 'lottery reviewing,' a charge now aimed increasingly at the more specialist literature as well. Even after review, papers that are technically sound are often rejected on the basis of lack of novelty or advance."[15]

PLOS wanted to work differently. The peer-review procedure at *PLOS ONE* is predicated on the idea of "technical soundness" in which papers are judged according to whether or not their methods and procedures are thought to be solid, rather than on the basis of whether their contents are judged to be important.[16] This is a model in which reviewers should not accept or reject a paper for its novelty or significance, but should only assess its scientific validity. Hence, *PLOS ONE* will accept replication studies, null results (where an experiment didn't work), and other forms of scientific output that might not be published elsewhere. Thus, *PLOS ONE* was designed to "initiate a radical departure from the stifling constraints of

12 Peter van den Besselaar et al., Explaining Gender Bias in ERC Grant Selection – Life Sciences Case, in: *STI 2018 Conference Proceedings*, Leiden University, 2018, 314. See also Christian Burgers/Camiel J. Beukeboom, Stereotype Transmission and Maintenance Through Interpersonal Communication: The Irony Bias, in: *Communication Research*, 43 (3/2016), 414-441, doi:10.1177/0093650214534975; Camiel J. Beukeboom/Christian Burgers, Linguistic Bias, in: *Oxford Research Encyclopedia of Communication*, 2017, doi:10.1093/acrefore/9780190228613.013.439.

13 Peter Woods, *Successful Writing for Qualitative Researchers*, London 2006, 140-146; David Coniam, Exploring Reviewer Reactions to Manuscripts Submitted to Academic Journals, in: *System*, 40 (4/2012), 544-553, doi:10.1016/j.system.2012.10.002; Brian Paltridge, *The Discourse of Peer Review: Reviewing Submissions to Academic Journals*, London 2017, 49-50.

14 The database was supplied to us electronically for offsite use and storage.

15 Catriona J. MacCallum, ONE for All: The Next Step for PLoS, in: *PLOS Biology*, 4 (11/2006), e401, doi:10.1371/journal.pbio.0040401.

16 PLOS, Journal Information, in: *PLOS ONE*, 2016, URL: http://www.plosone.org/static/information [last accessed: Mar. 31, 2021]. Note though that even this definition is contentious. See Richard Poynder, *PLoS ONE, Open Access, and the Future of Scholarly Publishing*, 2011, URL: https://richardpoynder.co.uk/PLoS_ONE.pdf [last accessed: Mar. 31, 2021].

this existing system." In this new model, it was claimed, "acceptance to publication [would] be a matter of days."[17]

We were provided by PLOS with a database consisting of 229,296 usable peer-review reports written between 2014 and 2016 from *PLOS ONE*. There were other reports in this database, but the identifiers assigned to them made it impossible to group these reports by review round and so these data were discarded. We wanted to know: how have the radical propositions that led to the creation of *PLOS ONE* affected actual practices on the ground in the title? Do PLOS reviewers behave as one might expect given the radicalism on which *PLOS ONE* was premised? And what can we learn about organisational change and its drivers? These broader questions are addressed in the book that came out of the project.[18]

In order to understand the composition of the archive and to communicate these findings in a way that does not cite any material directly, for reasons of data protection, we undertook a qualitative coding exercise (specifically domain and taxonomic descriptive coding) in which three research assistants collaboratively built a taxonomy of statements derived from the longer reviews.[19] In order to achieve intersubjective and, as far as possible, some intercultural linguistic assessment of the database, we had a diverse team of coders. Two of the research assistants were native English speakers based in London in the United Kingdom, although we note that the policed boundary of "native" and "non-native" speakers comes with both challenges for the specific study of peer review, but also with postcolonial overtones.[20] The third research assistant was an L2 English speaker (English as a second language) based in Lethbridge in Canada with significant social scientific background experience, including with this kind of coding work.

The goal of our coding exercise was to delve into the linguistic structures and semantic comment types that are used by reviewers, following previous work by Fortanet.[21] In order to militate against identity subjectivities in the coding process, each report was coded in triplicate – in which each research assistant worked at first individually but then regrouped to build collaborative consensus among the group on both sentiment and thematic classification – thereby constructing an

17 MacCallum, ONE for All: The Next Step for PLoS.

18 Eve et al.,*Peer Review and Institutional Change in Academia* .

19 David W. McCurdy/James P. Spradley/Dianna J. Shandy, *The Cultural Experience: Ethnography in Complex Society*, Long Grove, IL, 2005, 44-45.

20 251Karent Englander, Revision of Scientific Manuscripts by Non-Native English-Speaking Scientists in Response to Journal Editors' Language Critiques, in: *Journal of Applied Linguistics and Professional Practice*, 3 (2/2006), 129-161, doi:10.1558/japl.v3i2.129.

21 Inmaculada Fortanet, Evaluative Language in Peer Review Referee Reports, in: *Journal of English for Academic Purposes*, 7 (1/2008), 27-37, doi:10.1016/j.jeap.2008.02.004.

intersubjective agreement on the labels assigned for each term.[22] The downside of this approach is that, clearly, we traded accuracy for volume. This resulted in 78 triplicate tagged reports, consisting of 2,049 statements. Given the constraints on our resources, we hoped nonetheless that we could use the coded statements as a training resource for a neural network text classifier, to extrapolate up our claims about the archive.

Our coding exercise eventually built the following taxonomy of peer-review statements:

Table 1: The taxonomy of statements built for the Reading Peer Review project from the PLOS ONE database.

High-Level Category	Fine-Grained Category	Explication
Data	Data	A reference to results and/or data.
Data	Data commentary	A description of or commentary upon data. For instance, a reference to a chart's legend.
Data	Interpretation	Extrapolation from data. This category can overlap with data analysis/treatment.
Data	Analysis/treatment	How data are treated after collection. This includes data analysis and statistical analysis. It can also refer to secondary data (sets).
Data	Presentation	Includes reference to data display. Also includes comments on formatting, size of tables, redundancy of images, visibility of images, and size of the images.
Field of Knowledge	(Knowledge) Statement	A statement that the reviewer makes (about fact or community agreed notions). Does not apply to the reviewer paraphrasing the original article. Relates to knowledge claims by the reviewer and/or authors.

22 Such a triplicate coding approach had been used previously by Lutz Bornmann/Christophe Weymuth/Hans-Dieter Daniel, A Content Analysis of Referees' Comments: How Do Comments on Manuscripts Rejected by a High-Impact Journal and Later Published in Either a Low- or High-Impact Journal Differ?, in: *Scientometrics*, 83 (2/2010), 493-506, doi:10.1007/s11192-009-0011-4.

Field of Knowledge	Information for author(s)	Statements that indicate a reviewer's subjective opinion. E.g., "I consider it appropriate to…"
Field of Knowledge	Positioning	Reference to ways in which/ to what extent the authors position concepts/ideas in relation to others. Can also imply/require the Literature tag (see below).
Field of Knowledge	Literature	Explicit reference to secondary literature. Negative sentiment score in this category refers to misinterpretation or misrepresentation of literature, or lack of relevance of references employed.
Field of Knowledge	Revision	A comment on whether revisions have been made. A positive sentiment score in this category indicates revisions met while a negative means the opposite. This category also includes corrections and reference to subsequent/previous revisions.
Field of Knowledge	Holistic revision	Reviewer signals a range of issues to be fixed through revisions (referring to multiple categories).
Field of Knowledge	Fallibility	Instances where the reviewer admits they may not be correct in their opinion/criticism or admits inadequacy of and uncertainty around judgement.
Field of Knowledge	Tone	Tone of reviewer exhibits bias against non-western submission/language (patronising). Tone of reviewer exhibits *ad hominem* attack on author or team of researchers. Also used to denote overly familiar personal register/tone. Awarded appropriate sentiment score if tone implies praise or critique of manuscript.
Field of Knowledge	Potential/significance	A remark upon the significance of findings/data/results/work. This also includes the potential of contribution to knowledge or research; references to reproduction of experiments. Also used to flag poor scholarship and auto-plagiarism via a lack of novelty. Note that this category of "significance" should *not* be a criterion used for judgement of admission within the *PLOS ONE* ecosystem.

Expression	(English) Language	Reference to use of English, languages other than English, native/non-native speakers.
Expression	Typographical errors	Reference to surface level errors, including grammatical errors. Lack of consistency denotes strongly negative sentiment. Trivial typos are low sentiment score. Comments on punctuation are attributed using this tag.
Expression	Expression	Communicative quality - coherence of style and academic/scientific register. Choice of language. Rewording. Definitions of terms/acronyms.
Expression	Terminology	Use/deployment of subject-specific terminology. Can refer to accessibility of terms.
Expression	Cohesion	Comments on linkage between sections of paper in terms of correlation, structure and organisation.
Expression	Style	Comments on adhesion to house style.
Expression	Citation	Referencing and citation practice; includes lack of appropriate citation.
Expression	Summary	When a reviewer summarises or signals a section of paper. Also used as a form of transition before critique. Includes quotations from original text, including title.
Expression	Transition	A transitive statement which makes no reference to the manuscript. Includes notes to editors.
Methodology	Methodology	Broader approach to methods adopted. Also refers to rationale, justification or basis for research. Ethical issues/concerns.
Methodology	Statistics	In general and/or explicit reference to statistics including statistical tests. Explicit reference to or use of statistical tests such as Analysis of Variance (ANOVA), Student's T-Test, Pearson, correlation coefficients, Mann Whitney (package).
Methodology	Experimental design	Reference to a series of experiments, hypotheses, sample size, control groups, parameters, data collection tools, inferential/descriptive statistics, correlation, data modeling.

Methodology	Method	Refers to the description of method, including procedures, techniques, and discussion of advantageous alternatives.
Methodology	Limitations	Discussion of limitations.
Omission	Implied omission	Implies that something is missing without explicitly stating it.
Omission	Omission	Explicitly states that something is missing.
Omission	Accuracy	Comments on the accuracy of (data) description (& definitions). Can refer to factual or descriptive inaccuracy. Can also refer to (lack of) precision.
Omission	Elaboration	Request for more detail, information, clarification or precision. Different to omission in the sense that omission is about something that isn't there at all whereas this tag calls for supplementation.
Omission	Argument/analysis	Discussion of data/results. When there is "omission," it is unlikely that this tag will be used also.
Omission	Ambiguity	Reference to clarity, vagueness. Can connote positive (as clear, well worded etc.) as well as negative sentiment. Instances where something not clear to reviewer.
Omission	Argument	Pertains to clarity of argument - exposing point of view. Distinct from "argument/analysis" in that it deals with literature. Can also refer to the phrasing of an argument. Negative sentiment can refer to redundant or unconvincing argument. Explicit reference to logic or logical can imply this category. Also refers to the coherence of an argument. Claims implying criticism/agreement.
Omission	Implied criticism	Used for tagging questions from reviewers. Negative meaning/critique implicit. For instance, "Would this manuscript benefit from X?"
Section	Outcome	Publishability and suitability of results/data/findings. Relates to publishability of specific paper. Usually with reference to the admissibility criteria of *PLOS ONE*.

Section	Overarching comment	Used for tagging comments that broadly apply to the whole manuscript.
Section	Conclusion	Reference to the results of interpretation and/or analysis. Can also refer to results/findings. Reference to implications of results. Also refers to limitations of study.
Section	Abstract	Reference to the work's abstract.
Section	Appendix	Reference to an appendix in a work.

In order to conduct our computational reading test, we built a multi-class and multi-label text classifier based on the TenCent NeuralClassifier toolkit.[23] Although multi-class and multi-label text classification is a difficult task and even though we were only possessed of a relatively minimal, albeit robust, training set, the neural network was good at classifying certain types of input text. In particular, the network performed well at recognising requests for revision and/or outcome statements. For example, the generic statement "I do not recommend publication" was tagged by the network as pertaining to "revision" and "outcome." Some other types of broad statements were also accurately classified: "In particular I am left confused as to how the results fit in here" was marked as "ambiguity" and "cohesion" by the software.

However, the specific challenges of implementing an accurate classification system were many. First, the tagged data proved insufficient for these purposes. The labour-intensive processes of triplicate tagging gave us the confidence that we needed in the material that had been tagged, but this came at the expense of volume. Further, since each tagged statement was relatively short it was difficult to train natural-language processing toolkits to identify salient features; there is not

23 *An Open-Source Neural Hierarchical Multi-Label Text Classification Toolkit: Tencent/NeuralNLP-NeuralClassifier*, 2019, URL: https://github.com/Tencent/NeuralNLP-NeuralClassifier [last accessed: Mar. 31, 2021]. For more on paradigms of so-called "distant reading," see Franco Moretti, *Distant Reading*, London 2013; Franco Moretti, *Graphs, Maps, Trees: Abstract Models for Literary History*, London 2007; Ted Underwood, A Genealogy of Distant Reading, in: *Digital Humanities Quarterly*, 11 (2/2017), URL: http://www.digitalhumanities.org/dhq/vol/11/2/000317/000317.html [last accessed: Mar. 31, 2021]; Andrew Piper, *Enumerations: Data and Literary Study*, Chicago, IL, 2018; Nan Z. Da, The Computational Case against Computational Literary Studies, in: *Critical Inquiry*, 45 (3/2019), 601-639, doi:10.1086/702594; Martin Paul Eve, *Close Reading With Computers: Textual Scholarship, Computational Formalism, and David Mitchell's* Cloud Atlas, Stanford, CA, 2019; Ted Underwood, *Distant Horizons: Digital Evidence and Literary Change*, Chicago, IL, 2019.

a huge volume in each case for the network to identify. As above, there are also instances where we did not find and tag particular types of statement, such as those pertaining to ethics. Finally, since each statement was written by different authors (reviewers), with different primary languages, the strength of these linguistic differentiations – as opposed to the words used within different types of classificatory statements – appear to be pulled to the fore.[24] As such, this study is limited to a relatively small sample size with a relatively good accuracy level within that sample.

Hence, while the network appears to work well at classifying statements that have appeared in almost all reviews – for instance, the outcome example above – it performed poorly at identifying less frequent types, such as "fallibility." The network was unable, for example, to ascribe a label to the statement "I must confess that I am not an expert with respect to these methods," a clear assertion of fallibility. Further, various statements around originality were not tagged with any accuracy. For instance, "There is nothing technically wrong with the paper, but it is not that original" was marked as an "overarching comment," which is a fair assessment. However, no label noting that this was a statement about originality or novelty was ascribed, regardless of the training parameters that we fed to the network.

A further method for "distant reading" the corpus is, of course, to conduct a simple text search through the reviews. This is how we identified the "missing" statements on ethics to which we earlier referred. This can be useful to find examples of specific kinds of practice. For instance, to identify overly aggressive reports we used a simple tool, "grep" (globally search a regular expression and print), to look for instances of the word "useless" in the top-800 longest reports. This yielded harsh reports that included phrases such as "Fig 12 is almost useless"; "the null model seems somewhat useless"; "remove the repeated useless sentences"; "I found the [secondary subject matter] results to be EXTREMELY distracting, and essentially useless"; "this work appears to be all but useless" and so on. What such searching cannot tell us, though, is the prevalence of such practices. For instance, the above examples were found using an extremely simple keyword search pulled from the top of our heads. There will be many instances of *ad hominem* or vicious attack that use different terms and the only reliable way, at present, to identify these is to read and to tag the reports themselves.

In addition to this, capital letters (as in the above "EXTREMELY" example) are relatively easy to detect and sometimes indicate strong sentiment of one kind or another. However, detection of these is not as simple as a regular expression ("\b[A-Z][A-Z]+\b") as this will also pull out the many acronyms used in scientific practice

24 For more on authorship signals among others, see Sarah Allison et al., *Quantitative Formalism: An Experiment*, Stanford 2011, URL: https://litlab.stanford.edu/LiteraryLabPamphlet1.pdf [last accessed: Mar. 31, 2021]; Matthew L. Jockers, *Macroanalysis: Digital Methods and Literary History*, Urbana, IL, 2013.

("BDNF," "ROC" etc.). It is also not clear that capital letters denote strong sentiment in one direction or another; "WOW" can indicate "WOW this is a brilliant paper," but it can equally likely specify "WOW, this paper was terrible." Furthermore, on occasion capital letters are used to denote section headings and/or specific portions of a paper ("in the METHOD section"). In this way, the extraction of capital letters – without a pre-built blacklist of words to exclude – is likely to result in many false positives.

