History
The Journal of the Historical Association History

State of the Field: Digital History

C. ANNEMIEKE ROMEIN
Erasmus University Rotterdam/Ghent University/KNAW Huygens ING

MAX KEMMAN
Dialogic

JULIE M. BIRKHOLZ
Ghent University

JAMES BAKER
University of Sussex

MICHEL DE GRUIJTER
KB National Library of the Netherlands

ALBERT MEROÑO-PEÑUELA
Vrije Universiteit Amsterdam

THORSTEN RIES
Ghent University

RUBEN ROS
Utrecht University

STEFANIA SCAGLIOLA
Luxembourg University

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, focusing 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. The third section considers 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

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. Funding was provided by Fonds Wetenschappelijk OnderzoekHard Drive Philology / Source Code Philology. TracTracing the digital writing and coding process in German literature: M. Beyer, M. Speier, F. Kittler, J. Piringer, H. Bajohr / G. Weichbrodt.

scholarship, usually in indirect ways. With this article we aim to take stock of the digital approaches and methods used in historical scholarship in order to provide starting points for scholars seeking to understand the digital turn in the field and how and when to implement such approaches in their work.

The use of computers in historical scholarship is not new, although the impact on the field has shifted over time. 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 was seen as having strayed too far from the traditional questions and methods of history.[1] 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.[2] 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*.[3] Such 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'.[4]

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.[5] With the mass-digitisation of libraries and archives under way 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

[1] 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*, 31/4 (1 1998), pp. 141–48, <https://doi.org/10.1080/01615449809601196>.

[2] Onno Boonstra, Leen Breure and Peter Doorn, 'Past, present and future of historical information science', *Historical Social Research/Historische Sozialforschung*, 29/2 (2004), pp. 4–132, <https://www.jstor.org/stable/20761957>.

[3] Still online at <http://valley.lib.virginia.edu/> [accessed 18 Oct. 2019].

[4] 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 Pat Hudson and Mina Ishizu, *History by Numbers: An Introduction to Quantitative Approaches* (London, 2016).

[5] Jane Winters, 'Digital history', in Marek Tam and Peter Burke (eds), *Debating New Approaches to History* (London, 2018), pp. 277–300; Kristen Nawrotzki and Jack Dougherty, *Writing History in the Digital Age* (Ann Arbor, 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* (Abingdon, 2013), pp. 1–20.

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.[6]

Furthermore, this suggests that computers 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, including 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 who look across institutional and disciplinary boundaries to engage in historical practices with the methodological and epistemological concepts of other disciplines.[7] 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.[8] The ambition of such pursuits 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 'mainstream' and to reflect on the ever-developing influence of digital history. We do not claim that our discussion presents 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. But 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. The article, therefore, while discussing separate techniques, is centrally concerned with exploring how the digital humanities have evolved and what that evolution might have brought to historical scholarship. We begin by discussing techniques that generate and secure data and make them machine searchable, such as OCR/HTR (defined below) and born-digital archives, computer vision, scholarly editions, and linked open data, before moving on to examine how data is made more accessible by quantitative text and network analysis. We also

[6] Bob Nicholson, 'The digital turn', *Media History*, 19/1 (2013), pp. 59–73, <https://doi.org/10.1080/13688804.2012.752963>.

[7] Max Kemman, 'Boundary practices of digital humanities collaborations', *DHBenelux Journal*, 1 (2019), pp. 1–24, <http://journal.dhbenelux.org/journal/issues/001/Article-Kemman/Article-Kemman.pdf> [accessed 12 Feb. 2020].

[8] Susan Schreibman, Ray Siemens and John Unsworth, *Companion to Digital Humanities* (Oxford, 2004), <http://www.digitalhumanities.org/companion/> [accessed 12 Feb. 2020].

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

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 paragraphs, 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.

Optical Character Recognition and Handwriting Text Recognition

Written documentation is core to our work as historians. Neither printed nor handwritten texts 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 they had been trained on.

Whereas OCR is applied to standard fonts and 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 hands can be trained through OCR-programs,[9] the results generated, for example, by the READ-Coop's HTR tool Transkribus are promising.[10] What separates Transkribus – a commonly used platform for the automated recognition, transcription and searching

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

of historical documents, from OCR-engines is the learning curve.[11] For example, the more transcribed pages that are added, the better language patterns are understood; resulting in Character Error Rates (CER) between 10% and 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.[12]

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.[13] Increasingly, both techniques are used for digitising collections, with the quality and thus capacity to read/recognise texts continuing to improve incrementally with 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,[14] 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 front-end, 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 archives,[17]

[11] <http://transkribus.eu> [accessed 12 Feb. 2020].

