250words

Abstract:

Computing and the use of digital sources and resources is an everyday and essential practice in current academic scholarship. The present article gives a concise overview of approaches and methods within digital historical scholarship, focussing on the question: *How have the Digital Humanities evolved and what has that evolution brought to historical scholarship?* We begin by discussing techniques in which data are generated and machine searchable, such as OCR/HTR, born-digital archives, computer vision, scholarly editions, and Linked Data. In the second section, we provide examples of how data is made more accessible through quantitative text and network analysis. We close with a section on the need for hermeneutics and data-awareness in digital historical scholarship

The technologies described in this article have had varying degrees of effect on historical scholarship, usually in indirect ways. For example, technologies such as OCR and search engines may not be directly visible in a historical argument; however, these technologies do shape how historians interact with sources and whether sources can be accessed at all. It is with this article that we aim to start to take stock of the digital approaches and methods used in historical scholarship which may serve as starting points for scholars to understand the digital turn in the field and how and when to implement such approaches in their work.

State of the field: Digital History

C. Annemieke Romein, Max Kemman, Julie M. Birkholz, James Baker, Michel de Gruijter,

Albert Meroño-Peñuela, Thorsten Ries, Ruben Ros, Stefania Scagliola

The use of computers in historical scholarship is not new, although the impact on the field has shifted over time.[1] Notably, the 1960s saw the rise of quantitative history, often referred to as *cliometrics*, where historians used mainframe computers for statistical analysis. During the 1980s, the discipline lost its enthusiasm for quantitative histories, which had strayed too far from the traditional questions and methods of history.[2] The rise of personal computers, word processing software and relational databases for enabling qualitative research throughout the 1980s led to a new wave of work called 'history and computing', gaining traction in the mid-1990s.[3] The emergence of the Web in the 1990s also afforded digital projects such as one of the first online-first historical publications: *The Valley of the Shadow*.[4] These new digital projects where the historical narrative was combined with the expanded possibilities of digital technology, including scans of historical sources and non-linear narratives, gave rise to the term 'digital history'. Digital history, as such, has origins both in quantitative approaches to the historical record, as well as in the qualitative approaches born out of this 'cultural turn'.[5]

[1] The authors wish to thank the peer reviewers and editors for their constructive comments and reflections on draft versions of this article. All errors of judgement or fact remain our own.

[2] John F. Reynolds, 'Do Historians Count Anymore?: The Status of Quantitative Methods in History, 1975–1995', *Historical Methods: A Journal of Quantitative and Interdisciplinary History* 31, no. 4 (1 January 1998): 141–48, https://doi.org/10.1080/01615449809601196.

[3] Onno Boonstra, Leen Breure, and Peter Doorn, 'Past, Present and Future of Historical Information Science', *Historical Social Research / Historische Sozialforschung* 29, no. 2 (108) (2004): 4–132, https://www.jstor.org/stable/20761957.

[4] Still online at http://valley.lib.virginia.edu/, accessed 18 October 2019

[5] Although we trace digital history back to quantitative history, the mistrust of statistics in cultural history has contributed to a more qualitative emphasis in digital history. We have, therefore, left synergies with economic and demographic history outside the scope of this article, although we expect such synergies will be valuable to both communities. See Hudson, P., & Ishizu, M. (2016). *History by numbers: An introduction to quantitative approaches*. Bloomsbury Publishing.

While the practices of 'cliometrics' or 'history and computing' are not (yet) standard approaches in historical scholarship, this is not to say that historians have missed the so-called 'digital turn'. Most, if not all, historians use computers to search and store material, as well as prepare publications.[6] With the mass-digitisation of libraries and archives underway since the 1990s, an increasing number of sources can be identified and are accessible online, many to be downloaded and analysed on the historian's computer. These digitised sources are often treated as surrogates; similar, although not identical, to the sources, yet with increased accessibility. Some have argued that digitised sources are much more than digital surrogates; but that these collections of digitised sources should instead be seen as enriched (big) data.[7]

Furthermore, this posits that computers serve to do more than present sources as illustrations accompanying a written narrative, but also provide means to analyse these data in new ways. Under the signifier of 'digital history', historians experiment with tools, concepts and methods from other disciplines (e.g. computer science, and computational linguistics), to develop new perspectives on our past. In this sense, we can understand digital history not as a distinct discipline or field, but as a community of practice of researchers from different backgrounds that look across institutional and disciplinary boundaries to engage historical practices with the methodological and epistemological concepts of other disciplines.[8] Digital history is in this pursuit aligned with the broader field of digital humanities, which gained momentum since 2004 with the emergence of the journal *Companion to Digital Humanities*, wherein computational methods are implemented in pursuit of humanistic questions.[9] The ambition of such pursuits

[6] Jane Winters, 'Digital History', in *Debating New Approaches to History*, eds. Marek Tam and Peter Burke, Bloomsbury Academic (2018): 277-300; Kristen Nawrotzki, & Jack Dougherty, *Writing History in the Digital Age*, University of Michigan Press (2013), https://doi.org/10.3998/dh.12230987.0001.001; Toni Weller, Introduction: History in the digital age. In Toni Weller (Ed.), *History in the Digital Age,* Routledge (2013): 1-20.

[7] Bob Nicholson, 'The Digital Turn', *Media History* 19, no. 1 (1 February 2013): 59–73, https://doi.org/10.1080/13688804.2012.752963.

[8] Max Kemman, 'Boundary Practices of Digital Humanities Collaborations', *DHBenelux Journal* Volume 1 | Integrating Digital Humanities (2019): 1–24, http://journal.dhbenelux.org/journal/issues/001/Article-Kemman/Article-Kemman.pdf, accessed 12 Feb 2020.

[9] Susan Schreibman, Ray Siemens, and John Unsworth, *Companion to Digital Humanities (Blackwell Companions to Literature and Culture)*, Hardcover, Blackwell Companions to Literature and Culture (Oxford: Blackwell Publishing Professional, 2004), http://www.digitalhumanities.org/companion/, accessed 12 Feb 2020.

is to document how digital approaches can diffuse to the broader humanities and historical scholarship, to become part of the general toolkit of humanistic inquiry.

In this *state of the field* article, we discuss several techniques that are currently (widely) used within Digital History/Humanities. Our aim is to provide insight into several approaches that have (already) made an impact within the field or are expected to develop into what could be called 'main-stream' and to reflect on the ever-developing influence of DH in history. We do not claim our discussion to present a comprehensive review of all of the work in digital history; indeed, our discussion depends mostly on Western scholarship published in English. We furthermore focus on working with texts and images, as most work in digital history does. By starting from these common types of data for historical scholarship, and using our own experiences, we aim to trace how methods developed within digital history may transform historical inquiries in the broader historical discipline. Each of the sections, therefore, discusses a technique according to the question *How have the Digital Humanities evolved and what has that evolution brought to historical scholarship?* We begin by discussing techniques that generate and secure data and make them machine searchable, such as OCR/HTR and born-digital archives, computer vision, scholarly editions, and Linked Open Data before moving on to how data is made more accessible by quantitative text and network analysis. We also discuss the importance of hermeneutics and data-awareness. We hope this serves as a starting point for digitally curious scholars to position their research, as well as for those active in digital history to reflect on the future and impact of the digital on the field.

I. Generating and securing data for further analysis

Historians work with a broad range of sources: primary documents in text and image format; analogue, digitised and born-digital documents; architecture, cultural artefacts and documentation of non-tangible heritage. Making digitised and digital sources available is increasingly becoming a core element in many research projects. Documentation and preservation of primary sources through digital replicas of sources

and objects, scholarly editions and born-digital archives are essential to historical scholarship. In the following section, we will look at several digital documentation and preservation formats, in which primary sources may be made available, searchable and ready for further analytic processing.

From Optical Character Recognition to Handwriting Text Recognition

Written documentation is core to our work as historians. Neither printed, nor handwritten text are readable by a computer. A computer can only recognise these images as text if it is trained to do so. Initially, Optical Character Recognition (OCR) was developed so that text could be 'read' by those with reading challenges, a task performed by Edmund Edward Fournier d'Albe's *optophone* (1910s) which transformed characters into sounds. In the 1950s, David Shepard developed *Gismo* which first transformed text to computer-readable data. Raymond Kurzweil was active in inventing the first omni-font OCR-system, which he further developed into a system that would convert data into text to be read out loud to visually-impaired people. This approach leveraged the strength of computers: to recognise images based on the statistical likelihood of language patterns it had been trained on.