A further way of exploring the corpus at scale is to use the techniques of "topic modeling," a technique that finds co-occurring words and bundles them into so-called "topics." Hence, a "topic" in a recipe book might be: "sugar," "icing," "sweet," "jelly." Topic modeling generally uses a process called "Latent Dirichlet Allocation" (LDA) in order to cluster together terms that probabilistically co-occur in similar contexts. This is a useful way to explore a dataset and to infer the groups of terms that most frequently crop up together; that is, which "topics" are explored within a corpus (a three-level hierarchical Bayesian model, in which each item of a collection is modeled as a finite mixture over an underlying set of topics).[25] As Ben Schmidt notes, this approach "does a good job giving an overview of the contents of large textual collections; it can provide some intriguing new artifacts to study; and it even holds […] some promise for structuring non-lexical data like geographic points."[26]

However, LDA is also a dangerous method. This is because there is no way to infer *why* topics have been grouped together. In particular, surprising groupings that appear to exhibit coherence may not be as well bound as we would like to think. As Schmidt continues,

> still, excitement about the use of topic models for discovery needs to be tempered with skepticism about how often the unexpected juxtapositions LDA creates will be helpful, and how often merely surprising. A poorly supervised machine learning algorithm is like a bad research assistant. It might produce some unexpected constellations that show flickers of deeper truths; but it will also produce tedious, inexplicable, or misleading results.[27]

That said, as an exploratory exercise that others may wish to take further, we produced a twenty-topic model using the MALLET tool based on the same corpus of 800 reports using the default hyperparameters and with stop words excluded. The results for this are shown in Table 2.

---

25 David M. Blei/Andrew Y. Ng/Michael I. Jordan, Latent Dirichlet Allocation, in: *Journal of Machine Learning Research*, 3 (2003), 993-1022.

26 Benjamin Schmidt, Words Alone: Dismantling Topic Models in the Humanities, in: *Journal of Digital Humanities*, 2 (1/2012), URL: http://journalofdigitalhumanities.org/2-1/words-alone-by-benjamin-m-schmidt/ [last accessed: April 1, 2021].

27 Schmidt, Words Alone: Dismantling Topic Models in the Humanities.

Table 2: A topic model of the 800 longest reports in our database of reviews at PLOS ONE.

Topic Number	Topic Terms
1	authors manuscript paper study comments data review results current previous discussion work addressed major provide reviewer research information important studies
2	interests competing samples genetic dna populations population gene pcr table strains sequences figure structure loci sequencing individuals analysis chromosomes number
3	paper data case make point time clear important find fact understand general evidence e.g i.e model number high approach literature
4	line lines page sentence paragraph change suggest section figure results text table discussion reference manuscript remove delete add replace information
5	genes gene expression analysis number sequences sequence genome rna authors expressed species biological fig methods transcripts data results transcriptome proteins
6	model method models paper approach parameters system distribution set network number dataset parameter distributions author proposed performance simulations equation networks
7	species study habitat lines area model spatial population areas line fish variables sites models distance size individuals data prey year
8	study treatment patients group participants trial intervention studies outcome pain analysis groups clinical outcomes research control care reported patient measures
9	patients study blood clinical studies disease plasma levels authors activity acute group tissue serum negative ace sensitivity mir cortisol healthy
10	social behavior females males male authors individuals scer_scrt lcl study female group behaviors human sex calls behaviour pointing sexual attention
11	fire area page e.g subject trees specific stand bone signals motion intensity study science frequency subjects biochemistry atmospheric stands fires

12	food hsv animals infection mice diet authors bees response dose resistance treatment weight pigs bacteria immune intake group larvae virus
13	participants task authors condition experiment effect stimuli results memory performance responses effects experiments stimulation response conditions experimental visual trials learning
14	cells authors cell figure fig expression data shown protein experiments levels control show mrna mice state manuscript results effect antibody
15	species phylogenetic taxa tree xxx based diversity sequences analysis clade character genus specimens support trees present phylogeny group taxonomic found
16	age health risk table page women population study prevalence model factors children results years variables cases analysis paragraph hiv year
17	null partly disease n/a cancer vaccine hpv page cervical women vaccination safety doi group gardasil adverse rate don't map human
19	data results authors analysis table methods study differences significant discussion statistical figure time section values effect test information sample size
19	species water soil temperature change growth plant plants concentrations climate samples concentration biomass sites fish site study conditions carbon community
20	fig protein binding manuscript light proteins figure structure shown images domain mutant site sequence image residues cry region structures pax

Some of these topics appear easy to interpret. Group one, comprising "authors," "manuscript," "paper" and so on cluster meta-statements about the paper, its submission, and the review process. It is curious, though, that "important" should find its way into the work here (although "not important" would also trigger this, so no sentiment value should be inferred). Certainly, there are multiple contexts within which the word "important" can appear. For instance, "it is important that the author address these points" is as likely a statement as "this paper is extremely important." Nonetheless, given that *PLOS ONE* specifically disavows importance from its criteria, it is significant that the term should appear so prominently among statements that are otherwise common in opening gambits.

Topic four, by contrast, clearly pertains to the mechanics of a paper and suggested corrections. Its functional emphasis on the "line," "sentence," "page," "figure," "table" and so forth – coupled with "suggest," "add," "replace," and "delete" is the archetypical set of terms that we find in revision requests. In our experience of tagging, such language is prevalent during line-by-line commentaries that usually take the form of "line 123: suggest adding X."

Several of the topics relate to subject matter that is clearly of disciplinary interest to and prominent within *PLOS ONE*. Topics two and five, for instance, are

concerned with genetics. Topic seven appears to be biology; topics eight and nine circle around medicine and clinical trials; topic ten relates to reproduction, mating, and sexuality; topic twelve seems to indicate dietary behaviours; topic fifteen is about biological taxonomies; topic sixteen is on ageing; and so on.

Of course, anyone who knows anything about *PLOS ONE* might have guessed that such terms would cluster together and be found as separate strata. For us, the more useful indicators are not the subject groupings, which one would expect, but the functional parameters. We can anticipate scenarios under which knowledge of the distinct linguistic layer of line-by-line corrections, for instance, could be extracted and formed into editorial "to-do" lists. We could also imagine automatic detection of appraisal of novelty and importance, and a flagging system that could warn the editor of such an approach (and that it should not be used in the judgement of articles). The challenge, as ever with topic modeling, though, is that the topics that seem clearly thematically clustered are obvious, while the ones that exhibit less coherence (say, topic 20) are baffling.

3. Conclusions

The resources required to train a neural network for accurate multi-class and multi-label identification over the whole corpus were greater than those available to us. Indeed, the quality of the classification engine is directly proportional to the volume and accuracy of the training data. While our exercise yielded insights – particularly through LDA and plain-text search methods – to classify accurately the whole corpus and then to make deductive statements with any certainty requires a great deal more work at the corpus preparation stage. In short, while our experiment in using machine learning to examine the entire corpus of reviews might have worked well for certain types of statement, such as those pertaining to outcome, the uncertainty around, and low levels of, accuracy mean that any quantitative analysis based on the broader corpus, read at distance, would be unacceptably imprecise. Nonetheless, the moments of success in the network seem to indicate that those with broader resources for tagging and access to a large corpus of review reports might, in future, see some benefit in using this approach. For instance, we can envisage situations where such a network could detect hostile tone and warn the reviewer that s/he is being overly harsh or *ad hominem*. We could also imagine situations in which such a classifier could distinguish reviews that were structured in an unusual/idiosyncratic manner. While this would not rule out the review from being useful, it could give an indication that the reviewer is inexperienced or working away from norms of the form. That said, if the network were used by publishers to insist on normative practices in review, then this could stifle new ways of writing and operating.

Our experiments in using machine learning to read at scale taught us that, in essence, the results are only ever as good as the data that are fed in. Garbage in? Garbage out. We did not have garbage; our training data were robust, but they were not voluminous. However, to build an ultra-robust and massive corpus that would have made the AI methods work at scale would have required more labour effort than we could afford. This is perhaps the lesson that our AI approach taught us: it is resourcing and people that are the scarcities, not technology.

Bibliography

Aktas, Rahime Nur/Cortes, Viviana, Shell Nouns as Cohesive Devices in Published and ESL Student Writing, in: *Journal of English for Academic Purposes*, 7 (1/2008), 3-14, doi:10.1016/j.jeap.2008.02.002.

Alberts, Bruce/Hanson, Brooks/Kelner, Katrina L., Reviewing Peer Review, in: *Science*, 321 (5885/2008), 15, doi:10.1126/science.1162115.

Allison, Sarah, et al., *Quantitative Formalism: An Experiment*, Stanford 2011, URL: https://litlab.stanford.edu/LiteraryLabPamphlet1.pdf [last accessed: Mar. 31, 2021].

An Open-Source Neural Hierarchical Multi-Label Text Classification Toolkit: Tencent/NeuralNLP-NeuralClassifier, 2019, URL: https://github.com/Tencent/NeuralNLP-NeuralClassifier [last accessed: Mar. 31, 2021].

Bakanic, Von/McPhail, Clark/Simon, Rita J., Mixed Messages: Referees' Comments on the Manuscripts They Review, in: *The Sociological Quarterly*, 30 (4/1989), 639-654, URL: http://www.jstor.org/stable/4121469.

Baldwin, Melinda, Scientific Autonomy, Public Accountability, and the Rise of "Peer Review" in the Cold War United States, in: *Isis*, 109 (3/2018), 538-558, doi:10.1086/700070.

Baldwin, Melinda, In Referees We Trust?, in: *Physics Today*, 70 (2/2017), 44-49, doi:10.1063/PT.3.3463.

Bardy, A. H., Bias in Reporting Clinical Trials, in: *British Journal of Clinical Pharmacology*, 46 (2/1998), 147-150, doi:10.1046/j.1365-2125.1998.00759.x.

Batagelj, Vladimir/Ferligoj, Anuška/Squazzoni, Flaminio, The Emergence of a Field: A Network Analysis of Research on Peer Review, in: *Scientometrics*, 113 (1/2017), 503-532, doi:10.1007/s11192-017-2522-8.

Besselaar, Peter van den, et al., Explaining Gender Bias in ERC Grant Selection – Life Sciences Case, in: *STI 2018 Conference Proceedings*, Leiden University, 2018, 346-352.

Beukeboom, Camiel J./Burgers, Christian, Linguistic Bias, in: *Oxford Research Encyclopedia of Communication*, 2017, doi:10.1093/acrefore/9780190228613.013.439.

Blei, David M./Ng, Andrew Y./Jordan, Michael I., Latent Dirichlet Allocation, in: *Journal of Machine Learning Research*, 3 (2003), 993-1022.

Bonjean, Charles M/Hullum, Jan, Reasons for Journal Rejection: An Analysis of 600 Manuscripts, in: *PS*, 11 (1978), 480-483.

Bornmann, Lutz, Peer Review and Bibliometrics: Potentials and Problems, in: Jung Cheol Shin/Robert K. Toutkoushian/Ulrich Teichler (eds.), *University Rankings: Theoretical Basis, Methodology and Impacts on Global Higher Education*, Dordrecht, 2011, 145-164, doi:10.1007/978-94-007-1116-7.

Bornmann, Lutz, Scientific Peer Review, in: *Annual Review of Information Science and Technology*, 45 (1/2011), 197-245, doi:10.1002/aris.2011.1440450112.

Bornmann, Lutz/Daniel, Hans-Dieter, The Effectiveness of the Peer Review Process: Inter-Referee Agreement and Predictive Validity of Manuscript Refereeing at *Angewandte Chemie*, in: *Angewandte Chemie International Edition*, 47 (38/2008), 7173-7178, doi:10.1002/anie.200800513.

Bornmann, Lutz/Weymuth, Christophe/Daniel, Hans-Dieter, A Content Analysis of Referees' Comments: How Do Comments on Manuscripts Rejected by a High-Impact Journal and Later Published in Either a Low- or High-Impact Journal Differ?, in: *Scientometrics*, 83 (2/2010), 493-506, doi:10.1007/s11192-009-0011-4.

Brainard, Jeffrey/You, Jia, What a Massive Database of Retracted Papers Reveals about Science Publishing's "Death Penalty," in: *Science | AAAS*, 25.10.2018, URL: https://www.sciencemag.org/news/2018/10/what-massive-database-retracted-papers-reveals-about-science-publishing-s-death-penalty [last accessed: Mar. 31, 2021].

Brembs, Björn/Button, Katherine/Munafò, Marcus, Deep Impact: Unintended Consequences of Journal Rank, in: *Frontiers in Human Neuroscience*, 7 (2013), 1-12, doi:10.3389/fnhum.2013.00291.

Budden, Amber E., et al., Double-Blind Review Favours Increased Representation of Female Authors, in: *Trends in Ecology & Evolution*, 23 (1/2008), 4-6, doi:10.1016/j.tree.2007.07.008.

Burgers, Christian/Beukeboom, Camiel J., Stereotype Transmission and Maintenance Through Interpersonal Communication: The Irony Bias, in: *Communication Research*, 43 (3/2016), 414-441, doi:10.1177/0093650214534975.

Ceci, Stephen J./Peters, Douglas P., Peer Review: A Study of Reliability, in: *Change*, 14 (6/1982), 44-48.

Chawla, Dalmeet Singh, Thousands of Grant Peer Reviewers Share Concerns in Global Survey, in: *Nature*, 15.10.2019, doi:10.1038/d41586-019-03105-2.

Chubin, Daryl E./Hackett, Edward J., *Peerless Science: Peer Review and U.S. Science Policy*, Albany, NY, 1990.

Coniam, David, Exploring Reviewer Reactions to Manuscripts Submitted to Academic Journals, in: *System*, 40 (4/2012), 544-553, doi:10.1016/j.system.2012.10.002.

Connor, Ull/Mauranen, Anna, Linguistic Analysis of Grant Proposals: European Union Research Grants, in: *English for Specific Purposes*, 18 (1/1999), 47-62, doi:10.1016/S0889-4906(97)00026-4.

Cronin, Blaise, Vernacular and Vehicular Language, in: *Journal of the American Society for Information Science and Technology*, 60 (3/2009), 433-433, doi:10.1002/asi.21010.

Da, Nan Z., The Computational Case against Computational Literary Studies, in: *Critical Inquiry*, 45 (3/2019), 601-639, doi:10.1086/702594.

Dall'Aglio, Paolo, Peer Review and Journal Models, in: *ArXiv*, 2006, URL: http://arxiv.org/abs/physics/0608307 [last accessed: Mar. 31, 2021].

Daniel, Hans-Dieter, *Guardians of Science: Fairness and Reliability of Peer Review*, Weinheim, 1993.

Dickersin, Kay, et al., Publication Bias and Clinical Trials, in: *Controlled Clinical Trials*, 8 (4/1987), 343-353, doi:10.1016/0197-2456(87)90155-3.

Dickersin, Kay/Min, Yuan-I./Meinert, Curtis L., Factors Influencing Publication of Research Results: Follow-up of Applications Submitted to Two Institutional Review Boards, in: *JAMA*, 267 (3/1992), 374-378, doi:10.1001/jama.1992.03480030052036.

Englander, Karent, Revision of Scientific Manuscripts by Non-Native English-Speaking Scientists in Response to Journal Editors' Language Critiques, in: *Journal of Applied Linguistics and Professional Practice*, 3 (2/2006), 129-161, doi:10.1558/japl.v3i2.129.

Ernst, E./Kienbacher, T., Chauvinism, in: *Nature*, 15.08.1991, 560, doi:10.1038/352560b0.

Eve, Martin Paul, *Close Reading With Computers: Textual Scholarship, Computational Formalism, and David Mitchell's* Cloud Atlas, Stanford, CA, 2019.

Eve, Martin Paul, et al., *Peer Review and Institutional Change in Academia*, Cambridge 2021.

Fanelli, Daniele, Do Pressures to Publish Increase Scientists' Bias? An Empirical Support from US States Data, in: *PLoS ONE*, 5 (4/2010), doi:10.1371/journal.pone.0010271.

Fang, Ferric C./Bowen, Anthony/Casadevall, Arturo, NIH Peer Review Percentile Scores Are Poorly Predictive of Grant Productivity, in: *ELife*, 5 (2016), e13323, doi:10.7554/eLife.13323.