[12] Additional on HTR: Guenter Muehlberger et al., 'Transforming scholarship in the archives through handwritten text recognition', *Journal of Documentation*, 75/5 (2019), pp. 954–76, <https://doi.org/10.1108/JD-07-2018-0114>.

[13] 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 layout, the computer cannot recognise structure (yet). 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.

[14] Named Entity Recognition (NER) are pre-defined categories within unstructured texts, for example (but not limited to): persons, locations, time expressions.

[15] Thorsten Ries and Gábor Palkó, 'Born-digital archives', *International Journal of Digital Humanities*, 1/1 (2019), pp. 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*, 47/3 (2019), pp. 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, *The SAGE Handbook of Web History* (London, 2018).

[16] Patrícia Falcão and Tom Ensom, 'Conserving digital art', in Tula Giannini and Jonathan P. Bowen (eds), *Museums and Digital Culture: New Perspectives and Research* (Cham, 2019), pp. 231–51, <https://doi.org/10.1007/978-3-319-97457-6_11>.

[17] Abigail De Kosnik, *Rogue Archives: Digital Cultural Memory and Media Fandom* (Cambridge, MA, 2016); Sharon Webb, '"Digital archives in communities – practice and preservation" : a summary (or at least an attempt) – Digital Preservation Coalition', <https://www.dpconline.org/blog/digital-archives-in-communities; Ian Milligan, 'Finding Community in the Ruins of Geo

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, 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, and the resulting possible digital forensic analytical angles complicate data recovery. They equally complicate 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 documenting findings for the research public fairly complicated.[23] Digital archivists need to deal with challenges ranging from considering the ethics of dark archives, saving the content of online communities and cultures to the archaeological recovery of long-gone websites from offline backups. They also have to consider and document possible misrepresentations, lacunae and imbalances in these born-digital archive collections.[24]

As a consequence, researchers and archivists working with born-digital archives not only need data-mining and visualisation tools, such as

Cities:Distantly Readinga WebArchive'(Institute of Electrical and Electronics Engineers, 2015), https://uwspace.uwaterloo.ca/handle/10012/11650> [accessed 30 Oct.2019].

[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*, 16/3 (2013), pp. 197–214, <https://doi.org/10.1080/13645579.2013.774172>.

[20] Niels Brügger, *The Archived Web* (Cambridge, MA, 2018); Eveline Vlassenroot et al., 'Web archives as a data resource for digital scholars', *International Journal of Digital Humanities*, 1/1 (2019), pp. 85–111, <https://doi.org/10.1007/s42803-019-00007-7> [accessed 12 Feb. 2020].

[21] Trevor Owens, 'Historic iPhones: personal digital media devices in the collection', *Trevor Owens* (blog), 15 Nov. 2013, <http://www.trevorowens.org/2013/11/historic-iphones-personal-digital-media-devices-in-the-collection/> [accessed 12 Feb. 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 Feb. 2020].

[23] Ibid.

[24] Johan van der Knijf, 'Recovering '90s data tapes. experiences from the KB Web Archaeology project', paper on iPres 2019, <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].

Archives Unleashed,[25] and the data-mining functionality in *BitCurator,*[26] but also need 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, such as. Susan Thomas and Jeremy L. John, the major shift that marked the rise of the born-digital studies was the publication of Matthew Kirschenbaum's seminal book *Mechanisms: New Media and the Forensic Imagination*.[28] In the following years, Kirschenbaum's and Doug Reside's studies became paradigmatic showcases for digital forensic work on personal born-digital archives, as well as for forensic standards in born-digital primary records archiving. Their work was accompanied by large international projects on born-digital archiving in the GLAM sector (Galleries, Libraries, Archives and Museums), leading to the development of archival sector-specific methods and toolsets (such as BitCurator). 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]

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] All this suggests that matters are moving in a direction where forensic detection of digital disinformation, 'deep fake' and forgery, automated content generation and bots, online threat, malware and hacking will play an increasingly important role in born-digital preservation, archiving and web history research.[32] Ecological considerations about the carbon footprint of data management will probably also become a focus for researchers.[33]