Whereas OCR is applied to standard fonts, a finite number of characters and texts printed on a bright background, Handwriting Text Recognition (HTR) has to overcome the extensive variation in handwriting. To be able to decipher handwriting, several techniques needed to be combined: the statistical analysis of language patterns, artificial intelligence combined with deep learning, and human training. Although individual - very regular - hands can be trained through OCR-programs[10], the results generated, for example, by the READ-Coop's HTR tool Transkribus are promising.[11] What separates Transkribus - a commonly used "platform for the automated recognition, transcription and searching of historical documents, from OCR-engines, is the learning curve.[12] For example, the more transcribed pages that are added, the greater 'proficiency' that that language-patterns are understood; resulting in Character Error

[10] E.g. Kraken, Tesseract, ABBYY FineREADER.
[11] https://read.transkribus.eu/about/, accessed 12 Feb 2020.
[12] http://transkribus.eu, accessed 12 Feb 2020.

Rates (CER) between 10-25% on previously unseen handwritten material, and less than 10% when applied to similar hands (e.g. clerical texts/ paid scribes), and less than 5% when trained on an individual hand.[13]

Consequently, both OCR and HTR have had an enormous impact on the conversion of printed and written texts into machine-readable textual data, offering - first and foremost - the possibility of searching texts.[14] Increasingly both techniques are used for digitising collections, with the quality and thus capacity to read/recognise texts continuing to improve incrementally to the improvements in digital imaging. Whether we will recognise this as an independent step within the processing of formerly paper documents to data, or if (and possibly: when) OCR/HTR will come to be integrated within a data-pipeline that will incorporate many other techniques - such as Named Entity Recognition - is difficult to predict.

Born-digital Archives

The textual sources used in historical work are not solely physical; they are also born-digital archives. The *Internet Archive* (since 1996) and its frontend, the *Wayback Machine*, are undoubtedly the most well-known born-digital archives, yet born-digital archives are much more diverse.[15] Personal archives, institutional repositories, the preserved collections of digital art in museums and galleries,[16] digital community

[13] Additional on HTR: Guenter Muehlberger et al., 'Transforming Scholarship in the Archives through Handwritten Text Recognition', *Journal of Documentation* 75, no. 5 (2019): 954–76; https://doi.org/10.1108/JD-07-2018-0114.

[14] At this point, the conversion is (mainly) into plain text; that is, in short, also the downside of both processes to this date: the original layout markup is lost in the conversion. While the original authors and/ or printers would have had a reason behind the computer, they cannot recognise structure. An OCR-tool as ABBYY FineReader does recognise if a text is printed in bold, italics or in a larger font, but it does not yet digest this into information on titles, tables or even paragraphs - it merely notes differences in features.

[15] Thorsten Ries and Gábor Palkó, 'Born-Digital Archives', *International Journal of Digital Humanities* 1, no. 1 (1 April 2019): 1–11, https://doi.org/10.1007/s42803-019-00011-x; Lise Jaillant, 'After the Digital Revolution: Working with Emails and Born-Digital Records in Literary and Publishers' Archives', *Archives and Manuscripts* 47, no. 3 (2 September 2019): 285–304, https://doi.org/10.1080/01576895.2019.1640555. For the current state of the field of Web History, see: Niels Brügger and Ian Milligan (Eds.): The SAGE Handbook of Web History. London: Sage 2018.

[16] Patrícia Falcão and Tom Ensom, 'Conserving Digital Art', in *Museums and Digital Culture: New Perspectives and Research*, ed. Tula Giannini and Jonathan P. Bowen, Springer Series on Cultural Computing (Cham: Springer International Publishing, 2019), 231–51, https://doi.org/10.1007/978-3-319-97457-6_11.

archives,[17] national web archives,[18] and social media archives[19] offer research opportunities for historical, art historical, and literary scholarship and have already generated an impressive volume of research, notably in web history.[20] As James Baker argues that from a digital forensics perspective mobile phones,[21] the Internet of Things, and cloud data will soon become part of the historical record that historians will want to access to reflect on the past.[22]

With all these different types of born-digital archives, digital preservation practitioners, archivists and researchers face specific challenges and complexities. The data volume of born-digital archives; hardware, software, standards and context obsolescence become challenges and complexities for preservation over time. The broad spectrum, variety and historical fluidity of digital materiality,[23] and the resulting possible digital forensic analytical angles complicate data recovery, born-digital analysis of creation history, provenance, metadata and hidden embedded content and structures of digital primary sources by requiring historical forensic analytical knowledge and tools, which ultimately make documentation of findings to the research public fairly complicated. Digital archivists need to deal with challenges ranging from considering and balancing the ethics of the formation of dark archives and saving the content of online communities and cultures to the archaeological discovery, recovery of long gone

[17] Abigail De Kosnik, *Rogue Archives: Digital Cultural Memory and Media Fandom* (MIT Press, 2016); Sharon Webb, '"Digital Archives in Communities – Practice and Preservation" : A Summary (or at Least an Attempt) - Digital Preservation Coalition', accessed 30 October 2019, https://www.dpconline.org/blog/digital-archives-in-communities; Ian Milligan, 'Finding Community in the Ruins of GeoCities: Distantly Reading a Web Archive' (Institute of Electrical and Electronics Engineers, 2015), https://uwspace.uwaterloo.ca/handle/10012/11650.

[18] See project RESAW (https://resaw.eu/, https://resaw.eu/web-archives/) and the list of IIPC members (http://netpreserve.org/about-us/members/), accessed 26 Oct 2019.

[19] Rob Procter, Farida Vis, and Alex Voss, 'Reading the Riots on Twitter: Methodological Innovation for the Analysis of Big Data', *International Journal of Social Research Methodology* 16, no. 3 (1 May 2013): 197–214, https://doi.org/10.1080/13645579.2013.774172.

[20] Niels Brügger, *The Archived Web* (MIT Press, 2018); Eveline Vlassenroot et al., 'Web Archives as a Data Resource for Digital Scholars', *International Journal of Digital Humanities* 1, no. 1 (1 April 2019): 85–111, https://doi.org/10.1007/s42803-019-00007-7.

[21] Owens, Trevor, 'Historic IPhones: Personal Digital Media Devices in the Collection', *Trevor Owens* (blog), 15 November 2013, http://www.trevorowens.org/2013/11/historic-iphones-personal-digital-media-devices-in-the-collection/ Owens, Trevor, accessed 12 Febr 2020..

[22] James Baker, 'Digital Forensics in the House of Lords: Six Themes Relevant to Historians (Part One)', *Blog of the Software Sustainability Institute* (blog), 29 March 2019, https://software.ac.uk/blog/2019-03-29-digital-forensics-house-lords-six-themes-relevant-historians-part-one, accessed 12 Febr 2020.

[23] Baker.

websites from offline backups[24] and the reflection and documentation of possible misrepresentations, lacunae and imbalances in these born-digital archive collections.

As a consequence, researchers and archivists working with born-digital archives not only need data-mining and visualisation tools (such as *Archives Unleashed,*[25] and the data-mining functionality in *BitCurator*[26]) but also to understand and analyse primary born-digital sources as documents in their own right.[27] While the beginnings of born-digital preservation date back to the endeavour of the *Internet Archive* and the work of a few pioneering archivists in the 1990s and 2000s (e.g. Susan Thomas, Jeremy L. John), the major shift that marked the rise of the born-digital studies is Matthew Kirschenbaum's seminal book *Mechanisms. New Media and the Forensic Imagination*.[28] In the following years, Kirschenbaum's and Doug Reside's work became paradigmatic academic use cases for personal archives and digital primary records archived according to digital forensic standards, which were accompanied by large international projects on born-digital archiving in the archival and Galleries, Libraries, Archives, and Museums (GLAM) sector, leading to the development of the *Bitcurator* Linux distribution and its archival toolset. This work showed that in-depth knowledge of computing history and digital forensic, 'e-palaeographic' skills[29] are needed when archivists and researchers secure, preserve, curate and interpret the distributed and fragile forensic materiality of born-digital historical primary records.[30]

---

[24] Johan van der Knijf, 'Recovering '90s Data Tapes. Experiences From the KB Web Archaeology project', paper on iPres 2019, URL: https://ipres2019.org/static/pdf/iPres2019_paper_9.pdf (accessed 26 Oct 2019), see also: https://www.bitsgalore.org/2019/09/09/recovering-90s-data-tapes-experiences-kb-web-archaeology (accessed 26 Oct 2019).

[25] Ian Milligan, 'The Archives Unleashed Project', accessed 30 October 2019, https://archivesunleashed.org/. (accessed 26 Oct 2019).