Fish, Stanley, No Bias, No Merit: The Case against Blind Submission, in: *PMLA*, 103 (5/1988), 739-748.

Fogg, Louis/Fiske, Donald W., Foretelling the Judgments of Reviewers and Editors, in: *American Psychologist*, 48 (3/1993), 293-294, doi:10.1037/0003-066X.48.3.293.

Fortanet, Inmaculada, Evaluative Language in Peer Review Referee Reports, in: *Journal of English for Academic Purposes*, 7 (1/2008), 27-37, doi:10.1016/j.jeap.2008.02.004.

Fyfe, Aileen, et al., Managing the Growth of Peer Review at the Royal Society Journals, 1865-1965, in: *Science, Technology, & Human Values* 45 (3/2019), 405-429, doi:10.1177/0162243919862868.

Giannoni, Davide Simone, Medical Writing at the Periphery: The Case of Italian Journal Editorials, in: *Journal of English for Academic Purposes*, 7 (2/2008), 97-107, doi:10.1016/j.jeap.2008.03.003.

Gillespie, Gilbert W./Chubin, Daryl E./Kurzon, George M., Experience with NIH Peer Review: Researchers' Cynicism and Desire for Change, in: *Science, Technology, & Human Values*, 10 (3/1985), 44-54, doi:10.1177/016224398501000306.

Godlee, Fiona/Smith, Jane/Marcovitch, Harvey, Wakefield's Article Linking MMR Vaccine and Autism Was Fraudulent, in: *BMJ*, 342 (2011), c7452, doi:10.1136/bmj.c7452.

Goodman, Steven N., Manuscript Quality before and after Peer Review and Editing at Annals of Internal Medicine, in: *Annals of Internal Medicine*, 121 (1/1994), 11, doi:10.7326/0003-4819-121-1-199407010-00003.

Grimaldo, Francisco/Marušić, Ana/Squazzoni, Flaminio, Fragments of Peer Review: A Quantitative Analysis of the Literature (1969-2015), in: *PLOS ONE*, 13 (2/2018), e0193148, doi:10.1371/journal.pone.0193148.

Grimaldo, Francisco/Paolucci, Mario/Sabater-Mir, Jordi, Reputation or Peer Review? The Role of Outliers, in: *Scientometrics*, 116 (3/2018), 1421-1438, doi:10.1007/s11192-018-2826-3.

Hames, Irene, *Peer Review and Manuscript Management of Scientific Journals Guidelines for Good Practice*, Malden, MA, 2007.

Harwood, Nigel, "I Hoped to Counteract the Memory Problem, but I Made No Impact Whatsoever": Discussing Methods in Computing Science Using I, in: *English for Specific Purposes*, 24 (3/2005), 243-267, doi:10.1016/j.esp.2004.10.002.

Harwood, Nigel, "Nowhere Has Anyone Attempted … In This Article I Aim to Do Just That": A Corpus-Based Study of Self-Promotional I and We in Academic Writing across Four Disciplines, in: *Journal of Pragmatics*, Focus-on Issue: Marking Discourse, 37 (8/2005), 1207-1231, doi:10.1016/j.pragma.2005.01.012.

Herron, Daniel M., Is Expert Peer Review Obsolete? A Model Suggests That Post-Publication Reader Review May Exceed the Accuracy of Traditional Peer Review, in: *Surgical Endoscopy*, 26 (8/2012), 2275-2280, doi:10.1007/s00464-012-2171-1.

Ingelfinger, Franz J., Peer Review in Biomedical Publication, in: *The American Journal of Medicine*, 56 (1974), 686-692.

Ioannidis, John P. A., Effect of the Statistical Significance of Results on the Time to Completion and Publication of Randomized Efficacy Trials, in: *JAMA*, 279 (4/1998), 281-286, doi:10.1001/jama.279.4.281.

Jockers, Matthew L., *Macroanalysis: Digital Methods and Literary History*, Urbana, IL, 2013.

Kravitz, Richard L., et al., Editorial Peer Reviewers' Recommendations at a General Medical Journal: Are They Reliable and Do Editors Care?, in: *PLOS ONE*, 5 (4/2010), e10072, doi:10.1371/journal.pone.0010072.

LaFollette, Marcel C., *Stealing into Print: Fraud, Plagiarism, and Misconduct in Scientific Publishing*, Berkeley, CA, 1992.

Lamont, Michèle, *How Professors Think: Inside the Curious World of Academic Judgment*, Cambridge, MA, 2009.

Lillis, Theresa/Curry, Mary Jane, The Politics of English, Language and Uptake: The Case of International Academic Journal Article Reviews, in: *AILA Review*, 28 (2015), 127-150, doi:10.1075/aila.28.06lil.

Link, Ann M., US and Non-US Submissions: An Analysis of Reviewer Bias, in: *JAMA*, 280 (3/1998), 246-247, doi:10.1001/jama.280.3.246.

Lloyd, Margaret E., Gender Factors in Reviewer Recommendations for Manuscript Publication, in: *Journal of Applied Behavior Analysis*, 23 (4/1990), 539-543, doi:10.1901/jaba.1990.23-539.

Lock, Stephen, *A Difficult Balance: Editorial Peer Review in Medicine*, Philadelphia, PA, 1986.

MacCallum, Catriona J., ONE for All: The Next Step for PLoS, in: *PLOS Biology*, 4 (11/2006), e401, doi:10.1371/journal.pbio.0040401.

Mahoney, Michael J., Publication Prejudices: An Experimental Study of Confirmatory Bias in the Peer Review System, in: *Cognitive Therapy and Research*, 1 (2/1977), 161-175, doi:10.1007/BF01173636.

McCurdy, David W./Spradley, James P./Shandy, Dianna J., *The Cultural Experience: Ethnography in Complex Society*, Long Grove, IL, 2005.

Meruane, Omar Sabaj/Vergara, Carlos González/Pina-Stranger, Álvaro, What We Still Don't Know About Peer Review, in: *Journal of Scholarly Publishing*, 47 (2/2016), 180-212, doi:10.3138/jsp.47.2.180.

Moore, Samuel, et al., Excellence R Us: University Research and the Fetishisation of Excellence, in: *Palgrave Communications*, 3 (2017), doi:10.1057/palcomms.2016.105.

Moretti, Franco, *Distant Reading*, London 2013.

Moretti, Franco, *Graphs, Maps, Trees: Abstract Models for Literary History*, London 2007.

Mulligan, Adrian/Hall, Louise/Raphael, Ellen, Peer Review in a Changing World: An International Study Measuring the Attitudes of Researchers, in: *Journal of the American Society for Information Science and Technology*, 64 (1/2013), 132-161, doi:10.1002/asi.22798.

Murray, Dakota, et al., Author-Reviewer Homophily in Peer Review, in: *BioRxiv* 400515 (2019), doi:10.1101/400515.

Mustaine, Elizabeth Ehrhardt/Tewksbury, Richard, Reviewers' Views on Reviewing: An Examination of the Peer Review Process in Criminal Justice, in: *Journal of Criminal Justice Education*, 19 (3/2008), 351-365, doi:10.1080/10511250802476178.

Okike, Kanu, et al., Single-Blind vs Double-Blind Peer Review in the Setting of Author Prestige, in: *JAMA*, 31 (12/2016), 1315-1316, doi:10.1001/jama.2016.11014.

Paltridge, Brian, *The Discourse of Peer Review: Reviewing Submissions to Academic Journals*, London 2017.

Peters, Douglas P./Ceci, Stephen J., Peer-Review Practices of Psychological Journals: The Fate of Published Articles, Submitted Again, in: *Behavioral and Brain Sciences*, 5 (2/1982), 187-195, doi:10.1017/S0140525X00011183.

Petty, Richard E./Fleming, Monique A./Fabrigar, Leandre R., The Review Process at PSPB: Correlates of Interreviewer Agreement and Manuscript Acceptance, in: *Personality and Social Psychology Bulletin*, 25 (2/1999), 188-203, doi:10.1177/0146167299025002005.

Pierie, Jean-Pierre E.N., Walvoort, Henk C./Overbeke, A. John P.M., Readers' Evaluation of Effect of Peer Review and Editing on Quality of Articles in the Nederlands Tijdschrift Voor Geneeskunde, in: *The Lancet*, 348 (9040/1996), 1480-1483, doi:10.1016/S0140-6736(96)05016-7.

Piper, Andrew, *Enumerations: Data and Literary Study*, Chicago, IL, 2018.

PLOS, Journal Information, in: *PLOS ONE*, 2016, URL: http://www.plosone.org/static/information [last accessed: Mar. 31, 2021].

Poynder, Richard, *PLoS ONE, Open Access, and the Future of Scholarly Publishing*, 2011, URL: https://richardpoynder.co.uk/PLoS_ONE.pdf [last accessed: Mar. 31, 2021].

Retraction Watch, URL: https://retractionwatch.com/ [last accessed: Mar. 31, 2021].

Ross, Joseph S., et al., Effect of Blinded Peer Review on Abstract Acceptance, in: *JAMA*, 295 (14/2006), 1675-1680, doi:10.1001/jama.295.14.1675.

Ross-Hellauer, Tony, What Is Open Peer Review? A Systematic Review, in: *F1000Research*, 6 (2017), 588, doi:10.12688/f1000research.11369.2.

Rothwell, Peter M./Martyn, Christopher N., Reproducibility of Peer Review in Clinical Neuroscience, in: *Brain*, 123 (9/2000), 1964-1969, doi:10.1093/brain/123.9.1964.

Schmidt, Benjamin, Words Alone: Dismantling Topic Models in the Humanities, in: *Journal of Digital Humanities*, 2 (1/2012), URL: http://journalofdigitalhumanities.org/2-1/words-alone-by-benjamin-m-schmidt/ [last accessed: April 1, 2021].

Shatz, David, *Peer Review: A Critical Inquiry*, Issues in Academic Ethics, Lanham, MD, 2004.

Shehzad, W., How to End an Introduction in a Computer Science Article: A Corpus-Based Approach, in E. Fitzpatrick (ed.): *Corpus Linguistics Beyond the Word: Corpus Research from Phrase to Discourse*, Amsterdam 2015, 227-241, URL: https://brill.com/view/title/30110 [last accessed: Mar. 31, 2021].

Silbiger, Nyssa J./Stubler, Amber D., Unprofessional Peer Reviews Disproportionately Harm Underrepresented Groups in STEM, in: *PeerJ*, 7 (2019), e8247, doi:10.7717/peerj.8247.

Smigel, Erwin O./Ross, H. Laurence, Factors in the Editorial Decision, in: *The American Sociologist*, 5 (1/1970), 19-21.

Smith, Richard, Classical Peer Review: An Empty Gun, in: *Breast Cancer Research*, 12 (4/2010), S13, doi:10.1186/bcr2742.

Smith, Richard, Peer Review: A Flawed Process at the Heart of Science and Journals, in: *Journal of the Royal Society of Medicine*, 99 (4/2006), 178-182.

Squazzoni, Flaminio, Peering Into Peer Review, in: *Sociologica*, 3 (2010), 1-27, doi:10.2383/33640.

Sternberg, Robert J., et al., Getting in: Criteria for Acceptance of Manuscripts in Psychological Bulletin, 1993-1996, in: *Psychological Bulletin*, 121 (2/1997), 321-323, doi:10.1037/0033-2909.121.2.321.

Sugimoto, Cassidy R./Cronin, Blaise, Citation Gamesmanship: Testing for Evidence of Ego Bias in Peer Review, in: *Scientometrics*, 95 (3/2013), 851-862, doi:10.1007/s11192-012-0845-z.

Swales, John, *Genre Analysis: English in Academic and Research Settings*, Cambridge 1990.

Tennant, Jonathan/Ross-Hellauer, Tony, The Limitations to Our Understanding of Peer Review, in: *SocArXiv*, 2019, doi:10.31235/osf.io/jq623.

Tewksbury, Richard/Ehrhardt Mustaine, Elizabeth, Cracking Open the Black Box of the Manuscript Review Process: A Look Inside, in: *Justice Quarterly*, *Journal of Criminal Justice Education*, 23 (4/2012), 399-422, doi:10.1080/10511253.2011.653650.

Travis, G. D. L./Collins, H. M., New Light on Old Boys: Cognitive and Institutional Particularism in the Peer Review System, in: *Science, Technology, & Human Values*, 16 (3/1991), 322-341, doi:10.1177/016224399101600303.

Tregenza, Tom, Gender Bias in the Refereeing Process?, in: *Trends in Ecology & Evolution*, 17 (8/2002), 349-350, doi:10.1016/S0169-5347(02)02545-4.

Underwood, Ted, A Genealogy of Distant Reading, in: *Digital Humanities Quarterly*, 11 (2/2017), URL: http://www.digitalhumanities.org/dhq/vol/11/2/000317/000317.html [last accessed: Mar. 31, 2021].

Underwood, Ted, *Distant Horizons: Digital Evidence and Literary Change*, Chicago, IL, 2019.

Wakefield, A. J.et al., RETRACTED: Ileal-Lymphoid-Nodular Hyperplasia, Non-Specific Colitis, and Pervasive Developmental Disorder in Children, in: *The Lancet*, 351 (9103/1998), 637-641, doi:10.1016/S0140-6736(97)11096-0.

Weller, Ann C., Editorial Peer Review: Its Strengths and Weaknesses, in: *Journal of the Medical Library Association*, 90 (1/2002).

Woods, Peter, *Successful Writing for Qualitative Researchers*, London 2006.

Zuckerman, Harriet/Merton, Robert K., Patterns of Evaluation in Science: Institutionalisation, Structure and Functions of the Referee System, in: *Minerva*, 9 (1/1971), 66-100, doi:10.1007/BF01553188.

Chapter 6: Supervised and Unsupervised: Approaches to Machine Learning for Textual Entities

Tobias Hodel, University of Bern

Abstract

Applications that feed text into machine learning algorithms have existed for more than a decade. But it took multiple developments to make machine learning an exciting methodological approach to questions grounded in the humanities. The latest developments in handwritten text recognition (HTR) show the capabilities of supervised deep learning. However, the success of the technology comes with a price: It generates a set of methods that are complicated to grasp in theory and difficult to train algorithms in, algorithms that are not comprehensible to humans at all. By focusing on the two most frequently used approaches in machine learning (unsupervised and supervised), this paper lays out ways to critically use machine learning algorithms in the humanities. At the same time, we argue that these approaches help us to understand the epistemological assumptions of our disciplines and our methods.

Topic modeling used on large corpora of text leads to new insights into what topics occur, as well as the tendencies of a corpus. The approach uses unsupervised machine learning, through which a set of algorithms identify what words appear together frequently and so might indicate a topic. Topic modeling puts scholars at the end of the process, where they must still interpret the output of the algorithms.

In deciphering handwriting, supervised deep learning approaches have led to astonishing results, but also to new problems induced by the algorithm. The algorithm tries to adapt to the desired output, raising epistemological questions about transcribing and transliterating. The scholar is only able to alter the input, not how the algorithm manipulates it.

Based on these two examples, this paper promises a deeper understanding of a technology that is currently remodeling the way we do our research and that will increasingly intervene in our scholarship and even our daily lives in the future.

1. Machine Learning and the Humanities

It is pretty safe to assume that, in retrospect and from a computer-historical perspective, the 2010s and most probably also the 2020s will be seen as the era of the application of Artificial Intelligence (AI), or, more precisely, machine learning.[1]

From a scholarly standpoint, the arrival of AI has not, on the surface, led to a complete rethinking of research activities. Instead, established procedures have been altered slowly and sometimes impalpably. To search for literature, scholars have relied on finding aids in libraries for more than five centuries.[2] With search engines and a wide array of database infrastructures, however, the process of finding relevant primary and secondary sources is being completely changed. These changes have not been reflected in the products of our research: papers and books.[3]

The next phase of AI in the humanities is currently building up from within the humanities, as scholars have for some years now experimented with machine learning. With the advent of Tensorflow and other quite intuitively usable libraries that allow us to build deep and machine learning systems without higher degrees in engineering, we find ourselves in the first wave of applications and solutions that provide scholars with algorithms that are, at least to some degree, 'self-taught'.[4]

The growth of machine learning in the humanities disciplines has been fueled by the promises of AI, including the automatic recognition, identification and linking of text, entities, and even concepts. Algorithms that could meet these challenges not only benefit the shareholder value of companies using such tools to create adapted advertisements, but also scholars working on recently edited or digitized texts and corpora built on the web.