[25] Ian Milligan, 'The Archives Unleashed Project', <https://archivesunleashed.org/> [accessed 26 Oct. 2019].
[26] Bitcurator Consortium: *Bitcurator*, <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, 2018), pp. 593–606.
[28] Matthew G. Kirschenbaum, *Mechanisms: New Media and the Forensic Imagination* (Cambridge, MA, 2008).
[29] The term of the 'e-palaeographer' was coined by R. J. Morris (eds), 'Electronic documents and the history of the late twentieth century: black holes or warehouses?', in Edwards Higgs (ed.), *History and Electronic Artefacts* (Oxford, 1998), pp. 31–8, at p. 33.
[30] Matthew Kirschenbaum, 'The .txtual condition: digital humanities, born-digital archives, and the future literary', *Digital Humanities Quarterly*, 7/1 (2013), <http://digitalhumanities.org/dhq/vol/7/1/000151/000151.html>.
[31] Anne Helmond, 'Track the trackers', <https://wiki.digitalmethods.net/Dmi/DmiWinterSchool2012TrackingTheTrackers> [accessed: 26 Oct. 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*, 1/1 (2019), pp. 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*, 22/1 (2017), <https://doi.org/10.5210/fm.v22i1.7263>;

Computer Vision

While 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, and it is in the context of these developments that computer vision has been used in the service of historical scholarship.[34] The questions asked 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?

A significant milestone in the use of these techniques for historical 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, notably *The Illustration Archive* (2015) which used crowdsourcing, machine tagging and similarity matching to enhance the discovery of images, to link them and to make legible, in visual terms, the larger patterns in pre-twentieth century book illustration. To isolate illustrations for use in their digital archive, *The Illustration Archive* team used page-level XML (see the discussion of digital scholarly editions below) 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, 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.[37]

Keith L. Pendergrass et al., 'Toward Environmentally Sustainable Digital Preservation', *The American Archivist*, 82/1 (2019), pp. 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, <https://www.techradar.com/uk/news/intel-is-bringing-facial-recognition-to-toky-2020> [accessed 12 Feb. 2020]; James Griffiths, *The Great Firewall of China* (London, 2019).

[35] While 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 Feb. 2020]; Cerys Jones et al., 'Leonardo brought to light: multispectral imaging of drawings by Leonardo Da Vinci', *Zenodo*, 12 March 2018, <https://doi.org/10.5281/zenodo.1208430> [accessed 12 Feb. 2020].

[36] Lev Manovich, 'How to compare one million images?', in David M. Berry (ed.), *Understanding Digital Humanities* (London, 2012), pp. 249–78, <https://doi.org/10.1057/9780230371934_14>.

[37] <http://illustrationarchive.cf.ac.uk/> [accessed 28 Oct. 2019]; Julia Thomas, *Nineteenth-Century Illustration and the Digital: Studies in Word and Image* (London, 2017),

Work on historical images is advancing quickly. The use of convolutional neural networks, a machine-learning approach commonly used to detect and classify features of visual inputs, is powering recent step-changes in computer vision. Wevers and Smits' landmark 2019 work showed how this technique could be used to enrich our understanding of trends in historical corpora.[38] Taking over a century of Dutch newspapers as their source material, Wevers and Smits detected their non-textual elements, 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). In doing so, Wevers and Smits provide a much-needed pathway towards a scalable and historically relevant computational analysis of images by informational content. This offers the prospect of a digital history and 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 just as central to today's digital scholarly editions.[39] The main difference, however, is that, freed from the constraints of the printing press, 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.[40]

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,

<https://doi.org/10.1007/978-3-319-58148-4>; 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/>; 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 Sept. 2018, <https://doi.org/10.5281/zenodo.1412137>.

[38] Melvin Wevers and Thomas Smits, 'The visual digital turn: using neural networks to study historical images', *Digital Scholarship in the Humanities*, <https://doi.org/10.1093/llc/fqy085> [accessed 30 Oct. 2019].

[39] 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].