[26] Bitcurator Consortium: *Bitcurator*, URL: https://bitcurator.net/ (accessed 26 Oct 2019).

[27] Jane Winters, 'Web archives and (digital) history: a troubled past and a promising future?', in *The SAGE Handbook of Web History* (London: Sage, 2018), 593–606.

[28] Matthew G. Kirschenbaum, *Mechanisms: New Media and the Forensic Imagination* (MIT Press, 2008).

[29] The term of the 'e-palaeographer' was coined by Edward Higgs and R.J. Morris, eds., 'Electronic Documents and the History of the Late Twentieth Century: Black Holes or Warehouses?', in *History and Electronic Artefacts* (Clarendon Press, 1998), 33.

[30] Matthew Kirschenbaum, 'The .Txtual Condition: Digital Humanities, Born-Digital Archives, and the Future Literary', *Digital Humanities Quarterly* 007, no. 1 (1 July 2013), http://digitalhumanities.org/dhq/vol/7/1/000151/000151.html.

An important recent development in this sub-field is the focus on methods to introduce critical source appraisal, data criticism and more in-depth analysis to web history research;[31] thus moving in a direction where forensic detection of digital disinformation, 'deep fake' and forgery, automated content generation and bots, online threat, malware[32] and hacking will play an increasingly important role in born-digital preservation, archiving and web history research. Ecological considerations about the carbon footprint of data management will likely also become a focus for researchers.[33]

Computer Vision

Whilst text has been central to the identity of the Digital Humanities; historical scholarship is not limited to the study of text. The ability of machines to comprehend digital images has made remarkable strides in recent years,[34] and it is in the context of these developments that computer vision has been used in the service of historical scholarship. These questions tend to address scale:[35] Which digital images are available? How are images similar? How can large-scale visual analysis be used to understand change over time in the production, use, and content of visual culture?

---

[31] Anne Helmond, 'Track the Trackers', accessed 30 October 2019, https://wiki.digitalmethods.net/Dmi/DmiWinterSchool2012TrackingTheTrackers (accessed: 26 October 2019); Trevor Owens and Grace Helen Thomas, 'The Invention and Dissemination of the Spacer Gif: Implications for the Future of Access and Use of Web Archives', *International Journal of Digital Humanities* 1, no. 1 (1 April 2019): 71–84, https://doi.org/10.1007/s42803-019-00006-8.

[32] Jonathan Farbowitz, *More Than Digital Dirt: Preserving Malware in Archives, Museums, and Libraries*, 2016, http://archive.org/details/16sThesisFarbowitzFinal.

[33] Zack Lischer-Katz, 'Studying the Materiality of Media Archives in the Age of Digitization: Forensics, Infrastructures and Ecologies', *First Monday* 22, no. 1 (2017), https://doi.org/10.5210/fm.v22i1.7263; Keith L. Pendergrass et al., 'Toward Environmentally Sustainable Digital Preservation', *The American Archivist* 82, no. 1 (1 March 2019): 165–206, https://doi.org/10.17723/0360-9081-82.1.165.

[34] Services that launched less than five years ago – such as Microsoft's much derided #HowOldRobot or Flickr's auto-tagger – now seem primitive when compared with the present day use of facial recognition technology to replace sports tickets and to oppress populations; Mike Moore, 'Intel Rolling out Facial Recognition Tech at Tokyo 2020 Olympics', TechRadar, accessed 30 October 2019, https://www.techradar.com/uk/news/intel-is-bringing-facial-recognition-to-toky-2020, accessed 12 Febr 2020 ; Griffiths, James, *The Great Firewall of China* (London: Zed, 2019).

[35] Whilst we do not discuss the use of spectral imaging to analyse the histories of individual paintings and drawings, we note these methods have enabled important findings, see Henri Neuendorf, 'X-Ray Analysis Reveals Joshua Reynolds Repainted Rembrandt Masterpiece', artnet News, 5 March 2015, https://news.artnet.com/exhibitions/x-ray-analysis-reveals-joshua-reynolds-repainted-rembrandt-masterpiece-27350, accessed 12 Febr 2020; Cerys Jones et al., 'Leonardo Brought to Light: Multispectral Imaging of Drawings by Leonardo Da Vinci', (12 March 2018), https://doi.org/10.5281/zenodo.1208430.

A significant milestone in the use of these techniques for history research was Lev Manovich's 'How to Compare One Million Images' (2012), in which digital images (as opposed to data points that represent them) are plotted by their visual characteristics – measures of brightness, saturation, hue – as a means of observing visual patterns at scale.[36] Since then, "word and image" scholars have made significant interventions such as *The Illustration Archive* (2015) which used crowdsourcing, machine tagging, and similarity matching the enhance the discovery of images, to link them, and to make legible – in visual terms – the larger patterns in pre-twentieth century book illustration.[37] To isolate illustrations for use in their digital archive, *The Illustration Archive* team used page-level XML (see the next section) containing the x and *y* coordinates for every element on each digital image. Using these XML features of the placement and size of images over time, between genres, and across single volumes, Will Finley tracked the printing of illustrations between 1780 and 1860,[38] enabling him to articulate the broader patterns of book illustration and to assert the importance of publishers to how book knowledge was constructed in the interplay between word and image.[39]

Work on historical images is advancing quickly, using convolutional neural networks – a machine learning approach commonly used to detect and classify features of visual inputs, and that is powering recent step-changes in computer vision – Wever and Smits landmark 2019 work showed how such an approach could be used to enrich our understanding of trends in historical corpora.[40] Taking over a century of Dutch newspapers as their source material, Wevers and Smits detected their non-textual elements,

---

[36] Lev Manovich, 'How to Compare One Million Images?', in *Understanding Digital Humanities*, ed. David M. Berry (London: Palgrave Macmillan UK, 2012), 249–78, https://doi.org/10.1057/9780230371934_14.

[37] http://illustrationarchive.cf.ac.uk/, accessed 28 October 2019; Julia Thomas, *Nineteenth-Century Illustration and the Digital: Studies in Word and Image*, The Digital Nineteenth Century (Palgrave Macmillan, 2017), https://doi.org/10.1007/978-3-319-58148-4.

[38] William Finley, 'Making an Impression: An Assessment of the Role of Print Surfaces Within the Technological, Commercial, Intellectual and Cultural Trajectory of Book Illustration, c. 1780-c.1860' (PhD, University of Sheffield, 2018), http://etheses.whiterose.ac.uk/23081/.

[39] William Finley, 'Data and Code For PhD Thesis- Making an Impression: An Assessment of the Role of Print Surfaces Within the Technological, Commercial, Intellectual and Cultural Trajectory of Book Illustration c.1780-c.1860.' (Zenodo, 9 September 2018), https://doi.org/10.5281/zenodo.1412137.

[40] Melvin Wevers and Thomas Smits, 'The Visual Digital Turn: Using Neural Networks to Study Historical Images', *Digital Scholarship in the Humanities*, accessed 30 October 2019, https://doi.org/10.1093/llc/fqy085.

charted their growth over time, and semi-automatically classified images by their visual characteristics and informational content. By taking this approach Wevers and Smits were able to cluster images by their arrangement (e.g. advertisements featuring a particular visual style), by their subjects (e.g. groups of people), or by their genre (e.g. chess problems). Wevers and Smits provide a much-needed pathway towards a scalable and historically relevant computational analysis of images by informational content, towards digital history the uses machines to analyse the information content of images rather than textual proxies for those images.

Digital scholarly editions

Scholarly editions preserve and make available the content of primary historical sources for a community of specialists and the interested public. They usually provide explanatory information in the commentary, and may additionally feature expert information such as bibliographical data, information about provenance and materiality of the sources. The same motivations that drove, for example, the Library of Alexandria's third century BC critical edition of the works of Homer, remain today in digital scholarly editions.[41] The main difference, however, is that, freed from the constraints of the printing press,[42] a digital edition can create searchable and linkable connections between textual features, include a variety of both static and interactive visualisations, and be complemented with a virtually unlimited critical apparatus and commentary.

