CHAPTER 7

All the Work that Makes It Work

Digital Methods and Manual Labour

Johan Jarlbrink

Automation is a temptation and a promise, and perhaps a threat. Old jobs disappear as robots and software do what human workers used to. Is this also the case with research within the humanities? Computers can process datasets of texts so large that it would take several lifetimes for scholars just to read it. Computers are excellent in finding patterns that are hard to recognise for human eyes and brains. What should researchers do when computers are much better in doing what scholars used to?

In this chapter, I will argue that digital research is far from automatised.[1] A human being is still needed to make sense of results, of course. I will focus on something else, not on the creative ways in which scholars interpret data outputs, but on the dull tasks that make data outputs possible. Most datasets need cleaning, editing and error checking. The outcome of automatic processes needs to be examined by someone who goes through the results; sometimes it needs to be corrected manually. Such procedures are often left out completely or only mentioned in brief when digital methods are discussed. Yet, they have a significant impact on results and need to be taken seriously.

I will mainly focus on various forms of text analysis, based on my own experiences and what colleagues have told me, as well as cases described in the

How to cite this book chapter:
Jarlbrink, J. (2020). All the work that makes it work: Digital methods and manual labour. In M. Fridlund, M. Oiva, & P. Paju (Eds.), *Digital histories: Emergent approaches within the new digital history* (pp. 113–126). Helsinki: Helsinki University Press. https://doi.org/10.33134/HUP-5-7

literature. The cases are meant to shed light on an often neglected part of digital methodologies, but the mundane aspects of data cleaning and curation are also significant beyond the field of digital humanities. Such procedures can be understood as 'a crucial part of the materiality of how scholarly and scientific work is done'.[2] The manual work needed to feed, improve and evaluate digital processing belongs to a long history of little tools and (supposedly) insignificant back-end operations that have made different kinds of research output possible. Digital scholarship, as traditional archival research and experimental work in laboratories, involves material and conceptual actors as well as human ones.[3]

In the first section, I will give a short background and explain why I think manual digital work matters. Three empirical sections will exemplify various kinds of manual operations. First, I describe human-assisted computational analysis in the humanities in the 1950s, 1960s and 1970s. Second, I present my own experiences from a project based on 19th-century newspapers. Third, I tell the story of how a colleague of mine used digital Named Entity Recognition (NER) in combination with pen and paper.

Invisible Work

Glimpses of the manual work that makes digitisation and computational analysis possible are sometimes given by accident. Google Books preserves a large part of our printed cultural heritage in a digital form, but also some of the hands that were needed to operate the scanners and handle the printed volumes. Just as the secretaries of the early 20th century, who left traces of themselves in the typewritten texts only as a result of errors, accidents make Google employees become visible in the digital database. Index fingers covered in condom-like pink gloves are included in many of the images now available online. They serve as a reminder of the people and work that feed the digital infrastructures.[4] Part of the workforce digitising printed materials is less visible. Much of the post-processing needed to produce useful digital surrogates is being outsourced to companies hiring low-wage workers in Cambodia and India.[5]

This kind of hidden work makes digitisation seem more straightforward and automatised than it is. The same goes for various forms of computer-assisted analysis. Tamraparni Dasu and Theodore Johnson has stated that:

> In our experience, the tasks of exploratory data mining and data cleaning constitute 80% of the effort that determines 80% of the value of the ultimate data mining results. Data mining books … provide a great amount of detail about the analytical process and advanced data mining techniques. However they assume that the data has already been gathered, cleaned, explored, and understood.[6]

Much of the cleaning can be done with software. Even an easy-to-use program such as Microsoft Excel allows you to search and replace, filter, merge, separate

and delete different kinds of data. More advanced or custom-made tools allow you to fine-tune the process. Still, such procedures need to be monitored in order to ensure the quality of the outcome. Sometimes software fail, and sometimes they need assistance from human pattern recognition. With a limited dataset, it can be more efficient to correct and edit by hand instead of spending time finding and running a software that will require additional and manual error checking anyway.