[40] Marita Mathijsen, *Naar de letter. Handboek editiewetenschap.* (Den Haag, 2010), pp. 19–29 and i–vi, <https://www.dbnl.org/tekst/math004naar03_01/> [accessed 12 Feb. 2020].

scholars demand more open, reliable and standardised digital editions. Some vast text archives, such as *Gallica,* offer scholars scanned images of document pages, but the full-text layer may, if the result of automated OCR (see above) is unsatisfatory, not meet scholarly standards of a reliable, citable scholarly resource.[41] 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.[42] 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. 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.[43]

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 of 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.[44] Infrastructures like TEI have both democratised the practice of scholarly editing and given scholarly editors a platform from which to fulfil the intellectual ambitions of this enduring

[41] Thomas Crombez, 'Digitale deemstering. Auteursrecht en de digitalisering van boeken in Nederland en Vlaanderen', *Vooys. Tijdschrift voor letteren*, 37/3 (2019), pp. 48–9. Gallica contains .txt, .pdf and .jpg files of the source.

[42] <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 Feb. 2020].

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

[44] 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, instance.g. graph-based 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 supported by browsers in the long term.

genre of humanities practice. New infrastructures must be developed according to the same principles.[45]

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'; of which LOD served as a technique to describe knowledge.[46] LOD standards afford a way to make (meta)data on and of objects available and publicly accessible in a format readable by both humans and machines. Thus instead of referencing an unstructured description of a place, person or object, for example 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 triplet 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 Knowledge Graph, 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.[47] The structuring of information in this way is also the backbone of 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 1).

The 5-star Linked Data rating system 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 by cross-datasets through URIs.[48] This cross-dataset linkage encourages data reuse, preventing

[45] Patrick Sahle, '2. What is a scholarly digital edition?', in Matthew James Driscoll and Elena Pierazzo (eds), *Digital Scholarly Editing: Theories and Practices* (Cambridge, 2017), pp. 19–39, <http://books.openedition.org/obp/3397>.
[46] <https://www.w3.org/2001/sw/> [accessed 10 Oct. 2019].
[47] <https://googleblog.blogspot.com/2012/05/introducing-knowledge-graph-things-not.html> [accessed 30 Oct. 2019].
[48] <https://www.ted.com/talks/tim_berners_lee_on_the_next_web?language=nl> [accessed 10 Oct. 2019].

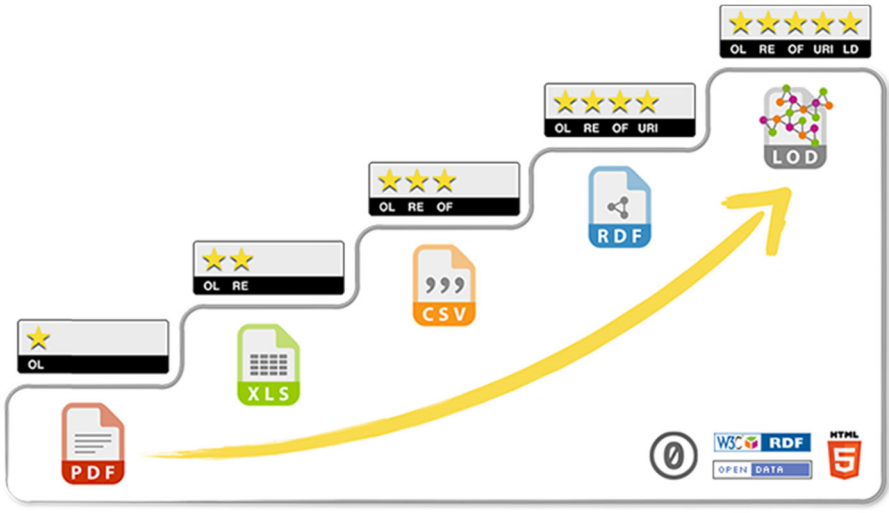

Figure 1 Open Data. [Colour figure can be viewed at wileyonlinelibrary.com]
Source: <https://5stardata.info/en/> [accessed 10 Oct. 2019].
Explanation: *OL = OpenLicence; *RE = machine REadable; *OF = OpenFormat; *URI = Uniform Resource Identifiers; *LD = LinkedData.

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 (Figure 2). 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.[49] 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. This consequently makes it difficult to implement an automatic process of RDF generation on complete collections; even so, a number of large-scale infrastructures are in progress.[50] Despite the potential of RDF, its use remains a technical barrier for many, a problem which has led to a discussion on emphasising usability for non-technical users through *Linked Open Usable Data.*

[49] Here is a 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.
[50] Rinke Hoekstra et al., 'The DataLegend ecosystem for historical statistics', *Journal of Web Semantics*, 50 (2018), pp. 49–61, <https://doi.org/10.1016/j.websem.2018.03.001>.