Although the subjects of AI and machine learning are often mentioned, we nonetheless lack much discussion of their methodological implications. We can tie

---

1 The author would like to thank the editor and the audience of the conference for its invaluable feedback. Thanks are as well due to Jake Purcell for lending his critical eye as well as his language skills. This article has benefited from work done within the READ project. READ has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement no. 674943.

2 Ann Blair, *Too Much to Know: Managing Scholarly Information before the Modern Age*,NewHaven 2010.

3 Concerning the looming overabundance of sources, see: Roy Rosenzweig, Clio Wired: The Future of the Past in the Digital Age, New York, 2011 3-27 (chapter Scarcity and Abundance? Preserving the Past). A comprehensive piece about using search engines in the humanities still needs to be written or I am just not aware of it.

4 For a broad and general introduction see Christof Schöch, Quantitative Analyse, in: Fotis Jannidis/Hubertus Kohle/Malte Rehbein (eds.), *Digital Humanities: Eine Einführung*, 279-298, Stuttgart 2017, doi:10.1007/978-3-476-05446-3_20. Tensorflow is an open-source library for machine learning: https://www.tensorflow.org/ [last accessed: April 2, 2021].

this absence to the complexity of the technology involved, as well as to the difference between machine learning and typical algorithms that follow strict orders and can be controlled much more easily. The capabilities of machine learning coupled with this black box of uncontrolled (or hard-to-control) parts leads to excitement, but also—and rightly so—to an unease about its widespread use in any discipline.

It's neither possible nor desirable to address all the problems and challenges posed by machine learning in one paper, so I will shed a light on the inner workings of this black box and put two approaches in perspective against the broader field of machine learning. In the upcoming years, we as scholars need to be able to actively engage in discussions about the potential and problems of AI at its current stage of implementation.

The paper will tackle two different approaches that fall under the umbrella of machine learning. In order to understand the differences, we will first briefly introduce how machine learning can be experienced and what trajectories are expected in the area of text analysis in the near future. From this theoretical and technical background, I will proceed to discuss two different applications of machine learning that show the gains and losses when using machine learning to analyze text. The goal of these two parts is not to showcase machine learning algorithms but to encounter methodological as well as epistemological consequences of the application of two very different forms of AI. Even before the linguistic turn, the humanities have been at the center of deliberations about what it means to "understand," and humanists have opted for hermeneutical rather than purely quantitative approaches.[5]

The paper tries to situate practices of the Digital Humanities in the quickly broadening field of critical algorithm studies. In addition to incorporating important questions brought up by recent research, I try not to focus on the analysis of implemented algorithms (like search engines)[6] or on cultural traits (like race),[7] but rather on the implementation of algorithmic solutions to problems grounded in the humanities. This approach is in line with developments within the digital humanities that emphasize the work of theorizing digital methods.[8]

---

5 On the hermeneutical method, see Hans-Georg Gadamer, *Hermeneutik I: Wahrheit Und Methode: Grundzüge Einer Philosophischen Hermeneutik*, Tübingen 2010.

6 Safiya Umoja Noble, *Algorithms of Oppression: How Search Engines Reinforce Racism*, New York 2018.

7 Ruha Benjamin, Ruha, *Race After Technology: Abolitionist Tools for the New Jim Code*, Medford 2019.

8 Ted Underwood, Theorizing Research Practices We Forgot to Theorize Twenty Years Ago, *Representations* 127 (1/2014), 64–72, doi:10.1525/rep.2014.127.1.64.

2. About the "Intelligence" of Machines

A Google search for "artificial intelligence" yields, in the middle of 2020, more than 700 million hits. Starting with Wikipedia pages—in languages that depend on your location—and brief introductory videos from YouTube, the first pages of search results consist only of explanations of the term and its possible meanings. Although the concept of AI has been around for more than fifty years, only in the last five to ten years has it been more than a buzzword. Since the arrival of faster central processing units (CPUs) and the application of graphical processing units (GPUs, typically used to render 2D and 3D imagery) to machine learning, the term AI has become more tangible, due to its actual impact on applications and ensuing discussions about its consequences.[9]

As the Google search demonstrated by putting very broad introductions at the top of the result hierarchy, the general understanding of AI is sketchy and often far from what algorithms at the moment can actually achieve. In order to get an idea of machine learning capabilities, we need to take a small detour to briefly think about what intelligence means in the context of machines. At the same time, if we compare the impressions based on the introductions with actual results of algorithms, we will see that the current state opens up a tremendous amount of possibilities, but also that we still advance on a step-by-step basis and shouldn't overestimate the role of machines (yet). Algorithms remain in a state of being useful "helpers," and nothing more.

Although the humanities do not deal only with textual elements, a multitude of disciplines use text as one of their main foundations for furthering our knowledge within and across disciplines. The use of AI to recognize and sort or cluster text is thus a typical approach to using these technologies in the humanities. From a technical point of view, we differentiate three types of machine learning: supervised, unsupervised, and reinforcement learning.[10] The difference between the three types lies in the role of the optimization process. In supervised learning,

---

9 Discussions about "fair AI" are currently being brought forward in a wide variety of research centers, probably most notably in the AI Now institute at New York University: https://ainowinstitute.org/ [last accessed: April 2, 2021]. See also (in German): iRights lab et al., Praxisleitfaden zu den Algo. Rules - Orientierungshilfen für Entwickler:Innen und ihre Führungskräfte, Gütersloh 2020, doi:10.11586/2020029.

10 In this paper, I will not go into detail about the differences between classical machine learning and deep learning. Both approaches are addressed since topic modeling can be included in the former and handwritten text recognition to the latter. See also Ted Underwood/ Matthew L. Jockers, Text-Mining the Humanities, in: Susan Schreibman/ Ray Siemens/John Unsworth, *A New Companion to Digital Humanities*, Chichester 2016, 291–306. We do not agree with Underwood and Jockers that topic modeling belongs to text mining, as opposed to machine learning.

the algorithm aligns input with a human-determined outcome, and training processes try to get the algorithm's actual output as close as possible to the desired result. In unsupervised learning, the result is not predetermined and then imitated. Instead, algorithms search for patterns, similarities, and clusters. Finally, reinforcement learning evolves out of feedback (automatic or manual) and develops an algorithm on the fly.[11] Since reinforcement learning is currently not used in the humanities (at least not to my knowledge), I will not deal with this last approach in this paper.

For all three approaches, it is difficult to speak of machine intelligence as compared with the human capacity to learn and adapt to circumstances, social settings, languages, etc. In general, it would be easy to dismiss the notion that, at the current stage, it is advisable to talk about intelligence in regards to machines and algorithms at all.[12] At the same time, it's remarkable how trained algorithms can perform challenging tasks in a shorter time than humans.

All machine learning algorithms are based on their defined input and output; as a consequence, we can look at those two crucial parts of the process. Until about a year ago (2019), most machine learning processes started from scratch, and models were built from available data, called "Ground Truth" denominating what is correct with a certainty. In order to measure output from algorithms, computer scientists coined the term "Ground Truth" which determines a perfect or desired result. In supervised learning such data, the Ground Truth is used for the training as well as the validation process. Currently, this procedure is evolving, since certain models have been made available and can be used for refinement training or as base models.[13] The reuse of models will lead to further problems in producing algorithms, since users of pretrained models will have to deal with bias induced from training data not available to them. Currently, first experiments with pre-trained models are being conducted, and reliable statements cannot yet be made.

---

11 Best known is maybe the Super Mario algorithm MARI/O, that masters the well-known Nintendo game in a miraculous manner: URL: https://www.youtube.com/watch?v=qv6UVOQ0F44&t=115s [last accessed: April 2, 2021].

12 For an introduction, see the articles in the June 2020 issue of Wired, esp. Elizabeth Spelke, It's Called Artificial Intelligence—but What Is Intelligence?, in: *Wired* 19.05.2020, URL: https://www.wired.com/story/its-called-artificial-intelligence-but-what-is-intelligence/ [last accessed: April 2, 2021]; Kelly Clancy, Is the Brain a Useful Model for Artificial Intelligence?, in: *Wired* 19.05.2020, URL: https://www.wired.com/story/brain-model-artificial-intelligence/ [last accessed: April 2, 2021].

13 This is especially true for modern languages. Concerning language models, see Jacob Devlin et al., BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding, in: *arXiv* [preprint], 24.05.2019, https://arxiv.org/abs/1810.04805 [last accessed: April 2, 2021]; Tom B. Brown et al., Language Models Are Few-Shot Learners, in: *arXiv* [preprint], 22.07.2020, URL: https://arxiv.org/abs/2005.14165 [last accessed: April 2, 2021].

The crucial question remains: How can results be judged, apart from via statistical reports? This problem will only be addressed briefly in this paper and needs more elaboration in the future. Ways to qualify the output of machine learning will be a key issue, due to the influence of the technology not only on scholarly work with documents (as data), but also due to the embeddedness of machine learning in the algorithms of our daily life. Quantification of results using statistical techniques, such as the F1-score (a comparison of an algorithm's recall and precision) or percentages of correctly identified characters (if we think about text recognition), is one indication of the capability of an algorithm, but it doesn't show problems, uncertainties, or bias induced by the approach. F1-scores, for example, tell us about the quality and quantity of intended results, but say nothing about unintended consequences due to any imprecisions.

To highlight the differences among machine learning approaches, I will provide two examples that use different types of machine learning (supervised and unsupervised) and deal with questions for which machine learning yields impressive results. For the unsupervised approach, I will look at topic modeling, and for the supervised counterpart, the application of deep learning to the recognition of handwritten documents. The aim is only to introduce briefly the two approaches from a rather theoretical point of view, not to provide a how-to guide for the two methods.[14]

3. Topic Modeling: Unsupervised Clustering

One of the main advantages computers have over humans is their ability to count and compare extremely quickly. With topic modeling, scholars use these two traits and try to apply them to text. The algorithms used for topic modeling work in two directions: First, they count the appearance of strings (called "tokens") in textual entities (e.g. a letter or a document). The expected term for "token" might instead be "word," but since this term is polysemic and could mean a string of characters (a token) or the semantic meaning of the string (in a sense the lemmatized token), we will use token instead. Second, the tokens are compared to other strings appearing in the same entity. Pairs, triples… clusters of tokens appearing often across different entities are understood to belong to a distinct topic. In this perspective, a topic is nothing other than a collection of tokens appearing in context. Whether a

14 In order to get acquainted to the methods, we recommend, for topic modeling: Shawn Graham/Scott Weingart/Ian Milligan, Getting Started with Topic Modeling and MALLET, in: *Programming Historian* 02.09.2012, URL: https://programminghistorian.org/en/lessons/topic-modeling-and-mallet [last accessed: April 2, 2021], and for Handwritten Text Recognition, see URL: https://transkribus.eu/wiki/index.php/How_to_Guides [last accessed: April 2, 2021].

calculated "topic" is really congruent to the human understanding of a topic needs further discussion. The goal of the algorithm is basically to build an understanding of the rate of the occurrence of tokens in the same entity. In a third step, the process gets inverted, and for every category (or "topic"), the percentage of a textual entity that consists of this topic is calculated.

Fig 6.1: Screenshot of a section of a visual topic modeling output. Vertically on the left are the documents specified (signature, volume, document); horizontally are topics indicated. Topic 16: Warfare, topic 17: Finance, topic 18: Poverty, topic 19 Foreign policy. The color indicates the presence of a topic in a document. Screenshot by the author.

Signature_Volume_Doc	Topic 16	Topic 17	Topic 18	Topic 19	Topic 20	Topic 21	Topic 22	Topic 23
MM_2_100_RRB_1848_0528.	0.36	0.00	0.00	0.08	0.00	0.00	0.00	0.16
MM_2_100_RRB_1848_0830.	0.31	0.02	0.00	0.00	0.00	0.00	0.00	0.00
MM_2_100_RRB_1848_0926.	0.31	0.00	0.00	0.20	0.00	0.00	0.00	0.00
MM_2_101_RRB_1848_1131.	0.31	0.12	0.19	0.00	0.02	0.00	0.00	0.00
MM_2_101_RRB_1848_1594.	0.30	0.00	0.00	0.00	0.00	0.00	0.00	0.00
MM_2_101_RRB_1848_1644.	0.30	0.00	0.00	0.00	0.00	0.00	0.00	0.00
MM_2_102_RRB_1848_1723.	0.31	0.00	0.02	0.00	0.00	0.00	0.00	0.00
MM_2_102_RRB_1848_1826.	0.32	0.00	0.00	0.00	0.00	0.00	0.00	0.00
MM_2_102_RRB_1848_2013.	0.30	0.00	0.00	0.00	0.00	0.00	0.00	0.00
MM_2_102_RRB_1848_2109.	0.93	0.04	0.00	0.00	0.00	0.00	0.00	0.00
MM_2_103_RRB_1849_0018.	0.30	0.06	0.00	0.00	0.01	0.00	0.03	0.15
MM_2_103_RRB_1849_0139.	0.93	0.00	0.00	0.00	0.00	0.00	0.00	0.00
MM_2_103_RRB_1849_0231.	0.31	0.00	0.00	0.00	0.03	0.00	0.00	0.10
MM_2_103_RRB_1849_0335.	0.32	0.00	0.00	0.00	0.00	0.00	0.00	0.00
MM_2_103_RRB_1849_0399.	0.30	0.00	0.00	0.00	0.00	0.00	0.00	0.03
MM_2_103_RRB_1849_0468.	0.30	0.00	0.00	0.00	0.09	0.00	0.00	0.00
MM_2_103_RRB_1849_0623.	0.36	0.00	0.00	0.00	0.11	0.00	0.00	0.00
MM_2_104_RRB_1849_0649.	0.30	0.00	0.00	0.45	0.00	0.00	0.00	0.03
MM_2_105_RRB_1849_1189.	0.36	0.00	0.01	0.00	0.00	0.00	0.00	0.15
MM_2_105_RRB_1849_1192.	0.31	0.00	0.00	0.00	0.00	0.00	0.00	0.01
MM_2_105_RRB_1849_1545.	0.30	0.39	0.00	0.00	0.00	0.00	0.00	0.00
MM_2_105_RRB_1849_1660.	0.32	0.00	0.00	0.00	0.00	0.00	0.00	0.00
MM_2_105_RRB_1849_1706.	0.36	0.08	0.00	0.00	0.02	0.00	0.00	0.06
MM_2_105_RRB_1849_1717.	0.30	0.12	0.00	0.00	0.00	0.00	0.00	0.14
MM_2_105_RRB_1849_1794.	0.31	0.02	0.00	0.00	0.00	0.00	0.00	0.25
MM_2_105_RRB_1849_2020.	0.30	0.00	0.00	0.00	0.00	0.00	0.00	0.07
MM_2_106_RRB_1849_2070.	0.31	0.00	0.00	0.00	0.00	0.00	0.00	0.24
MM_2_106_RRB_1849_2087.	0.30	0.32	0.00	0.00	0.00	0.00	0.00	0.00
MM_2_106_RRB_1849_2310.	0.32	0.00	0.00	0.00	0.00	0.00	0.00	0.22

The best way to describe the algorithm figuratively is to understand topic modeling as a scissor that cuts a text into single words (literally done in some cases)[15] and throws them into small bags (this is also where the technical term "bag of words" comes from). Now you determine the chances of two words occurring in one of the bags. If put in a table, combined with the information about the most frequent tokens, this probability is the output of a topic model.

15 Dan Hirschman, Doing Things with Bags-of-Words, in: *Scatterplot* 19.02.2020, URL: https://scatter.wordpress.com/2020/02/19/doing-things-with-bags-of-words/ [last accessed: April 2, 2021].

An entity could be a letter, a book, a chapter, whatever seems to be a useful comparator. So, if we want to build a model of topics appearing in different Wikipedia articles—a classical example where topic models are being built—the single article would be the entity.[16] For this paper, I will resort to the term "document" to indicate an entity that is of interest to us, one that can be compared and eventually shows or teaches (from the Latin "docere") something.