We could call any digital form of a work a digital edition. Before 2000, most digital editions were produced by reproducing the contents of a manuscript or printed text with the aid of a word processor. Nowadays, scholars demand more open, reliable and standardised digital editions. Some vast text archives,

[41] For definitions: http://uahost.uantwerpen.be/lse/index.php/lexicon/scholarly-edition/ and http://uahost.uantwerpen.be/lse/index.php/lexicon/edition-digital/, accessed 12 Dec 2019.
[42] Marita Mathijsen, *Naar de letter. Handboek editiewetenschap.* (Den Haag: KNAW Press, 2010), 19–29 and i–vi, https://www.dbnl.org/tekst/math004naar03_01/, accessed 12 Febr 2020..

like *Gallica*[43], offer scholars scanned images of document pages, but the full-text layer may, if the result of automated OCR (see above), not meet scholarly standards of a reliable, citable scholarly resource. By contrast, online digital collections like the *Women Writers Project*, the *Oxford Text Archive*, the *Digital Library for Dutch Literature* or the German Text Archive (DTA), and online digital scholarly editions such as the Samuel Beckett Digital Manuscript Project, the Arthur Schnitzler Digital Critical Edition and Nietzsche Source make the texts available at scholarly quality standards and often offer additional analytical features and tools.[44] In order to facilitate this quality, these editions use a form of eXtensible Markup Language called TEI-XML to 'mark up' features of the text such as layout, variants, marginalia, text structures, and entities (people, places, things). The usability of a digital edition may be further improved by providing access to metadata as Linked Open Data (see below). The Text Encoding Initiative (TEI) - the first guidelines for which were released in 1990 - has become the most commonly used standard for scholarly markup of textual sources in digital editions. It is interoperable, relatively easy to learn, and can be flexibly extended in order to encode highly complex textual phenomena.[45]

One thing that remains unchanged in the digital era is the labour involved in producing scholarly editions: models like TEI take time, skill, and domain-specific knowledge to be used effectively for scholarly editions. Nevertheless, digital transformations have enlarged the possibilities in the field of scholarly editions enormously: from providing access to sophisticated, multi-layered texts to enabling distant reading between otherwise disparate sources. These developments are, fortunately, independent from TEI: digital scholarly editions encoded in this standard can be converted to a new standard if TEI loses its role as the *lingua franca* for digital scholarly editions[46]. Infrastructures like TEI have both democratised

[43] Thomas Crombez, 'Digitale deemstering. Auteursrecht en de digitalisering van boeken in Nederland en Vlaanderen'. In: *Vooys. Tijdschrift voor letteren*, vol. 37 (2019), issue 3, 48-49. Gallica contains .txt, .pdf and .jpg files of the source.

[44] https://www.wwp.northeastern.edu, https://www.ota.ox.ac.uk/, https://www.dbnl.org/. http://www.deutschestextarchiv.de/, https://www.beckettarchive.org/, https://www.cam.ac.uk/Schnitzler-Edition, http://www.nietzschesource.org/, accessed 12 Febr 2020.

[45] For XML in general: https://www.w3schools.com/xml/xml_whatis.asp.

[46] TEI-XML has not been conceptualised as an eternal standard, it was implemented first in SGML, then migrated to XML. Since structural limitations of XML markup, e.g. cumbersome solutions for the problem of overlapping tag brackets, have never been sufficiently solved, the community is working on alternatives, For instance graph-based

the practice of scholarly editing and given scholarly editors a platform from which to fulfil the intellectual ambitions of this enduring genre of humanities practice. New infrastructures must be developed according to the same principles.[47]

Linked Open Data

In addition to text and images, historians are starting to discover the benefits of *Linked (Open) Data* (LOD). In 2006, Tim Berners-Lee, the inventor of the Web, wrote a memo on the Semantic Web which: 'provides a common framework that allows data to be shared and reused across application, enterprise, and community boundaries.'[48]; of which LOD served as a technique to describe knowledge. LOD standards afford a way to make (meta)data on and of objects available and (ideally) publicly accessible in a format readable by both humans and machines. Thus instead of referencing an unstructured description of a place, person or object (e.g. a dictionary entry or book), linked data through standards such as the *Resource Description Framework* (RDF) provides a standardised structure to organise, store and link information on these entities. For example, historical statements such as "Dante wrote The Divine Comedy" could be expressed as a triple consisting of:

- a *subject* (":Dante"),
- a *predicate* (":wrote"),
- and an *object* (":The_Divine_Comedy").

Each of these items is represented with unique identifiers (Uniform Resource Identifiers - URIs) that machines can read and retrieve. One of the best-known examples using such statements is Google's

editions, and variants of linked data: RDFa markup, JSON(-LD). Sustainability of scholarly editions is a big issue in this field of research, and an alternative solution to converting TEI to other standards is building editions as plain HTML 'minimal computing' (Gil, Visconti), 'prêt-à-porter' (Pierazzo) editions, which will be long-term supported by browsers.

[47] Patrick Sahle, '2. What Is a Scholarly Digital Edition?', in: *Digital Scholarly Editing: Theories and Practices*, ed. Matthew James Driscoll and Elena Pierazzo, Digital Humanities Series (Cambridge: Open Book Publishers, 2017), 19–39, http://books.openedition.org/obp/3397.

[48] https://www.w3.org/2001/sw/, accessed 10 Oct 2019.

Knowledge Graph[49], which identifies whether a search term refers to a person or organisation, and provides relevant information to that entity in a "knowledge panel" in the results page. The structuring of information in this way is also the backbone to Wikidata, DBpedia, and Geonames, platforms that are increasingly seen as primary and secondary sources in historical work to verify dates, locations, birthplaces, or known occupations of individuals, organisations, and places.

LOD is also important to historical scholarship as it is seen as the gold standard for maximising the reuse of data, see Figure 2. Open Data.

Figure 2. Open Data. Source: https://5stardata.info/en/ [retrieved: 10-10-2019].

*Explanation: *OL= OpenLicence; *RE=machine REadable; *OF= OpenFormat; *URI= Uniform Resource Identifiers; *LD= LinkedData*

[49] https://googleblog.blogspot.com/2012/05/introducing-knowledge-graph-things-not.html, accessed 30 Oct 2019.

The 5-star Linked Data rating system[50] encourages people to publish data on the Web in an increasingly open, structured and linked manner; where the fifth star is only given if data is linked cross-datasets through URIs. This cross-dataset linkage encourages data reuse, preventing repetitive information; but also enables other URIs representing the same entity or concept to be published elsewhere and to be linked together. Linking all possible sorts of data has led to a massive amount of data which is the Linked Open Data Cloud (see Figure 3). For historical scholarship, this means that statements about a single entity can be taken from a large amount of sources spread over many archives in order to gain a bigger picture or to identify opposing views.

In addition, to the usefulness of storing information and thus querying it in this way, linked data is also important to historical scholarship as libraries, archives and museums are increasingly making their catalogues and distinct collections open through RDF.[51] Still, the process of converting a catalogue to RDF is laborious as a large share of metadata on collections is expressed in natural language and often with different metadata standards, making it difficult to implement an automatic process of RDF generation on complete collections; large scale infrastructures are in progress.[52] Despite the potential of RDF and affordances of implementing specific queries, its use remains a technical barrier for many; which has led to a discussion on emphasising usability for non-technical users through *Linked Open Usable Data.*

[50] https://www.ted.com/talks/tim_berners_lee_on_the_next_web?language=nl, accessed 10 Oct 2019.
[51] An non-exhaustive list of RDF data services from national libraries: the US Library of Congress, Linked Data Service id.loc.gov, the BnF data.bnf.fr, the BNE datos.bne.es, KB .the Short-Title Catalogue Netherlands.
[52] Rinke Hoekstra et al., 'The DataLegend Ecosystem for Historical Statistics', Journal of Web Semantics 50 (May 2018): 49–61, https://doi.org/10.1016/j.websem.2018.03.001.