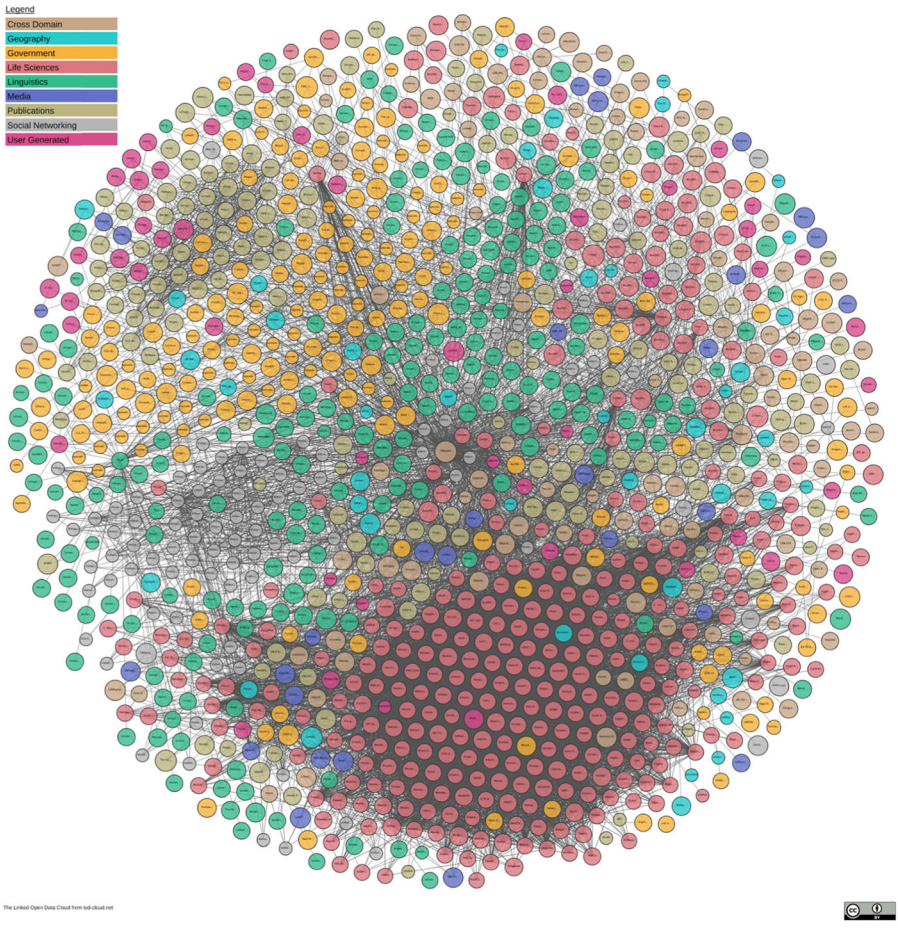

Figure 2 Linking open data cloud diagram 2020, by Max Schmachtenberg et al., <https://lod-cloud.net/> [accessed 12 Feb. 2020]. [Colour figure can be viewed at wileyonlinelibrary.com]

II

The increasing availability of digitised sources, either born-digital or made machine-readable, affords efficient assessment of sources. For example, it makes possible 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; that is, 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.[51] Announced in 2011, the tool was presented as a revolutionary new way of looking at culture.[52] Since then its capacity to offer a rapid overview of a word's frequency has become essential in studying historical phenomena.[53]

Frequency-based tools and methods are, however, not without their problems. Right from the outset many scholars pointed to the pitfalls of Google's Ngram Viewer Their critiques often apply to other frequency-based methods and fall into three categories.[54] 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 contextualisation through structured metadata: knowing who wrote what, when and in which context is essential to 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'.[55] 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

[51] Corpus linguists often refer to counted words as '*n*-grams': sequences of *n* words.
[52] Jean-Baptiste Michel et al., 'Quantitative analysis of culture using millions of digitized books', *Science*, 331/6014 (2011), pp. 176–82, <https://doi.org/10.1126/science.1199644>.
[53] Paul Caruana-Galizia, 'Politics and the German language: testing Orwell's hypothesis using the Google N-Gram Corpus', *Digital Scholarship in the Humanities*, 31/3 (2016), pp. 441–56, <https://doi.org/10.1093/llc/fqv011>.
[54] 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*, 10/10 (2015), e0137041, <https://doi.org/10.1371/journal.pone.0137041>.
[55] Jan Ifversen, 'About key concepts and how to study them', *Contributions to the History of Concepts*, 6/1 (2011), pp. 65–88, <https://doi.org/10.3167/choc.2011.060104>.