The problem with the generation of topic lists becomes clear once people are confronted with a specific topic model. As the main input, we need to define how many topics are to be found. However, according to the explanation of topic modeling above, clusters are only found near the end of the process, so the system obviously cannot know how many of these clusters to expect (though there are some ways to check if the number of clusters is statistically "ideal"). For a scholar, this input is counterintuitive, since we don't know at the beginning of the research process how many different topics to expect from a corpus. Working with students using topic modeling for the first time, I in general had the experience that they set the number of expected topics too low, since people were interested in very general topics. A smaller number of topics doesn't mean a more accurate depiction of the most common topics, but rather a clustering of very general terms (maybe on level of adverbs or adjectives).

For this experiment, I worked with a dataset of about 150,000 nineteenth-century documents, called *Regierungsratsprotokolle*, that record decisions by the highest executive of the canton of Zürich (a district in modern Switzerland) and its administrative predecessor. At first glance, the tokens in the generated list did not form coherent "topics," but rather were just lists of terms that seemed to have some relation. If we rethink the process of their generation, this result can be explained: The clearest expression of a topic might not occur that often in a corpus. Even in mass produced literature in genres like romance, we might not encounter the token "love" as often as "hugging" or "kissing," or "felt" or "feeling."[17] Probably the most striking topic in the *Regierungsratsprotokolle* deals with warfare, but the token "Krieg" is missing completely; instead we find "Militärs" (military), "Regierungsrath" (the executive, appearing in almost all generated topics), "Kriegsrathe" (war council), "Infanterie" (infantry), "Truppen" (troops) and so forth. This example is quite intuitively interpreted, but several other "topics" can only be identified from token

16 For an example of a combined approach of categorization and Latent Dirichlet Allocation, see Xu Kang et al., Incorporating Wikipedia Concepts and Categories as Prior Knowledge into Topic Models, in: *Intelligent Data Analysis* 21 (2/2017), 443-461, doi:10.3233/IDA-160021.

17 Regarding the use of topic modeling for literary genres, see Christof Schöch, Topic Modeling Genre: An Exploration of French Classical and Enlightenment Drama, in: *Digital Humanities Quarterly* 11/2 (2017), URL: http://www.digitalhumanities.org/dhq/vol/11/2/000291/000291.html [last accessed: April 2, 2021].

lists with difficulty, and for certain lists it is nearly impossible. With some background knowledge about the corpus, it should eventually be possible to make sense of the lists, but until we hit something like topic 22, the list for this topic reads as follows: "zeit, diesem, weise, während, dann, sollte, darauf, zwar, möglich, seit, namentlich, alle…" (time, this, ways, while, then, should, thereupon, possible, since, namely, all…). Even with in-depth knowledge about the corpus, it's not possible to interpret some of the topics since they consist of auxiliary vocabulary with limited testimony.

Approaches to streamlining topic modeling do exist: 1) Text can be pre-processed according to part of speech, especially using normalization by lemmatization, 2) Elaborate stop-word lists can be implemented, containing words that do not have "meaning." The lemmatization of tokens as a preprocessing step will lead to text that is less variable and to the federation of tokens, especially when plural and singular forms of nouns or verbs are treated as identical. For highly standardized languages, such as most western languages including modern English, automatic lemmatization is already possible and leads to excellent results (some of the approaches are rule-based; others, more efficient, tend to use machine learning as well). As soon as we switch our interest to most languages pre-1900 (most historical documents) or to languages that so far have not attracted commercial enterprises such as search engines, vendors, or social media giants, the situation is completely different. The lack of annotated corpora, the quantity of variations in spelling, and, frankly, the lack of interest from researchers yield subpar results and consequently we are denied this method. With the second approach, the usage of stop-word lists, we could also resort to using ready-made files, which might result in unconsciously erasing tokens. Maybe such approaches even take us into philosophical or theoretical-linguistic territories, since we need to determine what words do reveal, or at least nudge us towards, the meaning of a document.

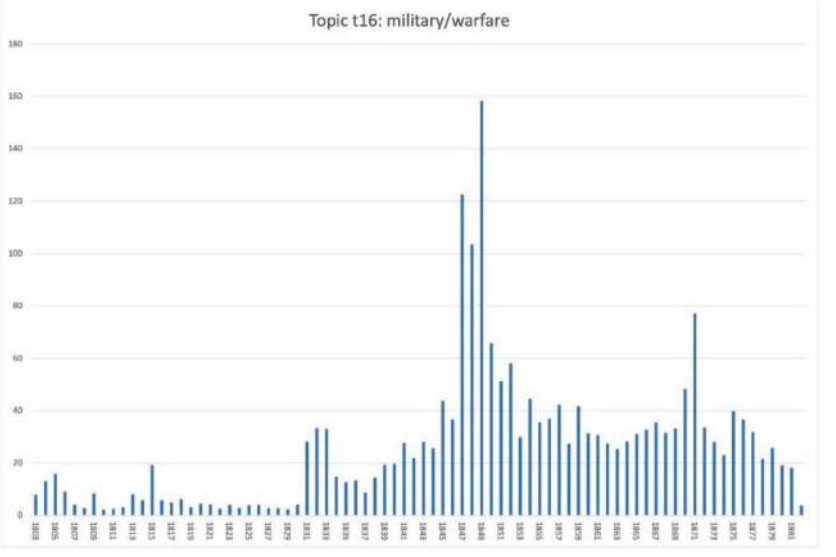
This leaves us for the moment in a difficult situation, since the algorithm indicates a variety of occurring topics only through a list in need of interpretation. It is through visualizations and repeated (re-)interpretation of topic lists that we will get an understanding of both the corpus we are analyzing and the method we are applying. In a sense, the cycle of interpretation can be understood as a hermeneutical cycle (here understood in a positive way) bringing us step-by-step closer to the corpus and thus the subject of our research.[18]

If we make temporal visualizations for the example corpus, we find striking traces of how some clusters indicate meaningful topics, helping us to understand what was decided by Zürich's executive in the 19th century. The topic indicating

18 For a broader introduction to topic modeling see also: Scott Weingart, Topic Modeling for Humanists: A Guided Tour, in: *The scottbot irregular* 25.07.2012, URL: http://www.scottbot.net/HIAL/?p=19113 [last accessed April 2, 2021].

"warfare" peaks in the middle/end of the 1840s, right at the time when Switzerland struggled towards a civil war (the so-called *Sonderbundskrieg*, in 1847). [19] And, even after the Swiss were pacified, the topic remained virulent: as we know, there were uprisings in the rest of Europe, especially in the bordering German lands in 1848. A second, smaller peak indicates the time of the French-German war (with battles in nearby Belfort) and the surrender of the Bourbaki army on Swiss soil in 1871. At the time, the Swiss army (including troops from Zürich) was mobilized and present at the borders.

Fig 6.2: Visualization of topic 16, concerning military/warfare in the corpus "Zürcher Regierungsratsprotokolle". The graph demonstrates how often a topic occurs (y-axis) in the documents produced in a certain year (x-axis).



Of course, historians knew about all those encounters, and searching for consequences of mobilization would have been possible in the minutes even without topic modeling. But the one topic only covers a small part of what can be found in the minutes. Some topics focus on the building of infrastructures to facilitate the use of railways. Quite frequently, financial decisions had to be made, another topic reflected in this group of documents.

19 A classical reference for the war in 1847 is Edgar *Der Sonderbundskrieg*, Zürich, 1947. For later research, see Pierre *La Guerre Du Sonderbund: La Suisse de 1847*, Neuchâtel 2018; Hans Rudolf Fuhrer/Jean Paul Loosli/Christian Moser, *Sonderbundskrieg 1847*, Wettingen 1997.

The main challenge—and, fittingly, this is a general problem in the humanities—is discerning relevant from irrelevant incidences, not on the level of sources but rather on a meta or mediated level. The algorithm forces the user to deal with processed documents—or data—and urges them to shift to a multitude of perspectives: From in-depth analysis and close reading of single pages in a source to a birds-eye view, scanning hundreds and thousands of pages, trying to make sense of a corpus.

The approach is of course not free of flaws. From a methodological point of view, we can identify a multitude of problems in addition to the exceptional opportunities. For one, the input, the corpus, very much influences the topics harvested. Furthermore, the number of topics needs to be defined beforehand, since the optimization process needs to optimize toward a value. The number of topics has therefore to be played around with, since some of the topics generated are mere lists of "non-topical vocabularies" that probably will not be taken into consideration when making statements about the content of the corpus.

In a sense, the approach can thus not be generalized. There are some (ongoing) attempts in that direction, but they raise a multitude of underlying questions (what's a general topic? are there topics independent of genres?) and consequently are mostly used within narrow typological frames.

From a technical standpoint and compared with deep learning approaches (we will tackle those later on), topic modeling is very much explainable and based on a set of clustering algorithms.[20]

Still, we recognize bias—not only in negative meaning—on different levels. First, the corpus biases the output in terms of topics. Documents from a narrow field will lead to a number of similar-seeming topics from that field (in the Regierungsrathsprotokolle "finances" as an ever important topic is present in at least three calculated topics). Second, the bias is induced by the user trying to make sense of the list of tokens. Depending on the knowledge of the corpora, expected topics will be read into the list.

The question arising from this is how to build data sets that are not biased. Is it possible to broaden the input, in order not to have the material focus on specific aspects? At the end of my consideration of topic modeling, we find ourselves right at the heart of issues of bias in machine learning. Topic modeling has often been advertised as a new way to approach digitized documents. The capability to visualize a multitude of documents quite easily would seem to make this method a

20 For latent dirichlet allocation, see David M. Blei, Introduction to Probabilistic Topic Models, in: *Communication of the ACM* (2011); David M. Blei/Andrew Y. Ng/Michael I. Jordan, Latent Dirichlet Allocation, in: *Journal of Machine Learning Research*, (3/2003), 993–1022.

preferred one for carving meaning out of a corpus. When considering the problems of the approach, however, the ease vanishes and critical stances open up.[21]

Despite the embedded problems, topic modeling offers a new way to treat sources heuristically, and it supports scholarly endeavors that want to deal with the abundance of sources that arises from mass digitization and born-digital data. Topic modeling allows this first steps toward big data in the humanities,[22] without necessarily performing quantitative analyses in later study stages.

In topic modeling, the clustering algorithm uses statistical methods that belong to the realms of machine learning. Although the algorithm starts at a more-or-less random point of word distribution, it is still very much understandable how the results are generated. It would even be possible, in theory at least, to replicate the result using analogue methods.

If we switch to deep learning, it would not be possible to do so. As we will see, there's a significant difference in the handling of input and output in the context of neural networks. Deep learning gained a lot of traction in the past decade and is currently implemented in a wide variety of algorithms (from image identification to self-driving cars, digital medicine and hiring applications). In the humanities, we currently stand at the starting point of this movement, with the first algorithms based on the networks being used on a regular basis, for example (and especially successfully) for handwritten text recognition.[23]

4. Handwritten Text Recognition: Supervised Training

With the advent of text recognition for print in the 1990s based on optical character recognition (OCR), similar results for handwriting seemed only a few years away. But the technology, based on the idea of isolating single characters, was never capable of delivering meaningful results, spare some success in recognizing neatly painted letters from the Middle Ages. Only in 2010 did the introduction of neural networks and the field of deep learning lead to a stark and astonishing improvement in handwritten text recognition (HTR), at that time as a proof of concept.

---

21 Ben Schmidt, Words Alone: Dismantling Topic Models in the Humanities, in: *Journal of Digital Humanities* 2 (1/2013), URL: http://journalofdigitalhumanities.org/2-1/words-alone-by-benjamin-m-schmidt/ [last accessed: April 2, 2021].

22 Shawn Graham/Ian Milligan/Scott Weingart, *Exploring Big Historical Data: The Historian's Macroscope*, London 2015, doi:10.1142/p981. A project using topic modeling reasonable on a large scale is: Impresso – Media Monitoring of the Past, URL: https://impresso-project.ch/) [last accessed: April 2, 2021].

23 See 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.

Eight years later, HTR entered the scene such that scholars could train and use specific models.

In the READ (short for "Recognition and Enrichment of Archival Documents") project, running from 2016 to 2019, scientists and scholars constructed a platform allowing them to train algorithms to recognize specific scripts with an error rate of around 3 percent.[24] This meant that, on the character level, out of 100 characters, 97 would be recognized correctly, including punctuation and space. Although the character error rate cannot be compared to a scholarly edition (typically around 0.1 to 0.2 percent character error rate), the result is still very much a legible and recognizable text that can be read, searched, and mined.[25]

Currently, we stand at the brink of training what are called general models that can recognize entire "styles" of handwriting, such as English in a Latin script of the 18th and 19th centuries or German current scripts of the 16th and 17th centuries. We thus can already conclude that deep learning can broaden access to historical material and helps us dive deeper into the subjects we're interested in. However, this comes at a price: By training and by selecting material for training, we strengthen machine learning algorithms, but we also bias them.

A very brief look at the architecture of the system of a deep learning algorithm, focused here on text recognition, explains the problem, since its success is based on training. In deep learning, we use systems modeled in ways similar to our understanding of the brain, consisting of a layered network of neurons. These cells are capable of either amplifying or reducing a signal coming from the preceding layer. The cells get "weighted" through training, meaning that the behavior of the cell (amplifying/reducing certain signals) will be optimized in a training process.

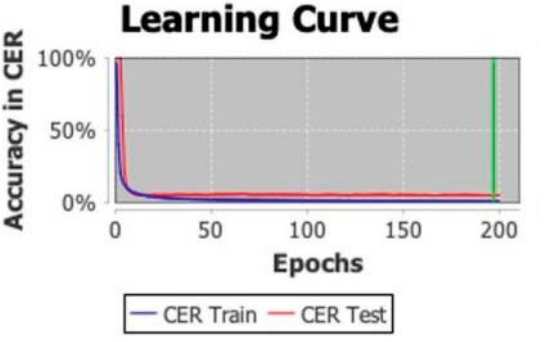
In the training process, the algorithm tries to align an input (in the case of text recognition, the image of a line of text) with a desired output (a string of characters). The training set is processed by the algorithm a certain number of times (called "epochs"), while optimizing the results towards the desired output. The main task of deep learning experts is thus not to know about the context of a certain problem like paleographical discussions for text recognition, but to decide on the number of layers, the (mathematical) optimizing processes, and the forms of decoding. This makes the technology very adaptable but also problematic.

---

24 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.

25 Since we discussed topic modeling earlier in the paper, it must be stated that the step from HTR to topic modeling is still quite difficult: Stephen Mutuvi et al., Evaluating the Impact of OCR Errors on Topic Modeling, in: *Lecture Notes in Computer Science*, 11279, Cham 2018, 3-14, doi:10.1007/978-3-030-04257-8_1.

Fig 6.3: Image from a learning curve, demonstrating how the neural networks reaches better results after each iteration. Own screenshot, made in Transkribus (text recognition platform and software): transkribus.eu.



If we train the algorithm on a set of pages from a specific hand, we will receive our first usable result after about 1,000 lines.[26] But, whatever decisions are made in the transcription process (for example, expanding abbreviations or the use of certain characters, such as long ſ vs. round s), the algorithm will "learn" the decisions implicitly. Even on the level of text recognition, this can become problematic, for example, in Latin scripts that tended to introduce a multitude of signs indicating abbreviation, signs that need to be expanded in context. One example would be "ꝝ" [Unicode A75D, Latin Small Letter Rum Rotunda], indicating the genitive plural ending of a word and resulting in different expansions depending on the gender (male/female/impersonal).[27] Since deep learning tends to rely on quantity, the instance most occurring will most probably be chosen. The inclusion or lack of a character will also lead to a different character set, making the recognition

26 For a more in-depth view on best practices (in German), see: Tobias Hodel, Best-Practices zur Erkennung alter Drucke und Handschriften – Die Nutzung von Transkribus large- und small-scale, in: Christof Schöch (ed.), *DHd 2020 Spielräume. Digital Humanities zwischen Modellierung und Interpretation*, Paderborn 2020, 84–87, doi:10.5281/zenodo.3666689 The recordings of recent Transkribus User Conferences might also give some insight: https://readcoop.eu/transkribus-user-conference-2018/ and https://readcoop.eu/transkribus-user-conference-2020/ [last accessed: April 2, 2021].