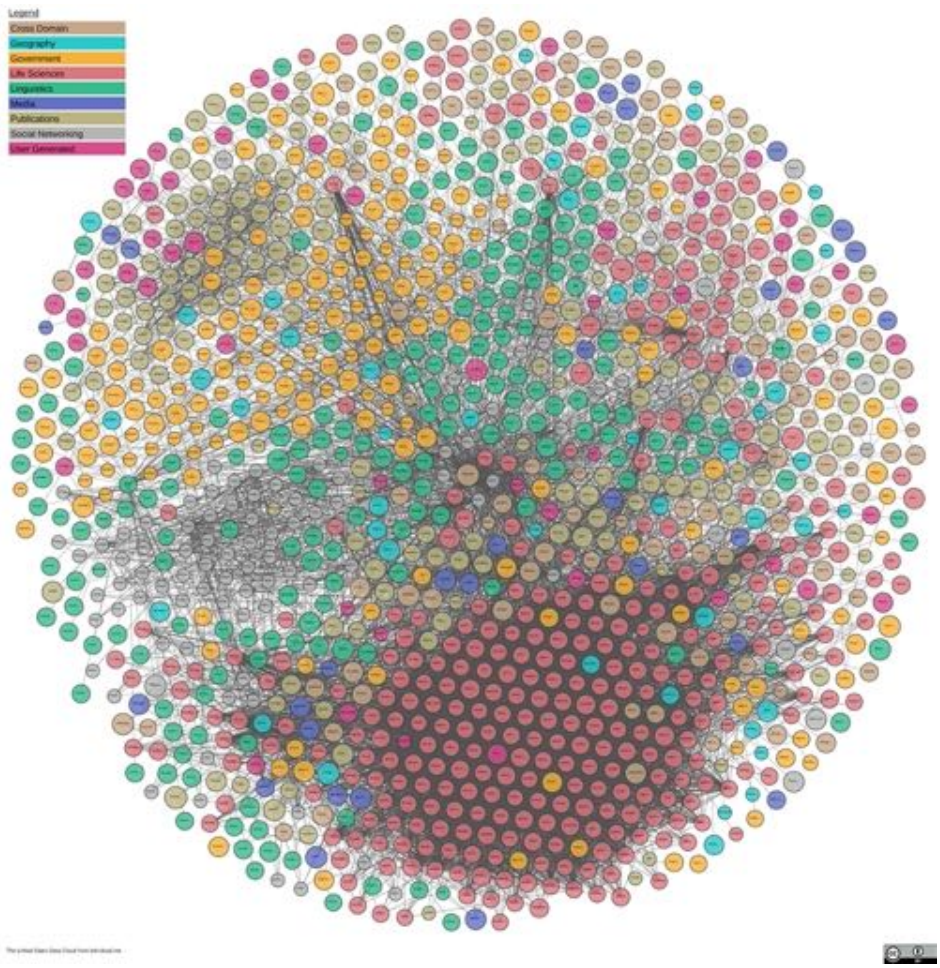

Figure 3. Linking open data cloud diagram 2020, by Max Schmachtenberg (et. al.) https://lod-cloud.net/ [situation on/ accessed 12 Febr 2020].

II Analysis

The increasing availability of digitised sources, either born-digital or made machine-readable, affords efficient assessment of sources. For example, the querying of terms through OCR enabled text, indexing and cataloguing of sources based on metadata, or the use of information or data from digital sources. In this section, we describe the possibilities for historical scholarship using quantitative text analysis as a means of understanding context and changes in language; as well as network analysis to investigate relational phenomenon.

Quantitative text analysis

Today millions of books, newspapers and letters are only ever a few clicks away. At the heart of historical text analysis lies the identification of linguistic patterns; e.g. where the frequency of keywords suggests phenomena that have changed over time. For many historians, it was the Google Books Ngram Viewer that first introduced them to *n*-gram frequency.[53] Announced in 2011 the tool was presented as a revolutionary new way of looking at culture.[54] Its capacity to rapidly offer an overview of a word's frequency has become essential in studying historical phenomena.[55]

Frequency-based tools and methods are, however, not without their problems. Since its inception in 2011, many scholars have pointed at the pitfalls of Google's Ngram Viewer.[56] Their critiques often apply to other frequency-based methods and fall into three categories. First, even the Google Books corpus, which is said to host 5% of all the books ever printed, does not represent 'language' or 'culture': it, like many corpora, is restricted in its representativity. Gauging the representativity of corpora requires careful

[53] Corpus linguists often refer to counted words as '*n*-grams': sequences of *n* words.
[54] Jean-Baptiste Michel et al., 'Quantitative Analysis of Culture Using Millions of Digitized Books', *Science* 331, no. 6014 (14 January 2011): 176–82, https://doi.org/10.1126/science.1199644.
[55] Paul Caruana-Galizia, 'Politics and the German Language: Testing Orwell's Hypothesis Using the Google N-Gram Corpus', *Digital Scholarship in the Humanities* 31, no. 3 (1 September 2016): 441–56, https://doi.org/10.1093/llc/fqv011.
[56] Eitan Adam Pechenick, Christopher M. Danforth, and Peter Sheridan Dodds, 'Characterizing the Google Books Corpus: Strong Limits to Inferences of Socio-Cultural and Linguistic Evolution', *PLOS ONE* 10, no. 10 (7 October 2015): e0137041, https://doi.org/10.1371/journal.pone.0137041.

contextualisation through structured metadata: knowing who wrote what, when, and in which context is essential in being able to explain changes in frequency.

In addition, there are multiple reasons why a word changes in frequency over time. Changing spelling conventions, the emergence of idioms or features of the data all determine the frequency of a word. Jumping to conclusions based on sudden changes is, therefore, a risky undertaking. Also, nothing guarantees that a word meant the same in the past. Mapping the changing frequency of a word becomes problematic if the same word meant something different in the past. Here, the detection of changes in the broader 'semantic field' of a word, as well as information on the composition of the data at a specific moment in time, can explain sudden ruptures.

In response to the potential problems associated with keyword frequency, recent approaches have transcended the level of individual words. The object of research shifts from the individual word to a broader 'semantic field'.[57] Instead of looking solely at the frequency of for example "foreign", one could also follow the 'behaviour' of all bigrams starting with 'foreign', such as "foreign bank" or "foreign trade" (Figure 1).[58] The second trend in historicising word meaning is the application of language modelling in digital history. Based on the context of a word, machine-learning techniques can quantify meaning. For example, the word 'king' is semantically similar to 'queen' because its 'neighbours' are similar ('palace', 'prince'). By applying this premise, computers are now able to identify words similar to a given keyword in specific temporal contexts.

[57] Jan Ifversen, 'About Key Concepts and How to Study Them', *Contributions to the History of Concepts* 6, no. 1 (1 June 2011): 65–88, https://doi.org/10.3167/choc.2011.060104.

[58] M. J. H. F. Wevers, 'Consuming America : A Data-Driven Analysis of the United States as a Reference Culture in Dutch Public Discourse on Consumer Goods, 1890-1990', Dissertation, 15 September 2017, http://dspace.library.uu.nl/handle/1874/355070; Mikko Sakari Tolonen et al., 'Spheres of "Public" in Eighteenth-Century Britain', 2018, https://researchportal.helsinki.fi/en/publications/spheres-of-public-in-eighteenth-century-britain, accessed 12 Febr 2020.

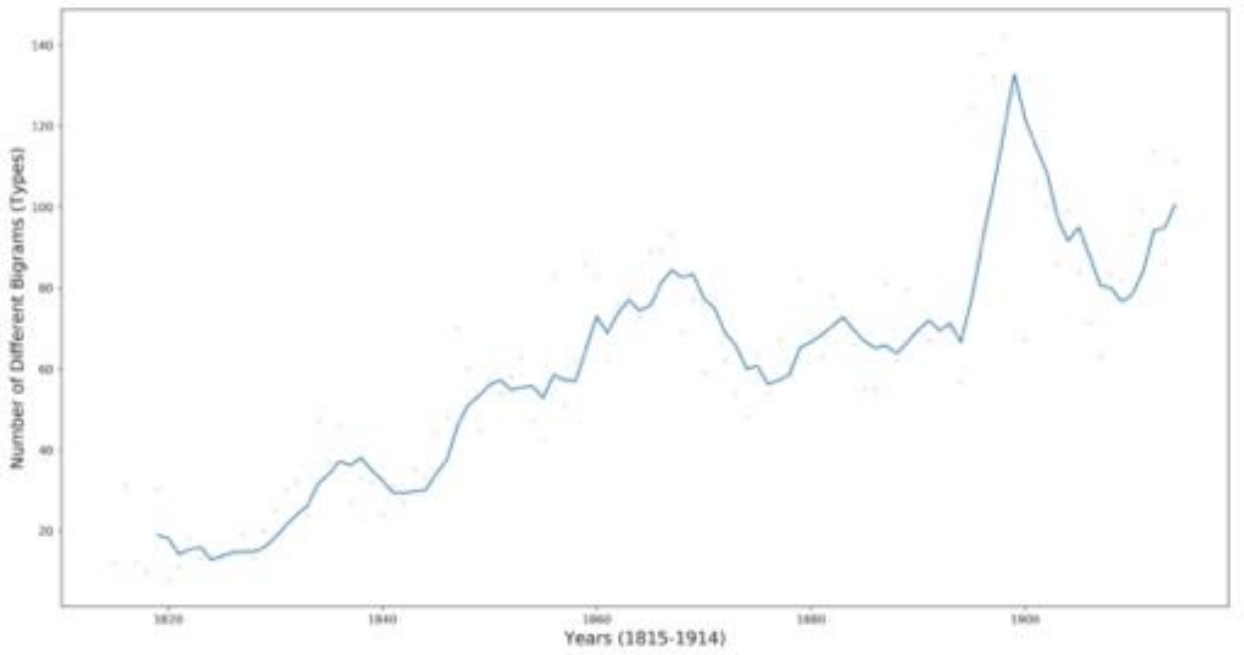

Figure 1. The (absolute) number of different bigrams that contain the adjective "binnenlandsche" ("domestic") in Dutch newspapers between 1815 and 1914.[59]