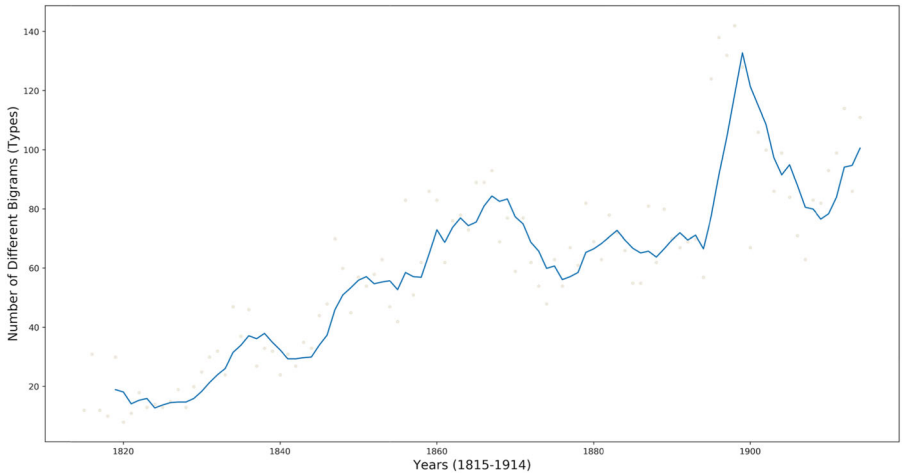

Figure 3 The (absolute) number of different bigrams that contain the adjective 'binnenlandsche' ('domestic') in Dutch newspapers between 1815 and 1914. [Colour figure can be viewed at wileyonlinelibrary.com]
Source: R. S. Ros, 'The birth of the foreign : a digital conceptual history of buitenland in Dutch newspapers 1815–1914', Master's thesis (2019), <http://dspace.library.uu.nl/handle/1874/382176> [accessed 12 Feb. 2020].

trade' (Figure 3).[56] 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.

Future research in historical textual data will probably involve better contextualisation through structured metadata. Full texts are not sufficient by themselves. To use them as historical data, researchers 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

[56] 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 (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 Feb. 2020].

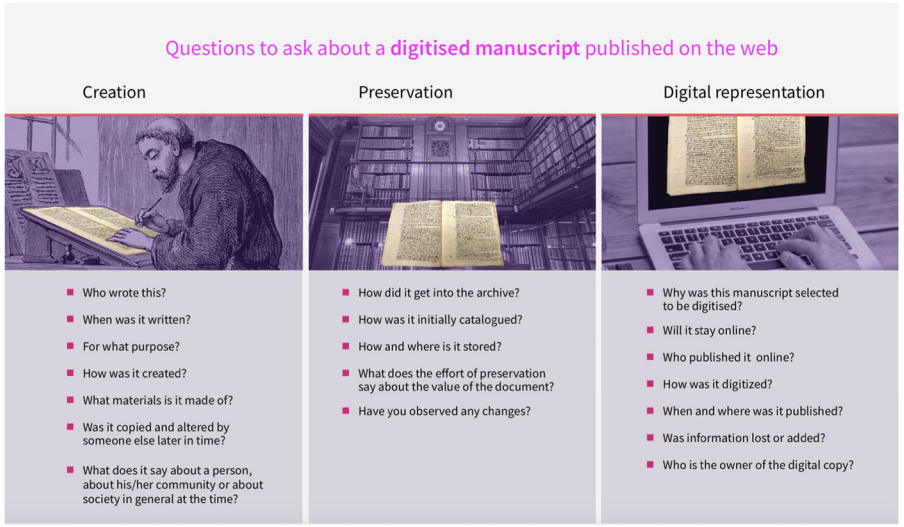

Figure 4 Visual aid showing the various contexts in which source criticism should be applied. Teaching platform for digital source criticism. [Colour figure can be viewed at wileyonlinelibrary.com]
Source: <https://ranke2.uni.lu/> [accessed 12 Feb. 2020].