27 See also the talk by Estelle Gueville/David Wrisley, Rethinking the Abbreviation: Questions and Challenges of Machine Reading Medieval *Scripta*, in: *Dark archives 2020 conference*, URL: https://www.youtube.com/watch?v=p38lvPRRNmA [last accessed: April 2, 2021].

process more or less likely to succeed, since a minimal number of occurrences of a certain character is necessary to provide a reliable identification.[28]

With regard to text recognition, the consequences of such flawed output won't create too many problems. However, the issue is exacerbated if we use the same technology for things like named entity recognition that lead to interpretations involving questions of identity (what/who is a person) or typology (what is a text)—basically, everything that falls in the context of natural language processing.[29]

One of the main problems in dealing with these kinds of deep learning algorithms is the black-box that exists in the process of training. Since none of them is based on any (let's call it contextual) knowledge of the field or corpus in question, the algorithms are trained on a particular input-output or specific corpus. Accordingly, the resulting models are highly biased by the material they are trained on. In each training session, the algorithm "learns" to deal with a world consisting only of the training material. In Foucauldian terms, we could speak of an episteme that lays in the model and is different from the episteme of all other models.[30]

Let's look, for example, at a recognition model that has been trained on letters from the 15th century. The corpus presented consists of letters sent from the city council of Bern to a bailiff residing in nearby Thun, a small city under the dominion of Bern. As is often the case with medieval administrative writing, the content is quite formulaic and repetitive. The letters were written by different scribes, and we collected about 100 such letters, which we transcribed by hand and used as the material to train an HTR model.[31] Even when we apply the model on some random part of an image with no text written on it whatsoever (as far as I can tell), the HTR model gives us as a result parts of a typical letter, with the salutation and signature even appearing in the correct order. We could therefore conclude that the model is trained and keen to recognize what it was trained to "read." Of course, the example here is forced, and the preceding algorithm—the layout analysis—would not actually identify the part of the image with no text as a text region. The result is also striking due to its coupling with a generated vocabulary. Nonetheless, it

28 In addition, current HTR systems have been produced with alphabet languages in mind, leading to problems in decoding in the cases where not 100 characters but some thousand signs are part of the character set.

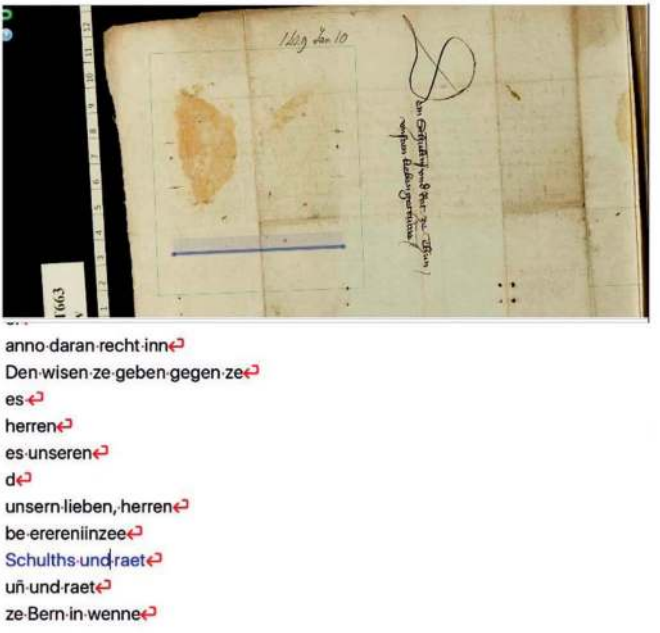
29 Tobias Hodel, Konsequenzen automatischer Texterkennung – Ein Aufriss zur Texterkennung mit machine learning, in: Vogeler, Georg (ed.), *DHd 2018. Kritik der digitalen Vernunft. Konferenzabstracts. Universität zu Köln, 26. Februar Bis 2. März 2018*, 249–51, Köln 2018, URL: http://dhd2018.uni-koeln.de/wp-content/uploads/boa-DHd2018-web-ISBN.pdf [last accessed: April 2, 2021].

30 Michel Foucault, *The Archaeology of Knowledge: And the Discourse on Language*, New York 1982.

31 The model was trained on 21,682 words and leads to a sub-par model with a character error rate on the validation set of 22.54 percent.

demonstrates how closely related the result of the recognition process is to the training set.

Fig 6.4: Screenshot from Transkribus, with the model recognizing (non-existent) text in the green square. The document transcribed is from the City Archives of Thun (BAT), BAT 663, 296v.



Since deep learning relies on large amounts of data to train specific models, we end up in a cycle of bias. In order to strengthen the model's ability to perform, in this case, text recognition, as much training data as possible is needed, usually gained from recognition processes that have undergone correction. The decisions of the annotators and the recognition process will thus fire back into the model and reinforce any problems. The whole issue becomes obvious if we look at the field's nomenclature. The term "Ground Truth" refers both to the material used to train a system and to the material used to evaluate the algorithm. The two sets are technically completely separated and not overlapping, with one subset used for training and another subset for evaluation. But both need to be prepared and checked by a human, thus the so-called Ground Truth is a quite far-reaching term for something influenced by someone (or rarely but ideally a group of people) who

needs to determine, in the case of text recognition, the correct transcription of a particular document.

What we see in the realms of deep learning is a reinforcement of bias introduced by the selection of training material. In a way, this issue is quite similar to what we observed in topic modeling, where the corpus at the beginning strongly influenced the calculated topics. Furthermore, deep learning is basically impossible to grasp as an algorithm. What happens in the neural network can only be observed; it is not possible to influence the process (save for creating a completely new alignment of the layers of the neural network). This barrier means that deep learning is in the difficult position that every model has to be checked for its bias and consequently for implicit and explicit problems through both input and output. Even epistemological questions must be brought up as part of the methodological discussion: What do we deem recognizable or sortable for an algorithm?

This brings us back to the initial question of what consequences machine learning approaches have for the humanities, looking at two very different approaches under the umbrella of Artificial Intelligence.

5. Addressing and (Not Yet) Solving Bias in Machine Learning: An Initial Conclusion

The most notable conclusion from the two approaches, text recognition and topic modeling, is they include—inevitably—bias. At the same time, the use of machine learning algorithms opens up a wide array of research possibilities that would have been unthinkable only ten years ago. For example, results from masses of handwritten documents were, before automatic recognition became a scalable process, accessible on only a very limited basis. And we only stand at the brink to grasp how we can use this new research tool.[32] Accessing masses of documents presorted by topic, or at least by clusters, opens up new ways to sift through material.

Topic modeling offers the opportunity to cluster together similar documents and extract characteristic tokens out of a cluster, allowing for an in-depth engagement with the documents as well as the method. In comparison, the controllability of topic modeling remains quite high, although the approach of unsupervised machine learning algorithms would seem to hint into another direction.

The case is quite the opposite if we look at the supervised deep learning approach, used here for handwritten text recognition. Although the model tries to adapt to and approximate a trained, desired output, the determining factors that

32 The consequences of the development of Google books are only starting to be recognized, see Lara Putnam, The Transnational and the Text-Searchable: Digitized Sources and the Shadows They Cast, in: *The American Historical Review* 121 (2/2016), 377–402, doi:10.1093/ahr/121.2.377.

lead to a specific output are unknown to the user and part of a black box consisting of artificial neurons defying interpretation.

Both approaches, supervised as well as unsupervised, insert forms of bias at different stages, whether through the corpus employed to build topics or through training material.

It is thus the connection between the capabilities of and the deeply embedded bias in machine learning that makes it a divided and torn approach in dire need of contextualization. Understanding the method—similar to the methods and theories of any discipline—is consequently a first step to interacting with its technological aspects critically and cracking open the parts of the method that are stored away in a black box.

The use of AI in scholarly endeavors is not only a way to yield results (be they recognized text or clustered topics), but moreover an intriguing means of engaging with a set of technologies that govern our everyday lives, including a multitude of processes we might be subjected to.

In conclusion, scholars need to treat machine learning methods as they would source material or research methods, by adding a layer of methodological critique to the research process. Alongside the publication of training and validation data (if possible), this will lead to a deeper level of interaction with a method that has been deemed inexplicable.

Algorithms and data should nonetheless be approached playfully, so that results of machine learning approaches are not taken too seriously, and data from the humanities can form a playground with a rich body of research for identifying and critiquing biased and thus problematic results.

Bibliography

Benjamin, Ruha, *Race After Technology: Abolitionist Tools for the New Jim Code*, Medford 2019.

Blair, Ann, *Too Much to Know: Managing Scholarly Information before the Modern Age*, New Haven 2010.

Blei, David M., Introduction to Probabilistic Topic Models, *Communication of the ACM*

Blei, David M./Ng, Andrew Y./Jordan, Michael I., Latent Dirichlet Allocation, in: *Journal of Machine Learning Research*

Bonjour, Edgar, *Der Sonderbundskrieg*, Zürich 1947.

Brown, Tom B., et al., Language Models Are Few-Shot Learners, in: *arXiv* [preprint], 22.07.2020, URL: arxiv:2005.14165 [last accessed: April 2, 2021].

Clancy, Kelly, Is the Brain a Useful Model for Artificial Intelligence?, in: *Wired* 19.05.2020, URL: https://www.wired.com/story/brain-model-artificial-intelligence/ [last accessed: April 2, 2021].

Devlin, Jacob, et al., BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding, in: *arXiv* [preprint], 24.05.2019, URL: arxiv:1810.04805 [last accessed: April 2, 2021].

Du Bois, Pierre, *La Guerre Du Sonderbund: La Suisse de 1847*, Neuchâtel 2018.

Foucault, Michel, *The Archaeology of Knowledge: And the Discourse on Language*

Fuhrer, Hans Rudolf/Loosli, Jean Paul/Moser, Christian, *Sonderbundskrieg 1847*, Wettingen 1997.

Gadamer, Hans-Georg, *Hermeneutik I: Wahrheit Und Methode: Grundzüge Einer Philosophischen Hermeneutik*, Tübingen 2010.

Graham, Shawn/Milligan, Ian/Weingart, Scott, *Exploring Big Historical Data: The Historian's Macroscope*, London 2015, doi:10.1142/p981.

Graham, Shawn/Weingart, Scott/Milligan, Ian, Getting Started with Topic Modeling and MALLET, in: *Programming Historian* 02.09.2012, URL: https://programminghistorian.org/en/lessons/topic-modeling-and-mallet [last accessed: April 2, 2021].

Gueville, Estelle/Wrisley, David, Rethinking the Abbreviation: Questions and Challenges of Machine Reading Medieval *Scripta*, in: *Dark archives 2020 conference*, URL: https://www.youtube.com/watch?v=p38lvPRRNmA [last accessed: April 2, 2021].

Hirschman, Dan, Doing Things with Bags-of-Words, in: *Scatterplot* 19.02.2020, URL: https://scatter.wordpress.com/2020/02/19/doing-things-with-bags-of-words/ [last accessed: April 2, 2021].

Hodel, Tobias, Best-Practices zur Erkennung alter Drucke und Handschriften – Die Nutzung von Transkribus large- und small-scale, in: Christof Schöch (ed.),

DHd 2020 Spielräume. Digital Humanities zwischen Modellierung und Interpretation, Paderborn 2020, 84–87, doi:10.5281/zenodo.3666689.

Hodel, Tobias, Konsequenzen automatischer Texterkennung – Ein Aufriss zur Texterkennung mit machine learning, in: Vogeler, Georg (ed.), *DHd 2018. Kritik der digitalen Vernunft. Konferenzabstracts. Universität zu Köln, 26. Februar Bis 2. März 2018*, 249–51, Köln 2018, URL: http://dhd2018.uni-koeln.de/wp-content/uploads/boa-DHd2018-web-ISBN.pdf [last accessed: April 2, 2021].

iRights lab, et al., Praxisleitfaden zu den Algo. Rules - Orientierungshilfen für Entwickler:Innen und ihre Führungskräfte, Gütersloh 2020, doi:10.11586/2020029.

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.

Mutuvi, Stephen, et al., Evaluating the Impact of OCR Errors on Topic Modeling, in: *Lecture Notes in Computer Science*, 11279, Cham 2018, 3-14, doi:10.1007/978-3-030-04257-8_1.

Noble, Safiya Umoja, *Algorithms of Oppression: How Search Engines Reinforce Racism*, New York 2018.

Putnam, Lara, The Transnational and the Text-Searchable: Digitized Sources and the Shadows They Cast, in: *The American Historical Review* 121 (2/2016), 377–402, doi:10.1093/ahr/121.2.377.

Rosenzweig, Roy, *Clio Wired: The Future of the Past in the Digital Age*, New York 2011.

Schmidt, Ben, Words Alone: Dismantling Topic Models in the Humanities, in: *Journal of Digital Humanities* 2 (1/2013), URL: http://journalofdigitalhumanities.org/2-1/words-alone-by-benjamin-m-schmidt/ [last accessed: April 2, 2021].

Schöch, Christof, Quantitative Analyse, in: Fotis Jannidis/Hubertus Kohle/Malte Rehbein (eds.), *Digital Humanities: Eine Einführung*, 279-298, Stuttgart 2017, doi:10.1007/978-3-476-05446-3_20.

Schöch, Christof, Topic Modeling Genre: An Exploration of French Classical and Enlightenment Drama, in: *Digital Humanities Quarterly* 11/2 (2017), URL: http://www.digitalhumanities.org/dhq/vol/11/2/000291/000291.html [last accessed: April 2, 2021].

Spelke, Elizabeth, It's Called Artificial Intelligence—but What Is Intelligence?, in: *Wired* 19.05.2020, URL: https://www.wired.com/story/its-called-artificial-intelligence-but-what-is-intelligence/ [last accessed: April 2, 2021].

Underwood, Ted, Theorizing Research Practices We Forgot to Theorize Twenty Years Ago, *Representations* 127 (1/2014), 64–72, doi:10.1525/rep.2014.127.1.64.

Underwood, Ted/Jockers, Matthew L., Text-Mining the Humanities, in: Susan Schreibman/ Ray Siemens/John Unsworth, *A New Companion to Digital Humanities*, Chichester 2016, 291–306.

Weingart, Scott, Topic Modeling for Humanists: A Guided Tour, in: *The scottbot irregular* 25.07.2012, URL: http://www.scottbot.net/HIAL/?p=19113 [last accessed April 2, 2021].

Xu, Kang, et al., Incorporating Wikipedia Concepts and Categories as Prior Knowledge into Topic Models, in: *Intelligent Data Analysis* 21 (2/2017), 443-461, doi:10.3233/IDA-160021.

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:

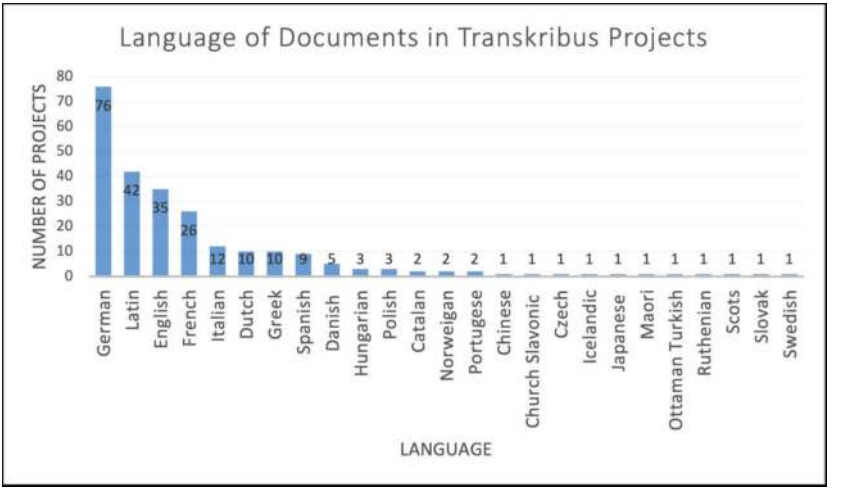
giving confidence in its coverage. There were active logins from 53 countries during that period, including Australia (1%), and fewer than 1% of users from Russia, Chile, Côte d'Ivoire, Cocos Islands, South Africa, India, People's Republic of China, Argentina, Iraq, Turkmenistan, El Salvador, Laos, and 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.

AFTERWORD: Towards a new Discipline of Computational Archival Science (CAS)

Richard Marciano, University of Maryland

As an afterword to this timely edited collection on Archives, Access and AI, I thought it might be helpful to amplify and extend some of the salient threads that were broached on the intersection of technology and archives.