Future research in historical textual data will likely involve better contextualisation through structured metadata. Full texts are not sufficient by themselves. To use them as historical data, they need additional information on their production and dissemination. Also, future research will transcend the level of words. Computational methods are increasingly able to model sentences, rhetorical tropes and discourses, which allows a more comprehensive grasp of historical language change. Combined with proper metadata, research into these "supra-lexical" units of analysis will hopefully complement a focus on the keyword(-search) and give a better insight into historical change. Besides the modelling of meaning on different linguistic levels, the detection of specific 'named entities' such as people, places, organisations is instrumental in gaining a better insight into historical texts.[60]

[59] R.S. Ros, 'The Birth of the Foreign : A Digital Conceptual History of Buitenland in Dutch Newspapers 1815-1914', Master thesis, 2019, http://dspace.library.uu.nl/handle/1874/382176, accessed 12 Febr 2020.
[60] C. Grover, S. Givon, R. Tobin and J. Ball, 'Named Entity Recognition for Historical Texts', Proceedings of the Sixth International Conference on Language Resources and Evaluation (Marrakech 2008), http://www.lrec-conf.org/proceedings/lrec2008/pdf/342_paper.pdf, accessed 12 Febr 2020.

Network Approach and Analysis

In retracing history, there is a need and interest to reconstruct the networks of the past. As research on social networks has shown, these networks matter: the position one has in a social network influences one's power and performance, as well as the structure of the relations that lend social, economic and political capital for individuals and organisations. Network analysis as a method has been used to analyze these structures and positions as a way of understanding relational phenomena.

Identifying historical networks is a laborious task, which traditionally has been done by hand in the archive, such as the work done by John F. Padgett and authors on the Medici networks in the early 1400s.[61] Researchers identify *nodes* and *edges*; where nodes can be individuals, organisations, or objects that can be related to another node via an *edge*- a connection or relationship (not dissimilar to the way *triples* work in Linked Open Data). For example, in the Mapping of the Republic of Letters project, correspondence between scholars in the late 17th and 18th centuries was projected as networks of senders and receivers to reconstruct communication flows during the Age of Enlightenment.[62]

The digitisation of archives and catalogues has afforded historical network research a new avenue for constructing networks. The increased access to metadata of archival materials (see: Linked Data), and digitisation and transcriptions of textual sources (see: Section I) have opened up an avenue of (semi-)automatic identification of historical networks, e.g. through written correspondence, manuscripts, printed materials such as books, newspapers or periodicals.[63] These approaches have resulted in the ability to

[61] John F. Padgett and Christopher K. Ansell, 'Robust Action and the Rise of the Medici, 1400-1434', *American Journal of Sociology* 98, no. 6 (1 May 1993): 1259–1319, https://doi.org/10.1086/230190.
[62] Giovanna Ceserani and Thea De Armond, 'British Architects on the Grand Tour in Eighteenth-Century Italy: Travels, People, Places', accessed 30 October 2019, https://purl.stanford.edu/ct765rs0222.
[63] Matje van de Camp and Antal van den Bosch, 'A Link to the Past: Constructing Historical Social Networks', in *Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis (WASSA 2.011)* (Portland, Oregon: Association for Computational Linguistics, 2011), 61–69, https://www.aclweb.org/anthology/W11-1708; Jana Diesner, 'From Texts to Networks: Detecting and Managing the Impact of Methodological Choices for Extracting Network Data from Text Data', *KI - Künstliche Intelligenz* 27, no. 1 (1 February 2013): 75–78, https://doi.org/10.1007/s13218-012-0225-0.

investigate more entities (more extensive networks), consider multiple types of relations (multiplex networks), and explore the dynamics of these networks (over multiple periods of time).

In addition to using computational approaches to identify networks, network analysis as a method provides an avenue to quantitatively analyse the characteristics of networks (whether inferred by hand or through computational techniques). Network analysis may include the analysis of the positions of nodes to assess relational power or the structure of a network to explain social capital and performance, where the network serves as a proxy for social structures. This method allows researchers to explore relational questions that complement our understanding of political, social and cultural phenomena in the past. The state-of-the-art on network analysis in historical scholarship depends on the period and domain; the Historical Network Research Network provides a systematic bibliography of network research in history that serves as a great starting point for positioning relational research questions in different periods, contexts, or entities.[64]

III Data awareness

Academics, and historians, in particular, need to be aware of the origin and authenticity of the data they use and of what has been in- and excluded in their selection. When dealing with analogue data, this task mainly concerns critically appraising the information that has been found and the strategy that has been chosen to identify the material. When dealing with digital sources, an additional task is required: interrogating the process through which the digital source has been made available. This implies being informed about the selection criteria for determining what is digitised, about alterations that occur during this process, and about how search algorithms determine which results appear on a historian's computer screen when conducting a search. This section is intended to raise awareness of data handling and possible pitfalls.

[64] Marten Düring, 'Historical Network Research. Network Analysis in the Historical Disciplines', 2017, http://historicalnetworkresearch.org/, accessed 12 Febr 2020.

Digital Hermeneutics or how to be critical about computer-code

The term 'hermeneutics', coined by the 19th-century German historian Droysen to emphasise the importance of 'interpretation' to construct historical knowledge,[65] has been reconceptualised in 'Digital Hermeneutics', in the light of the need to reflect on how computers influence the construction of scientific knowledge.

What is striking is that term refers to something 'new', while at the same time its etymology reveals its classical roots. *Digital* comes from the Latin *digitus* and refers to how numerals under ten were counted with fingers and *Hermes* was the god who delivered and interpreted messages in Greek mythology. When Mallery, Hurwitz and Duffy coined the phrase in 1986, they did so to understand the potential of computers in extracting meaning from classical texts.[66] Just as philology, the practice of applying source criticism to classical texts, is the origin of source criticism in the realm of history, which in turn contributed to the archival turn at the end of the 19th century, so was studying the relation between computers and human expression the beginning of a development that would eventually lead to the digital turn in humanities at the beginning of the 21st century.

The habitat of historians, who spend most of their time - often unconsciously - executing commands that make things happen on their screen, demands the integration of the principle of digital hermeneutics into the appreciation of the digital content that they retrieve through the web. This need is not a specific requirement for historians who engage with digital methods but applies to the historical community in its entirety. Scholarly work of historians is increasingly affected by the logic of digital library and archival information systems and of commercially driven strategies for selection and indexing of companies such as Google and Bing. Having a basic knowledge of how they function is now just as relevant as being able to identify bias in news coverage or forgeries in old manuscripts.

---

[65] Mueller, Philippe (2008), 'Understanding History: Hermeneutics and source criticism in historical scholarship', in: *Reading Primary Sources, the interpretation of texts from nineteenth and twentieth century*, (Eds.) Miriam Dobson and Benjamin Zieman, Routledge; https://doi.org/10.4324/9780203892213.

[66] Alberto Romele, Marta Severo, and Paolo Furia, 'Digital Hermeneutics: From Interpreting with Machines to Interpretational Machines', *AI & SOCIETY*, 30 June 2018, https://doi.org/10.1007/s00146-018-0856-2.

There is a difference, however, between historians who engage passively with historical content in digital form when they browse the web looking for literature and data, and those who are committed to a fully digital research process.[67] While the first will eventually produce a printed monograph, the second, still a minority, will use digitised or born-digital data, often neatly arranged in a database, analyse it with digital tools, and publish the results in the form of a website or a peer-reviewed publication supported by a dataset and code. Both categories can continue to do what historians have always done, question the origin and authenticity of a historical source by determining when it was created, by whom, for which purpose and with which means. Nevertheless, in the digital age, this perusal has to be complemented with a more technical and mathematical understanding of digital phenomena. Besides reflecting on why a particular collection of documents has been selected to be digitised and published on the web, a historian should also be able to identify the alterations and loss of context that occur when the collection is transformed from its analogue to its digital form.

[67] Zaagsma 2013, Zaagsma, Gerben "On Digital History". *BMGN - Low Countries Historical Review*, 128(4), pp.3–29, http://doi.org/10.18352/bmgn-lchr.9344.