entities' such as people, places and organisations is instrumental in gaining a better insight into historical texts.[57]

Network Approach and Analysis

One way of reconstructing and retracing history is through the reconstruction of the networks of the past. As research on social networks has shown, these networks mattered: the position one had in a social network influenced one's power and performance, as well as the structure of the relations that lent social, economic and political capital to individuals and organisations. Network analysis as a method has been used to analyse 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, for example in the work of Padgett and Ansell on the Medici networks in the early 1400s.[58] 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

[57] 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 Feb. 2020].
[58] John F. Padgett and Christopher K. Ansell, 'Robust action and the rise of the Medici, 1400–1434', *American Journal of Sociology*, 98/6 (1993), pp. 1259–1319, <https://doi.org/10.1086/230190>.

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 seventeenth and eighteenth centuries was projected as networks of senders and receivers to reconstruct communication flows during the Age of Enlightenment.[59]

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 (Linked Data), and digitisation and transcriptions of textual sources (Section I) have opened up an avenue of (semi-)automatic identification of historical networks, for example through written correspondence, manuscripts and printed materials such as books, newspapers or periodicals.[60] These approaches have resulted in the ability to investigate more entities (i.e. 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 an excellent starting point for positioning relational research questions in different periods, contexts or entities.[61]

III

Historians need to be aware of the origin and authenticity of the data they use and of what has been included and excluded in their preservation and 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

[59] Giovanna Ceserani and Thea De Armond, 'British architects on the Grand Tour in eighteenth-century Italy: travels, people, places', <https://purl.stanford.edu/ct765rs0222> [accessed 30 Oct. 2019].

[60] 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, OR, 2011), pp. 61–9, <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*, 27/1 (2013), pp. 75–78, <https://doi.org/10.1007/s13218-012-0225-0>.

[61] Marten Düring, 'Historical network research: network analysis in the historical disciplines' (2017), <http://historicalnetworkresearch.org/> [accessed 12 Feb. 2020].

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.

Digital Hermeneutics

The term 'hermeneutics', coined by the nineteenth-century German historian Droysen to emphasise the importance of 'interpretation' in constructing historical knowledge, has been reconceptualised in 'digital hermeneutics', in the light of the need to reflect on how computers influence the construction of scientific knowledge.[62]

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 in order to understand the potential of computers in extracting meaning from classical texts.[63] Just as in philology, the practice of applying source criticism to classical texts is the origin of source criticism in the realm of history. In turn this contributed to the archival turn at the end of the nineteenth 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 twenty-first 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. The 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.

[62] Philippe Mueller, 'Understanding history: hermeneutics and source criticism in historical scholarship', in Miriam Dobson and Benjamin Zieman (eds), *Reading Primary Sources, the interpretation of texts from nineteenth and twentieth century* (Abingdon, 2008), <https://doi.org/10.4324/9780203892213>.

[63] 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.[64] 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 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.

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.[65] 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' (Figure 4).[66] 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 can be transparent about this process, instead of leaving the computer's assumptions and limitations unarticulated.[67] In

[64] Gerben Zaagsma, 'On digital history', *BMGN – Low Countries Historical Review*, 128/4 (2013), pp. 3–29, <https://doi.org/10.18352/bmgn-lchr.9344>.

[65] L. Putnam, 'The transnational and the text-searchable: digitized sources and the shadows they cast', *The American Historical Review*, 121/2 (2016), pp. 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), pp. 540–569, <https://doi.org/10.3138/chr.694>.

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

[67] 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 and Jacco van Ossenbruggen, 'Toward a model for digital tool criticism: reflection as integrative practice, digital scholarship in the humanities', 12 Oct. 2018, <https://doi.org/10.1093/llc/fqy048>, for data criticism, see Frederick W. Gibbs, 'New forms of history: critiquing data and its representations', *the American Historian* no. 7 (Feb. 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', in Susan Schreibman and Ray Siemens (eds), *A Companion to Digital Literary Studies* (Oxford, 2008),

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.[68]

IV

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.[69] However, in view of 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 mainstream methodologies within history, without making digital history mainstream per se. Many, if not all, of the above-described methods, will inevitably become more commonplace in the historical discipline. Today it is hard to imagine conducting historical scholarship without technologies such as search engines, yet these technologies significantly impact historiography.[70] Furthermore, besides technological developments, a number of debates internal to digital history are likely to affect historical scholarship in the (near) future.