In her introductory editorial chapter, Lise Jaillant captures three main challenges faced by cultural heritage organizations: (1) dealing with scale, (2) unlocking "dark" archives, and (3) addressing the skills gap in data science and AI. In the process, she invites contributions that not only showcase compelling interdisciplinary case studies but also summon theoretical insights.

Let me expand on these three challenges, with a concrete example from my own teaching of computational techniques to library and information science graduate students. I will make this point by highlighting a specific cultural collection: a single digitized Historical City Directory ("Post Office Directories" in the UK) for the city of Charlotte, North Carolina. City Directories are an important source of genealogical information as they were often published annually. They also supplement Census data and other local records.

1. Revisiting the Computational Challenges Faced by Cultural Heritage Organizations

a. Dealing with scale

Scale in digitized cultural heritage collections should no longer come as a surprise. The city of Charlotte (North Carolina) 1911 Historical City Directory book, as scanned and OCR-ed by the Internet Archive, yields close to 2GB of data (Charlotte itself comprises a timeseries of 62 Directories spanning an 89-year period). Directories for the entire state of North Carolina cover over 100 cities, spanning over a 100-year period (from 1860 to 1969), with close to 1,000 directories in aggregate. This represents up to 2TB (Terabytes) of digital content for the state of North Carolina alone. A rough extrapolation to the entire United States, potentially leads to 100TB of data [or two hundred 500GB hard-drives], thus merely an order of magnitude under a 1PB (Petabyte) of data. Hence, **cultural**

collections are inherently "big data." When interconnecting city directories to intersecting historical collections such as Sanborn Fire Insurance Maps, Census data, vital records, redlining data, etc., we very quickly enter the Petabyte (PB) range.

Fig 8.1: 1911 City Directory for Charlotte, North Carolina (screen snapshot from the Internet Archive).

It is in this context of scale that scalability comes into play, or the ability of archival systems to handle a growing amount of information and processing. This emphasizes how the methods that might apply to small archival holdings may not be applicable to very large holdings: "Entrat" the conversation on *Applying AI to Archives*. Applications of AI and ML to archival collections are beginning to emerge, but as Lise Jaillant highlights there is still "a lack of compelling case studies" in this space.

b. Unlocking "dark" archives

If we accept the premise of cultural collections as "big data," the next natural step is to consider the use of computational treatments to unlock "dark" cultural archives. The challenge is to see if and how we can harness the best and latest advances in data science and demonstrate its usefulness and applicability to digital cultural assets.

It is worth reflecting on the term "dark archives" which is used throughout this book. In archival literature, dark archives have a specific designation. Per the

SAA Dictionary of Archives Terminology[1] they signify either "a repository that stores archival resources for future use but is accessible only to its custodian" or "a collection of materials preserved for future use but with no current access." The US National Archives 2017 Digital Strategy[2] plan for instance, discusses a digital records infrastructure capable of safely and securely preserving several Petabytes of data in their tape-based Dark Archive, with associated descriptive metadata. There are even gradations in the literature, introducing "light archives" and "dim archives", indicating intermediate levels of access. The way "dark" archives appear to be used in this book is in the context of using AI to improve accessibility.

Back to our example, we illustrate data science driven approaches to unlocking "dark" archives by interrogating the Internet Web Archive **[relates to 2. Bell web archives paper]**, and moving the 1911 Historical City Directory pages of Charlotte, North Carolina through two processing pipelines: (1) datafication [the top part of Fig 8.2], and (2) data analysis [the bottom part of Fig 8.2].

Fig 8.2: 1911 City Directory for Charlotte, North Carolina (screen snapshot from Richard Marciano's class syllabus).

For datafication, students are asked to go inside and steer what are too often considered "black box" processes **[relates to 8. Gooding black box paper]**

1 Society of American Archivists, Definition of Dark Archives, URL: https://dictionary.archivists.org/entry/dark-archives.html [last accessed: April 5, 2021].

2 The National Archives (UK), Digital Strategy, 03.2017, URL: https://www.nationalarchives.gov.uk/documents/the-national-archives-digital-strategy-2017-19.pdf [last accessed: April 5, 2021].

including: (1) digitization (image to unstructured text, i.e. the Optical Character Recognition ABBYYFineReader tool), (2) cleaning & transforming (unstructured to structured text, i.e. the data wrangling OpenRefine and Trifacta tools), and (3) text processing (Natural Language Processing/Named Entity Recognition text tagging, i.e. GATE/ANNIE NLP/NER tool).

For data analysis, the resulting enhanced structured text is ready to be: (4) represented spatially through the creation of maps (i.e. QGIS geographical information system tool), (5) visualized interactively through the creation of analytics dashboards (i.e. Tableau data analytics tool), and (6) modeled through social networks (i.e. NoSQL Neo4j graph database). AI and ML are experienced through steps 2. and 3. We also contrast printed and handwritten text extraction approaches **[relates to 6. Hodel HTR paper and 7. Terras HTR survey paper].**

This two-phased processing pipeline is meant to provide experiential learning pathways and demonstrate the meaning of unlocking "dark archives" through the creation of an iterative "Archives, Access and Artificial Intelligence" automation workflow **[relates to 3. Jaillant design thinking paper]**

What seems equally important is to train students to think beyond "dark archives" as well and give them exposure to "dark AI" **[relates to Lise Jaillant's discussion on the "Threat of Dark AI"]**. AI cannot be examined in isolation and needs to be contextualized within the entire records management. A striking illustration, is provided by Dr. Lyneise Williams, founder of the VERA Collaborative (Visual Electronic Representations in the Archive).[3] She provides a compelling case study of how the use of photograph digitization in particular can amplify marginalization or erasure, whether through limitations in the original source documents or limitations within the technologies. The latter may unintentionally obscure visual and written features, especially those related to race, gender, and/or class. Williams offers an art historical perspective on this phenomenon, demonstrating that technical limitations can lead to erasure and distortion of archival records involving underrepresented and/or marginalized communities.[4] If marginalized people are being erased from historical records, there is not much hope for AI/ML to change these outcomes. An open challenge that arises from this work is how AI and ML

3 See https://veracollaborative.com [last accessed: April 5, 2021].

4 Lyneise Williams, What Computational Archival Science Can Learn from Art History and Material Culture Studies, 12.12.2019, in: *2019 IEEE International Conference on Big Data*, Los Angeles, CA, URL: https://ai-collaboratory.net/wp-content/uploads/2020/02/Williams.pdf [last accessed: April 5, 2021].

approaches might help uncover hidden knowledge and/or mitigate erasures within archival collections related to racial erasure.[5]

c. Addressing the skills gap in data science and AI

While this book emphasizes training humanities researchers in quantitative and computational techniques, there are two other significant dimensions I would like to highlight: (a) Establishing a framework for thinking computationally when working with digital archives, and (b) Developing interdisciplinary collaboration team building skills:

- In a recently funded IMLS Symposium grant called CT-LASER, we explored developing a Framework for Mapping Computational Thinking (CT) to Library and Archival Science Education & Research (LASER)[6], explicitly using a set of computational practices covering: (1) data, (2) modeling and simulation, (3) computational problem solving, and (4) systems thinking. CT is a form of problem solving that uses modeling, decomposition, pattern recognition, abstraction, algorithm design, and scale.[7] We provided a summary of these twenty-two CT practices spread across these four practice verticals, and we demonstrated the remapping of these concepts to archival science.[8]

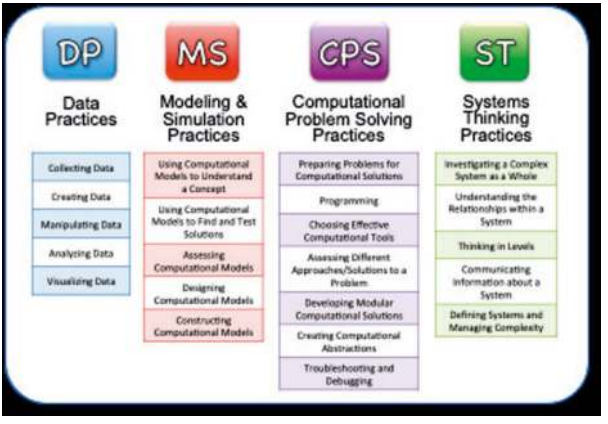
5 Lori A. Perine et al., Computational Treatments of the Legacy of Slavery (CT-LoS) "Reasserting Erased Memory," 12.12.2020, in: *2020 IEEE International Conference on Big Data*, Atlanta, in: https://ai-collaboratory.net/wp-content/uploads/2020/11/Perine.pdf [last accessed: April 5, 2021].

6 CT-LASER, final report, 01.10.2020, URL: https://ai-collaboratory.net/wp-content/uploads/2020/11/Final_Report_r.pdf [last accessed: April 5, 2021].

7 Jeannette M. Wing, Computational Thinking, in: *Communications of the ACM*, 49 (3/2006), 33–35, URL: https://www.cs.cmu.edu/ 15110-s13/Wing06-ct.pdf [last accessed: April 5, 2021].

8 Richard Marciano et al., Reframing Digital Curation Practices through a Computational Thinking Framework, 11.12.2019, in: *2019 IEEE International Conference on Big Data*, Los Angeles, CA, URL: https://ai-collaboratory.net/wp-content/uploads/2020/04/ReframingDC-UsingCT_final.pdf [last accessed: April 5, 2021].

Fig 8.3: Computational thinking taxonomy now mapped to working with digital archives (Screen snapshot from CT-LASER workshop talk, Apr. 2019).



More fundamentally, the project is addressing the integration of 'computational thinking' and 'archival thinking', as record-keeping innovation and technological development can only progress hand in hand. There is a need to accelerate opportunities for knowledge exchange and interdisciplinary synergies that will enable the infusion of archival concepts, principles, theories and methods with the computational and vice versa.[9]

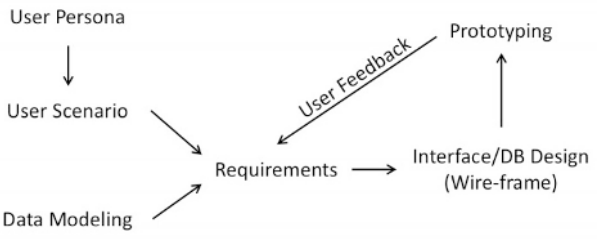
- At the University Maryland iSchool, from 2015 to 2020, I co-directed a digital curation innovation center, predicated on teaching students to work in interdisciplinary teams. During this period, we mentored over 300 students, across dozens of cultural and infrastructure projects, with a particular focus on big records and archival analytics with a mix of students with diverse backgrounds (humanities, information management, human computer interface,

9 William Underwood et al., Introducing Computational Thinking into Archival Science Education, in: *Proceedings of IEEE Big Data Conference 2018*, CAS Workshop, Seattle, WA, see 2761-2765, URL: https://ai-collaboratory.net/wp-content/uploads/2020/03/1.Underwood.pdf [last accessed: April 5, 2021].

library and archives.)[10-11] We observed that the exposure to learning how to work collaboratively in interdisciplinary teams was an indispensable skill that needed to be nurtured and developed early on.[12] The book chapter on Photoarchives **[relates to 1. Han AI applied to Photoarchives paper]** eminently illustrates the diversity of AI and Cultural Data collaborative teams, as it features five researchers with backgrounds in operations research and information engineering, mathematics, digital art history, statistics, and electrical and computer engineering.

In addition, and in support of Lise Jaillant's chapter on design thinking, our collaborative training approach emphasized an iterative design process in which ideation, prototyping, and testing are central. This relates to agile development and its iterative benefits that seem to be a natural fit with for how cultural materials are interrogated.

Fig 8.4: Iterative design process used in the Human Face of Big Data.



10 Student-Led "Datathon" Exploring Data, Investigating Methodologies, 28-29.10.2019, URL: https://ai-collaboratory.net/projects/ct-los/student-led-datathon-at-the-maryland-state-archives/ [last accessed: April 5, 2021].

11 Resistance at Tule Lake: A Conversation with the Filmmaker and iSchool Digital Curators (and Film Viewing), URL: https://ai-collaboratory.net/projects/ct-ja_ww2_camps/digital-curation-students-and-filmmaker-event/ [last accessed: April 5, 2021].

12 P. Nicholas et al., Establishing a Research Agenda for Computational Archival Science through Interdisciplinary Collaborations between Archivists and Technologists, in: *SAA 2020 Research Forum* (accepted for publication).

2. Towards a New Discipline of Computational Archival Science (CAS)

We posit the emergence of a new praxis we call *Computational Archival Science*. No one would dispute at this point in time the legitimacy of the fields of *Computational Social Science* ("Investigating social and behavioral relationships and interactions through: social simulation, modeling, network analysis, and media analysis"[13]), and *Computational Biology* ("The science of using biological data to develop algorithms or models to better understand biological systems"[14] Wikipedia). The latest addition to this computational turn may be *Computational Journalism* ("Finding and telling news stories, WITH, BY, or ABOUT algorithms"[15]).

For decades, archivists have been appraising, preserving, and providing access to digital records by using archival theories and methods developed for paper records. However, production and consumption of digital records are informed by social and industrial trends and by computer and data methods that show little or no connection to archival methods. As a matter of investigation, we have been exploring the foundations of CAS for the last five years. We captured this inquiry in a foundational paper that discusses the need to reexamine the theories and methods that dominate records practices, where we felt that this situation called for a formal articulation of a new trans-discipline, which we called *Computational Archival Science* (CAS).[16]

In this paper, our *working definition of CAS* is:

> A transdisciplinary field concerned with the application of computational methods and resources to large-scale records/archives processing, analysis, storage, long-term preservation, and access, with the aim of improving efficiency, productivity, and precision in support of appraisal, arrangement and description, preservation, and access decisions.
>
> The intent is to engage and undertake research with archival materials as well as apply the collective knowledge of computer and archival science to understand the ways that new technologies change the generation, use, storage, and preservation of records and the implications of these changes for archival functions and

13 https://en.wikipedia.org/wiki/Computational_social_science [last accessed: April 5, 2021].

14 https://en.wikipedia.org/wiki/Computational_biology [last accessed: April 5, 2021].

15 Nicholas Diakopoulos, Cultivating the Landscape of Innovation in Computational Journalism, CUNY Whitepaper, 04.2012, URL: http://cdn.journalism.cuny.edu/blogs.dir/418/files/2012/04/diakopoulos_whitepaper_systematicinnovation.pdf [last accessed: April 5, 2021].

16 Richard Marciano et al., *Archival Records and Training in the Age of Big Data*, in: J. Percell et al. (eds.), *Re-Envisioning the MLS: Perspectives on the Future of Library and Information Science Education*, Somerville, MA, 2018, 179-199, URL: https://ai-collaboratory.net/wp-content/uploads/2020/10/Marciano-et-al-Archival-Records-and-Training-in-the-Age-of-Big-Data-final.pdf [last accessed: April 5, 2021].

the societal and organizational use and preservation of authentic digital records. This suggests that computational archival science is a blend of computational and archival thinking.

Archival Concepts	Computational Methods
Going from paper catalog entries to digital catalogs, Matching records in distributed databases	Graph and Probabilistic Databases
Technology assisted review accessibility of presidential and federal e-mail accessioned into National Archives	Analytics, predictive coding to address PII
Provenance in terms of why, who and how	Abstraction and ontology construction
Appraisal	File Format Characterization, File Format policies, Bulk extractor (Identifies PII), Content Preview, Tagging
Classification of archival images	AI, Line detection, image segmentation
Recordkeeping	Auto-categorization, auto-classification, e-discovery, machine learning
Personally Identifiable Information (PII)	NLP, NER, sentiment analysis
Structured data interfaces to archival materials	APIs for cultural heritage materials, graph databases
Decentralized recordkeeping	Blockchain, secure computing, trustworthiness

This approach resonates with Lise Jaillant's discussion of *"AI for Good,"* where she highlights the value of developing AI in the context of archival principles including: respect des fonds, appraisal, authenticity, and original order.

We continue to explore the mapping of archival concepts to computational methods. Papers presented at our second 2017 CAS Workshop provided evidence of the following connections:

For more information on the body of work emerging from this CAS initiative, we invite the reader to explore our CAS portal (https://ai-collaboratory.net/cas), which now features five international IEEE Big Data workshops, over 30 workshops since 2016, and over 50 research papers and presentations.