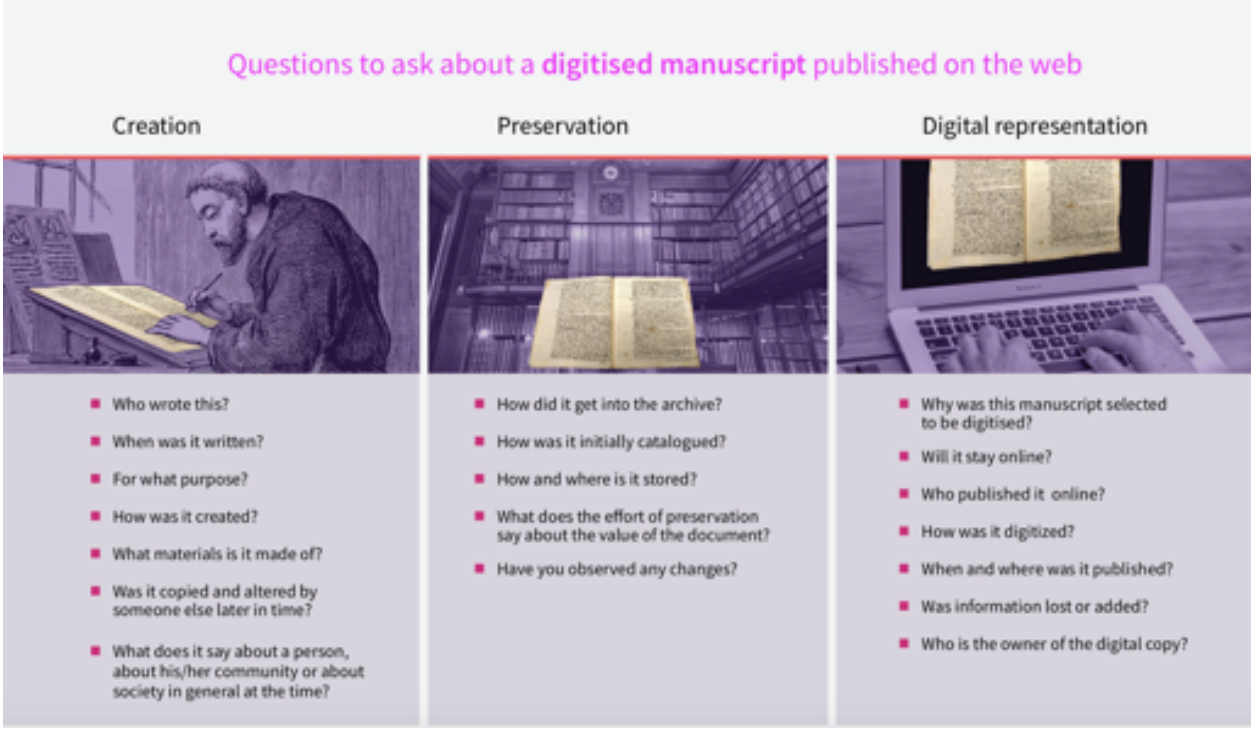

Figure 4. Visual aid showing the various contexts in which source criticism should be applied. Teaching platform for digital source criticism https://ranke2.uni.lu/ accessed 12 Febr 2020.

Another layer of manipulation that needs to be scrutinised is the selection bias of search engines that have permeated academic library systems and increasingly determine the literature that is consulted.[68] For those who 'go digital all the way', the critical appraisal of the digital dimension is more demanding, as the computer code itself needs to be criticised. As an algorithm - a command for steps that have to be taken to perform a specific task - is already a reduction of a complex reality, everything that is created through code - the data, the tool to process the data, and the website and interface to show the results of the analysis - should also be subject to 'source criticism'.[69] The choice of a particular computer language, database system or tool already steers the results in a particular direction. By applying digital hermeneutics, the historian

[68] Putnam, L, The Transnational and the Text-Searchable: Digitized Sources and the Shadows They Cast. *The American Historical Review*, *121*(2), (2016): 377–402 https://doi.org/10.1093/ahr/121.2.377; Ian Milligan, Illusionary Order: Online Databases, Optical Character Recognition, and Canadian History, 1997–2010. *Canadian Historical Review*, *94*(4) (2013): 540–569. https://doi.org/10.3138/chr.694

[69] See for an explanation and teaching aids on digital source criticism, the platform https://ranke2.uni.lu/, accessed 12 Febr 2020.

can be transparent about this process, instead of leaving the computer's assumptions and limitations unarticulated.[70] In practice, only historians with an interest in the epistemology of digital objects and processes will engage with this rigorous form of hermeneutics. For the majority, engaging with digital history will remain a hybrid mix of analogue and digital practices.[71]

IV Conclusions

Considering that computers are already ubiquitous in historical scholarship, several historians have argued that the phrase 'digital history' will disappear in the next decade or so.[72] However, considering the long history of the debates and the wide variety of technologies and debates within digital history, it is much more likely that some technologies will become main-stream methodologies within history, without making digital history main-stream per se. Many, if not all, of the above-described methods, will inevitably become more common-place in the historical discipline. Today it is hard to imagine conducting historical scholarship without technologies such as search engines, yet these technologies significantly impact

[70] See for an explanation on digital hermeneutics in practice, the website of the Doctoral Training Unit; Digital History and Hermeneutics: https://dhh.uni.lu/about-us/. See for digital tool criticism: Marijn Koolen, Jasmijn van Gorp, Jacco van Ossenbruggen; Toward a model for digital tool criticism: Reflection as integrative practice, Digital Scholarship in†the Humanities†, 12 October 2018, https://doi.org/10.1093/llc/fqy048, for data criticism see: Frederick W. Gibbs, New Forms of History: Critiquing Data and Its Representations, in: *the American Historian*, (no 7, February 2016) https://www.oah.org/tah/issues/2016/february/new-forms-of-history-critiquing-data-and-its-representations/ For algorithmic criticism see: Steven Ramsay, Algorithmic Criticism, *A Companion to Digital Literary Studies,* ed. Susan Schreibman and Ray Siemens. Oxford: Blackwell, 2008.: http://www.digitalhumanities.org/companionDLS/. For interface criticism see: *Interface Criticism;*Aesthetics Beyond Buttons (ed)Christian Ulrik Andersen & Søren Bro Pold, Interface Criticism, Aarhus Universitetsforlag, 2011. 296 p. (Acta Jutlandica, Vol. 2011/1). (Humanities Series I, Vol. 2011/1). https://www.oah.org/tah/issues/2016/february/new-forms-of-history-critiquing-data-and-its-representations/

[71] Gerben Zaagsma, 'On Digital History', *BMGN - Low Countries Historical Review* 128, no. 4 (16 December 2013): 3–29, https://doi.org/10.18352/bmgn-lchr.9344.

[72] Zaagsma.

historiography.[73] Furthermore, besides technological developments, a number of debates internal to digital history are likely to affect historical scholarship in the (near) future.

In partaking in digital history, as a methodology or practise, we engage in other research practices. Many digital history projects are conducted through cross-disciplinary collaboration between historians and computational experts (such as corpus linguists, data scientists, and research software engineers), as well as experts from GLAM-domains. This multifaceted nature of digital history research requires expertise to ask the right questions, to create a usable dataset, *and* to process the data in order to discuss the research questions. Therefore, it is increasingly difficult for historians to conduct digital historical scholarship independently. As such, digital history is likely to affect how historians publish their work - increasingly multi-authored (which this article is a reflection of) and in a digital format with accompanying accessible data- and how it is evaluated.[74]

The effect digital history will have on future historiography is thereby increasingly negotiated in such cross-disciplinary collaborations. Here historians are uncertain how they can use digital methods while computational experts are uncertain how digital methods can process historical data sets. This introduces the problem that historians as users of tools may not fully comprehend how they acquire their research results. Whether historians should blindly trust the output of a tool or discard the tool as epistemologically incompatible are both undesired consequences. As we have seen, some historians have consequently argued that historians will need to develop much more digital knowledge and learn to be programmers themselves. Others instead argue that tools should be made more understandable to historians.

---

[73] Tim Hitchcock, 'Confronting the Digital', *Cultural and Social History* 10, no. 1 (1 March 2013): 9–23, https://doi.org/10.2752/147800413X13515292098070; Putnam, 2016; Milligan, 2013.