In using digital history, as a methodology or practice, 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 in terms of multi-authored articles (of which this article is a reflection) and the digital format, with accompanying accessible data, and how it is evaluated.[71]

<http://www.digitalhumanities.org/companionDLS/>. For interface criticism, see Christian Ulrik Andersen & Søren Bro Pold, *Interface Criticism: Aesthetics beyond Buttons* (Aarhus, 2011), p. 296, <https://www.oah.org/tah/issues/2016/february/new-forms-of-history-critiquing-data-and-its-representations/>.

[68] Gerben Zaagsma, 'On digital history', *BMGN – Low Countries Historical Review*, 128/4 (2013), pp. 3–29, <https://doi.org/10.18352/bmgn-lchr.9344>.

[69] Ibid.

[70] Tim Hitchcock, 'Confronting the digital', *Cultural and Social History*, 10/1 (2013), pp. 9–23, <https://doi.org/10.2752/147800413X13515292098070>; Putnam, 'The transnational and the text-searchable'; Milligan, 'Illusionary order'.

[71] Arguing with Digital History working group, 'Digital History and Argument', white paper, Roy Rosenzweig Center for History and New Media (13 November 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,

The effect digital history will have on future historiography is thereby increasingly negotiated through cross-disciplinary collaborations. Here historians are uncertain how they can use digital methods while computational experts are uncertain how digital methods can process historical datasets. This introduces the problem that historians as users of tools may not fully comprehend how they acquire their research results. We would argue that it is undesirable both that historians should blindly trust the output of a tool or discard the tool as epistemologically incompatible. 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.

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 is already involved in educating students, the incorporation of digital history in the history curriculum is no trivial matter.[72] The technologies described in this article point to the broad directions of digital history, and nobody can be an expert in all.

Finally, there is an open debate on 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 for large amounts of born-digital material (e.g. email, WhatsApp messages, Facebook), although the Web ARChive (or WARC) standard is an honourable exception. Furthermore, the technologies used by historians themselves are not sustainable, as the software quickly becomes 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

---

<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/>.

[72] 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*, 32/1 (2019), p. 16, <https://doi.org/10.1629/uksg.465>; Sharon Webb and James Baker, 'Teaching history in a digital age', *Historical Transactions* (blog), 12 Sept. 2019, <https://blog.royalhistsoc.org/2019/09/12/teaching-digital-history/> [accessed 12 Feb. 2020].

in indirect 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.[73] However, these technologies shape how historians interact with sources and whether sources can be accessed at all.[74] 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' time; there are too many directions for future research questions and implementations of digital technology. External pressure towards increasing open access as well as technological developments such as artificial intelligence may furthermore stimulate digital history, with historians increasingly opening up the underlying sources and methods for use by the wider public or by computers.[75] There is one certainty: the field will look very different from today.

FURTHER READING

- Bod, R., *A New History of the Humanities: The Search for Principles and Patterns from Antiquity to the Present* (Oxford: Oxford University Press, 2013).
- Dougherty, J. and Nawrotzki, K., *Writing History in the Digital Age* (Ann Arbor: University of Michigan Press, 2013).
- Graham, S., Milligan, I. and Weingart, S., *Exploring Big Historical Data: The Historian's Macroscope* (Singapore: World Scientific Publishing Company, 2015).
- Guldi, J. and Armitage, D., *The History Manifesto* (Cambridge: Cambridge University Press, 2014).
- Jockers, M. L., *Macroanalysis: Digital Methods and Literary History* (Champaign: University of Illinois Press, 2013).

[73] Hitchcock, 'Confronting the digital'.
[74] Julia Laite, 'The emmet's inch: small history in a digital age', *Journal of Social History*, <https://doi.org/10.1093/jsh/shy118> [accessed 30 Oct. 2019]; Roopika Risam, *New Digital Worlds: Postcolonial Digital Humanities in Theory, Praxis, and Pedagogy* (Evanston, IL, 2019), <https://doi.org/10.2307/j.ctv7tq4hg>.
[75] <https://www.coalition-s.org/> [accessed 12 Feb. 2020].