In "Computational Thinking in Archival Science Research and Education,"[17] Bill Underwood examines noteworthy archival research projects and describes how we were able to identify instances of all twenty-two CT Practices from Fig 8.3.

17 William Underwood/Richard Marciano, Computational Thinking in Archival Science Research and Education, 11.12.2019, in: *2019 IEEE International Conference on Big Data*, Los Angeles, CA, in: https://ai-collaboratory.net/wp-content/uploads/2021/03/Underwood.pdf [last accessed: April 5, 2021].

3. On the Need to Create a Network of Practitioners and Scholars in CAS

In the context of a 2019-2020 AHRC-funded International Research Collaboration Network in Computational Archival Science (IRCN-CAS) between the U. Maryland, King's College London, the Maryland State Archives, and The National Archives (UK),[18] we further observed that: (1) The new ways in which the public and researchers wish to engage with archival materials, are disrupting to traditional archival theories and practices, (2) The application of computational methods and tools to the archival problem space needs to be further explored, and (3) The contextualization of records also needs to be explored, whether through: capturing metadata, enhancing records by semantic tagging, and linking records with other records.

This led us to conclude that the way forward would benefit from establishing an international computational network for librarians and archivists.[19] This prompted us to launch the AIC Collaboratory at The Alan Turing Institute in London, UK on January 20, 2020 at the CAS Symposium held there, bringing together partners from leading academic and cultural institutions from six continents, with explicit goals to: (1) EXPLORE the opportunities and challenges of "disruptive technologies" for archives and records management (digital curation, machine learning, AI, etc.), (2) LEVERAGE the latest technologies to unlock the hidden information in massive stores of records, (3) PURSUE multidisciplinary collaborations to share relevant knowledge across domains, (4) TRAIN current and future generations of information professionals to think computationally and rapidly adapt new technologies to meet their increasingly large and complex workloads, and PROMOTE ethical information access and use.

4. On the Need to Pilot a Collaborative Network for Integrating Computational Thinking into Library and Archival Education and Practice

Finally, in response to the Editor's comment on the "lack of compelling case studies" and the need to develop real-world examples within the academic or professional literature, we conclude this afterword on a collaborative case study note, inviting

18 https://computationalarchives.net/ [last accessed: March 23, 2021].

19 Richard Marciano et al., Establishing an International Computational Network for Librarians and Archivists, in: *iConference 2019 Blue Sky Papers series*, URL: http://hdl.handle.net/2142/103139 [last accessed: April 5, 2021].

the readers of this book to consider joining forces on an initiative meant to address these gaps.[20]

To support the development of shared and reusable case studies, AIC researchers recently launched a 2-year IMLS-funded grant to pilot an online collaborative network of educators and practitioners to enable the sharing and dissemination of computational case studies and lesson plans through a Jupyter Notebook interactive computational learning platform, called *CASES* [Computational Archival Science Educational System], see: https://cases.umd.edu.

This virtual network is really a network of networks with seventeen collaborators dedicated to mapping Computational Thinking to Archival and Library practices. This Network includes: (1) a *Core Network (CN)* of seven experts in digital archives, lesson plan evaluation, project management, computational thinking, library software integration, and ethics and representation in digital collections, (2) an *Educator Network (EN)* of four educators from MLIS programs (at all ranks), and (3) a *Practitioner Network (PN)* of seven librarians / archivists representing four diverse and under-represented American collections of African-, Asian-, and Puerto Rican -American lineage: (a) the Maryland State Archives *Legacy of Slavery Project*, (b) the Spelman College *Department of Drama and Dance Photographs*, (c) Densho's *WWII Japanese American Camps Collections*, and (d) the 2019 *Puerto Rican Summer Protests ("RickyRenuncia")*. We are calling this cluster of *Practitioner Network* collections ***"Representing America,"*** to emphasize its significance and impact of training future MLIS students and exposing them to the full diversity of the American experience. In addition, we will seek feedback from an *Advisory Network (AN)* consisting of: (1) five US experts [three Practitioners at Cultural Institutions: Smithsonian National Museum of American History, Harvard Library, the US Holocaust Memorial Museum, and *two iSchool Educators* from UCLA and Drexel], and (2) International experts from all six continents. This pilot network will lead to the publication of shared, interactive, and reusable case studies which will include AI/ML exemplars, and will need to be extended and sustained through larger networks of practitioners and scholars.

To accelerate the development of case studies in Archives using AI and ML in particular, we have launched a FARM Initiative on the Future of Archives and Records Management (see: https://ai-collaboratory.net/details-aic-farm-initiative/) which seeks to develop Jupyter Notebook-based additions to the CASES repository.

---

20 Piloting an Online National Collaborative Network for Integrating Computational Thinking into Library and Archival Education and Practice, URL: https://www.imls.gov/sites/default/files/project-proposals/re-246334-ols-20-full-proposal.pdf [last accessed: April 5, 2021].

Fig 8.5: The CASES website showing a carousel of notebooks for browsing (website screen snapshot).

It is through all of these types of community intervention that we believe rapid and meaningful progress will be achieved in creating enhanced digital scholarship predicated on the integration of archives, access, and AI.

Bibliography

CT-LASER, final report, 01.10.2020, URL: https://ai-collaboratory.net/wp-content/uploads/2020/11/Final_Report_r.pdf [last accessed: April 5, 2021].

Diakopoulos, Nicholas, Cultivating the Landscape of Innovation in Computational Journalism, CUNY Whitepaper, 04.2012, URL: http://cdn.journalism.cuny.edu/blogs.dir/418/files/2012/04/diakopoulos_whitepaper_systematicinnovation.pdf [last accessed: April 5, 2021].

Lee, Myeong, et al., Heuristics for Assessing Computational Archival Science (CAS) Research: The Case of the Human Face of Big Data Project, 12.2017, in: *IEEE Big Data 2017*, Boston, MA, see 2262-2270, URL: https://ai-collaboratory.net/wp-content/uploads/2020/04/Myeong_Lee.pdf [last accessed: April 5, 2021].

Marciano, Richard, et al., *Archival Records and Training in the Age of Big Data*, in: J. Percell et al. (eds.), *Re-Envisioning the MLS: Perspectives on the Future of Library and Information Science Education*, Somerville, MA, 2018, 179-199, URL: https://ai-collaboratory.net/wp-content/uploads/2020/10/Marciano-et-al-Archival-Records-and-Training-in-the-Age-of-Big-Data-final.pdf [last accessed: April 5, 2021].

Marciano, Richard, et al., Reframing Digital Curation Practices through a Computational Thinking Framework, 11.12.2019, in: *2019 IEEE International Conference on Big Data*, Los Angeles, CA, URL: https://ai-collaboratory.net/wp-content/uploads/2020/04/ReframingDC-UsingCT_final.pdf [last accessed: April 5, 2021].

Perine, Lori A., et al., Computational Treatments of the Legacy of Slavery (CT-LoS) "Reasserting Erased Memory," 12.12.2020, in: *2020 IEEE International Conference on Big Data*, Atlanta, in: https://ai-collaboratory.net/wp-content/uploads/2020/11/Perine.pdf [last accessed: April 5, 2021].

Piloting an Online National Collaborative Network for Integrating Computational Thinking into Library and Archival Education and Practice, URL: https://www.imls.gov/sites/default/files/project-proposals/re-246334-ols-20-full-proposal.pdf [last accessed: April 5, 2021].

Resistance at Tule Lake: A Conversation with the Filmmaker and iSchool Digital Curators (and Film Viewing), URL: https://ai-collaboratory.net/projects/ct-ja_ww2_camps/digital-curation-students-and-filmmaker-event/ [last accessed: April 5, 2021].

Society of American Archivists, Definition of Dark Archives, URL: https://dictionary.archivists.org/entry/dark-archives.html [last accessed: April 5, 2021].

Student-Led "Datathon" Exploring Data, Investigating Methodologies, 28-29.10.2019, URL: https://ai-collaboratory.net/projects/ct-los/student-led-datathon-at-the-maryland-state-archives/ [last accessed: April 5, 2021].

The National Archives (UK), Digital Strategy, 03.2017, URL: https://www.nationalarchives.gov.uk/documents/the-national-archives-digital-strategy-2017-19.pdf [last accessed: April 5, 2021].

Underwood, William, et al., Introducing Computational Thinking into Archival Science Education, in: *Proceedings of IEEE Big Data Conference 2018*, CAS Workshop, Seattle, WA, see 2761-2765, URL: https://ai-collaboratory.net/wp-content/uploads/2020/03/1.Underwood.pdf [last accessed: April 5, 2021].

Williams, Lyneise, What Computational Archival Science Can Learn from Art History and Material Culture Studies, 12.12.2019, in: *2019 IEEE International Conference on Big Data*, Los Angeles, CA, URL: https://ai-collaboratory.net/wp-content/uploads/2020/02/Williams.pdf [last accessed: April 5, 2021].

Wing, Jeannette M., Computational Thinking, in: *Communications of the ACM*, 49 (3/2006), 33–35, URL: https://www.cs.cmu.edu/ 15110-s13/Wing06-ct.pdf [last accessed: April 5, 2021].

Authors (by order of appearance in the volume)

Lise Jaillant is Senior Lecturer (Associate Professor) in Digital Humanities at Loughborough University, UK. She is currently leading two externally-funded international networks on artificial intelligence applied to digital archives: the UK/ Irish network AURA (www.aura-network.net) and the UK/ US network AEOLIAN (www.aeolian-network.net). For more information, see: www.lisejaillant.com

X.Y. Han is a PhD Student in Operations Research and Information Engineering at Cornell University, USA. He has been conducting research analyzing the mathematical geometry of deep neural networks.

Vardan Papyan is an assistant professor in the department of mathematics at the University of Toronto, Canada, cross-appointed with the department of computer science. He completed his postdoctoral studies in the department of statistics at Stanford University, and his PhD in the department of computer science at the Technion – Israel Institute of Technology.

Ellen Prokop is an image specialist at the National Gallery of Art, Washington, DC, USA. She previously served as Digital Art History Lead at the Frick Art Reference Library, New York.

David L. Donoho is Professor of Statistics at Stanford University, USA, with research interests in harmonic analysis, signal processing, deep learning and compressed sensing.

C. Richard Johnson, Jr. is the Geoffrey S. M. Hedrick Senior Professor of Engineering Emeritus at Cornell University, USA. His research since 2007 has focused on

computational art history, primarily in matching patterns in art supports to find rollmates among paintings on canvas and moldmates among prints on laid paper.

Mark Bell is Senior Digital Researcher at The National Archives, UK. His current research interests include: the application of machine learning in the archive, particularly in the UK Government Web Archive, the use of Handwritten Text Recognition for exploring digitised collections at scale, and the implications of machine learning including Explainable AI and representing uncertainty.

Tom Storrar is Head of Web Archiving at The National Archives, UK.

Jane Winters is Professor of Digital Humanities at the School of Advanced Study, University of London, UK. Her research interests include digital history, born-digital archives (particularly the archived web), the use of social media by cultural heritage institutions, and open access publishing.

Paul Gooding is Senior Lecturer in Information Studies at the University of Glasgow, UK. His research focuses on evaluating the impact of digital library collections on institutions and users, and how library and archival collections can be harnessed for innovative reuse in the Digital Humanities.

Martin Paul Eve is Professor of Literature, Technology and Publishing at Birkbeck, University of London, UK. He specialises in contemporary American fiction, histories and philosophies of technology, and technological mutations in scholarly publishing. He is well-known for his work on open access and HE policy.

Robert Gadie is a PhD student at the University of the Arts, London, UK, whose doctoral research focuses on the policy implications of artists' epistemological practice.

Victoria Odeniyi is a Post-doctoral Research Fellow at the University of the Arts, London.

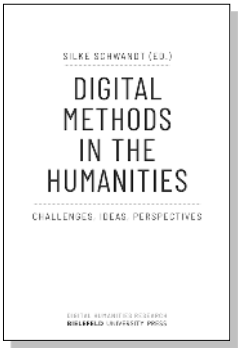
Shahina Parvin is a Post-doctoral Research Fellow in the Department of Sociology, Brandon University, Canada and Assistant Professor in Anthropology (on study leave) at Jahangirnagar University, Bangladesh.

Tobias Hodel is Assistant Professor in Digital Humanities at the University of Bern, Switzerland. His research interests include the theory of the digital humanities, machine learning in the humanities and critical algorithm studies.

Melissa Terras is Professor of Digital Cultural Heritage at the University of Edinburgh, UK. Her research interest is the digitisation of cultural heritage, including advanced digitisation techniques, usage of large-scale digitisation, and the mining and analysis of digitised content.

Richard Marciano is a professor in the College of Information Studies at the University of Maryland. His research interests center on digital curation, digital preservation, sustainable archives, cyberinfrastructure, and big data. He is the founder of the Advanced Information Collaboratory (AIC), https://ai-collaboratory.net/

Bielefeld University Press

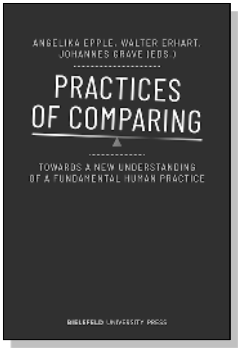


Silke Schwandt (ed.)

Digital Methods in the Humanities

Challenges, Ideas, Perspectives

2020, 312 p., pb., col. ill.
38,00 € (DE), 978-3-8376-5419-6
E-Book: available as free open access publication
PDF: ISBN 978-3-8394-5419-0

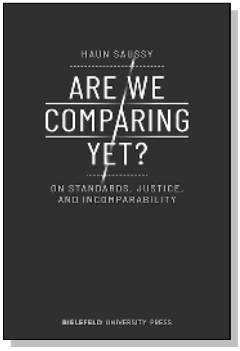


Angelika Epple, Walter Erhart, Johannes Grave (eds.)

Practices of Comparing

Towards a New Understanding of a Fundamental Human Practice

2020, 406 p., pb., col. ill.
39,00 € (DE), 978-3-8376-5166-9
E-Book: available as free open access publication
PDF: ISBN 978-3-8394-5166-3



Haun Saussy

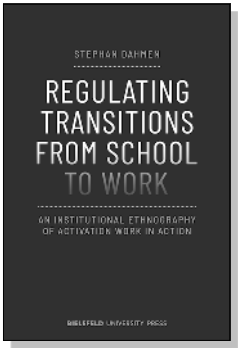
Are We Comparing Yet?

On Standards, Justice, and Incomparability

2019, 112 p., pb.
19,99 € (DE), 978-3-8376-4977-2
E-Book: available as free open access publication
PDF: ISBN 978-3-8394-4977-6

All print, e-book and open access versions of the titles in our list are available in the online shop www.bielefeld-university-press.de

Bielefeld University Press

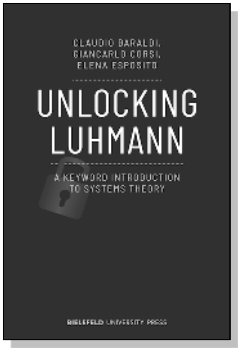


Stephan Dahmen

Regulating Transitions from School to Work

An Institutional Ethnography of Activation Work in Action

June 2021, 312 p., pb., ill.
36,00 € (DE), 978-3-8376-5706-7
E-Book: available as free open access publication
PDF: ISBN 978-3-8394-5706-1

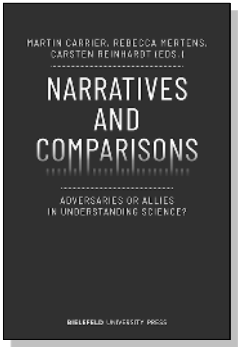


Claudio Baraldi, Giancarlo Corsi, Elena Esposito

Unlocking Luhmann

A Keyword Introduction to Systems Theory

April 2021, 276 p., pb.
40,00 € (DE), 978-3-8376-5674-9
E-Book: available as free open access publication
PDF: ISBN 978-3-8394-5674-3



Martin Carrier, Rebecca Mertens, Carsten Reinhardt (eds.)

Narratives and Comparisons

Adversaries or Allies in Understanding Science?

January 2021, 206 p., pb., col. ill.
35,00 € (DE), 978-3-8376-5415-8
E-Book: available as free open access publication
PDF: ISBN 978-3-8394-5415-2

All print, e-book and open access versions of the titles in our list are available in the online shop www.bielefeld-university-press.de