[74] Arguing with Digital History working group, "Digital History and Argument," white paper, Roy Rosenzweig Center for History and New Media (November 13, 2017): https://rrchnm.org/argument-white-paper/; E.L. Ayers, 'Guidelines for the Professional Evaluation of Digital Scholarship in History. Technical Report', American Historical Association, 2015, https://www.historians.org/teaching-and-learning/digital-history-resources/evaluation-of-digital-scholarship-in-history/guidelines-for-the-professional-evaluation-of-digital-scholarship-by-historians; C.A. Romein, *The Challenge of Using Digital and Digitised Sources for Journals and Articles: An ECR-Editor's View*, Wiley's Digital Humanities Fest 2019 http://wileyactual.com/wileyhumanitiesfest/2019/11/13/an-ecr-editors-view/

Related to this is the debate about how to educate students as practitioners of digital history, but also as citizens of digital societies. Considering the rapid rate of technological change, and how much there already is to educate students on, the incorporation of digital history in the history curriculum is no trivial matter.[75] The technologies described in this article point to the broad directions of digital history, and nobody can be an expert in all.

Finally, an open debate is how to preserve the output of digital history sustainably . While libraries and archives have developed standards for preserving digitised material, this is not yet the case of large amounts of born-digital material (e.g. email, WhatsApp messages, Facebook), though the Web ARChive (or WARC) standard is an honourable exception. Furthermore, the technologies used by historians themselves are not sustainable, as the software is quickly outdated, abandoned, and non-functional. How to preserve digital historical scholarship results, and the processes by which to achieve effective preservation is an active area of research among historians, GLAM professionals, and computational experts.

In this article, we have only superficially described the current state of digital history. While research questions still lead historical scholarship, new methods for assembling, processing, and analysing sources as data are being implemented to investigate these questions. At the same time, we argue that scholars in digital history need to be critical of how algorithms influence the outcomes of research. The technologies described in this article have had varying degrees of effect on historical scholarship, usually in in-direct ways. Technologies such as OCR and search engines are often not directly visible in a historical argument, especially since historians tend to cite the physical archival sources.[76] However, these technologies shape

[75] T. Mills Kelly, *Teaching History in the Digital Age*, 2013, http://hdl.handle.net/2027/spo.12146032.0001.001; Anna-Maria Sichani et al., 'Diversity and Inclusion in Digital Scholarship and Pedagogy: The Case of *The Programming Historian*', *Insights* 32, no. 1 (8 May 2019): 16, https://doi.org/10.1629/uksg.465; Sharon Webb and James Baker, 'Teaching History in a Digital Age', *Historical Transactions* (blog), 12 September 2019, https://blog.royalhistsoc.org/2019/09/12/teaching-digital-history/, accessed 12 Feb 2020..

[76] Hitchcock, 2013

how historians interact with sources and whether sources can be accessed at all.[77] Other technologies have not yet diffused to the broader historical discipline; it is consequently too early to tell how they will impact research. As such, we cannot predict what the state of the field will be like in ten years; there are too many directions for future research questions and implementations of digital technology. External pressure towards increasing open access[78] as well as technological developments such as artificial intelligence may furthermore stimulate the digital history, with historians increasingly opening-up the underlying sources and methods for use by the wider public or by computers. There is one certainty: the field will look very different from today.

Further reading

- Bod, R. (2013). *A new history of the humanities: The search for principles and patterns from antiquity to the present*. Oxford University Press.
- Dougherty, J., & Nawrotzki, K. (2013). *Writing history in the digital age*. University of Michigan Press.
- Jockers, M. L. (2013). *Macroanalysis: Digital methods and literary history*. University of Illinois Press.
- Graham, S., Milligan, I., & Weingart, S. (2015). *Exploring big historical data: The historian's macroscope*. World Scientific Publishing Company.
- Guldi, J., & Armitage, D. (2014). *The history manifesto*. Cambridge University Press.

Short bio's on contributors

[77] Julia Laite, 'The Emmet's Inch: Small History in a Digital Age', *Journal of Social History*, accessed 30 October 2019, https://doi.org/10.1093/jsh/shy118; Roopika Risam, *New Digital Worlds: Postcolonial Digital Humanities in Theory, Praxis, and Pedagogy* (Northwestern University Press, 2019), https://doi.org/10.2307/j.ctv7tq4hg.
[78] https://www.coalition-s.org/, accessed 12 Febr 2020.

Dr James Baker is a Senior Lecturer in Digital History and Archives at the University of Sussex (UK) and at the Sussex Humanities Lab. He is a Software Sustainability Institute Fellow, a Fellow of the Royal Historical Society, a convenor of the Institute of Historical Research Digital History seminar, a member of The Programming Historian Editorial Board and a Director of ProgHist Ltd. He is an expert in the authority of the digital record, the history of knowledge organisation, historical interactions with information technologies, and the history of the printed image. His research is funded by the Arts and Humanities Research Council (UK), British Academy, British Council, and the European Commission. Contact: james.baker@sussex.ac.uk.

Dr Julie M. Birkholz is a digital humanities network researcher. She is currently a Post-doctoral researcher on the ERC WeChangEd Research Project at the Department of Literary Studies, Ghent University, Belgium. Her research focuses on the use of the network perspective in digital collections, and specifically periodicals- from building digital pipelines, to extracting and analysing networks from text; where she uses a quantitative social network analysis approach to systematically compare contexts, structures and outcomes of understanding the networks of social actors of the past. Contact: Julie.Birkholz@UGent.be.

Michel de Gruijter, MA is Coordinator Editorial Office for the Digital Library of Dutch Literature (DBNL) and was Project Advisor for the Researcher-in-Residence in 2019, both at the National Library of the Netherlands (KB). He is specialised in full-text digitisation of printed material and holds an MA degree in Early Modern Dutch Literature. Contact: michel.degruijter@kb.nl.

Dr Max Kemman is a researcher/consultant at the Dutch firm Dialogic - Innovation & Interaction, where he works on research questions related to ICT innovations in the public sphere. In 2019 he defended his PhD thesis *Trading Zones of Digital History* at the University of Luxembourg. His PhD research encompassed an ethnographic study of historians collaborating with computational experts, investigating

the question of how historian's practices are affected by such cross-disciplinary interactions. Contact: kemman@dialogic.nl.

Dr Albert Meroño-Peñuela is a Postdoc Research Fellow at the Knowledge Representation & Reasoning Group of the Vrije Universiteit Amsterdam. He studies the construction of knowledge bases and their role in Digital Humanities scholarly processes. His broad interests include knowledge graphs, Linked Data, and Web query languages. He is also involved in CLARIAH, the most extensive research infrastructure for Humanities in the Netherlands, and serves in the CLARIAH Technical Board as coordinator of the LOD Interest Group Contact: albert.merono@vu.nl.

Dr Thorsten Ries is a Postdoc Research Fellow at Ghent University (FWO), currently Guest Professor at Antwerp University in German Literature, and a Research Associate at the Sussex Humanities Lab. He is a specialist in Born-digital Genetic Criticism, Digital Forensics, Born-digital Archives and Digital History. He uses digital forensic methods to analyse born-digital material in literary archives and explored forensic perspectives in digital historical scholarship during a Marie-Sklodowska-Curie Fellowship at the University of Sussex. He received his PhD in German Literature from Hamburg University and Ghent University (Joint PhD; 2013). Contact: Thorsten.Ries@UGent.be.

Dr C. Annemieke Romein is an early modernist and digital historian. She is currently a Postdoctoral researcher on a Rubicon-project (NWO) called 'Law and Order: Low Countries?!' at the Department of History, Ghent University; she is also an associate researcher of the Erasmus University Rotterdam. From May until October 2019, she was a Researcher-in-Residence at the KB National Library of the Netherlands on a DH-project called: 'Entangled Histories'. All of her current projects revolve around legislation in the early modern era (±1500-1750), (automatic) meta-dating these texts and connecting them through time (topic-wise) and space (geographically). Contact: info@caromein.nl.

Ruben Ros, MA, is a postgraduate student at Utrecht University. He focuses on computational methods and conceptual history. In his Master's thesis, he traced the emergence of the Dutch concept of 'buitenland' in digitised newspapers. He has worked as a research assistant at Helsinki Computational History Group (COMHIS) and the Digital Humanities Lab of the KNAW Humanities Cluster (Amsterdam) in the Netherlands. Contact: ruben@rubenros.nl.

Dr Stefania Scagliola is a postdoctoral researcher at the Centre for Contemporary and Digital History at the University of Luxembourg, where she created the teaching platform for digital source criticism www.ranke2.uni.lu. She is a specialist in digital pedagogy, oral history and military history and has initiated several projects in which new technology is integrated into traditional historical scholarship. Contact: Stefania.scagliola@uni.lu