Algorithms solve problems according to specified rules. That is why they may be of limited use if a dataset is noisy and patterns are irregular. 'Signals are always surrounded by noise, even to the extent that we cannot always decipher which is which.'[7] Hadley Wickham explains (alluding to Leo Tolstoy) that 'tidy datasets are all alike but every messy dataset is messy in its own way'.[8] A dataset can be corrupt in numerous ways, but there is only one way in which it is flawless. The multiple possibilities of errors, and the irregularity of their occurrence, can make it difficult to specify the rules on how to solve problems algorithmically. In some cases, the fastest way may be to do some of the work manually.

As Dasu and Johnson point out, cleaning has a significant impact on results. Yet, detailed discussions on cleaning and error-checking processes are rare in introductions and chapters on methodology. Introductions usually describe digital tools, not manual or semi-manual tasks.[9] The role of digital tools and models is often discussed in terms of black boxes, with an input and an output and an obscure software in the middle. Such black boxes must be opened up in order to make research processes transparent.[10] Manual and semi-manual procedures can be said to represent another black box, however, perhaps even more opaque. They can be difficult to describe in a transparent way since they rely on human pattern recognition, a sensitivity to individual cases and the ability to make informed distinctions between information and noise.

A History of Manual Labour

As Markus Krajewski has pointed out in his media history of service, before digital servers there were human servants: human calculators, research assistants and secretaries.[11] The birth of automatised data processing did not do away with them. When Vannevar Bush speculated on the future research potentials of computers in 1945, he described a machine that 'will take instructions and data from a roomful of girls armed with simple keyboard punches, and will deliver sheets of computed results every few minutes'.[12] Father Robert Busa is often referred to as the first scholar to use the capabilities of computers within the humanities. However, his project also involved 'a roomful of girls'. His interest started in the 1940s when he studied the preposition 'in' in the works of Thomas Aquinas. This research would clearly benefit from the technologies developed to speed up data processing in business and administration. Busa partnered with IBM and during the following decades they constructed

an index of the full vocabulary in the works of Aquinas, published in 1974 as *Index Thomisticus*. In words that echo in recent publications on distant reading, Busa stated that: 'What had first appeared as merely intuition, can today be presented as an acquired fact: the punched card machines carry out all the material part of the work of (making a concordance).'[13]

The process was far from automatic though. The mainframe computers available at the time required 'a constant procession of human servants'.[14] In 1964, Busa had a team of 60 people assisting him with editing, programming and machine operations. Around 35 staff members were required for key-punching texts, verifying, listing, sight-checking and punch-card processing (the data was later transferred to magnetic tapes). In 1951, he estimated that it would take four years to complete the index. The reason why the project was not finished until the mid-1970s was mainly the laborious work of pre-editing and proofreading. 'Busa calculated that the thirty years of work he and others had spent on it amounted to roughly one million man hours.'[15] The foundational project of what would become digital humanities was truly a manifestation of the manual work needed to process data with computers.

The labour-intensive process did not discourage other scholars from using computers in their research (perhaps because those who introduced new methods seldom emphasised the importance of manual tasks). When the *Index Thomisticus* was completed in 1974, Busa was no longer alone. Linguists were among the early adopters, as well as some historians. Swedish historians were introduced to the idea that 'Clio faces automation' in an article by Carl Göran Andræ from 1966. Andræ explained that modern computers provided solutions to problems related to massive source materials. With data coded onto punch cards, or optical and machine-readable paper forms, it was possible to sort large amounts of data mechanically or electronically. In many cases, the systems were used as search engines, but they could also perform statistical calculations. The examples he gave included databases of coded newspaper articles, correlations between election results and census data, and the geographical distribution of unions and memberships in popular movements. Andræ concluded, as Busa before him, that: 'The mechanical work can now be left to computers.'[16]

Details on the actual research process are rare in publications by the first generation of computer-using Swedish historians.[17] Assistants, secretaries and machine operators may have been essential parts of the research process, but they were rarely acknowledged in the end results. Some clues can be found, however, and the impression they give is quite different from Andræ's optimistic view. The most laborious tasks concerned coding, in this case referring to the transfer of data from source documents to machine-readable formats (punch cards or optical markings on paper forms). A Swedish pioneer, the press historian Stig Hadenius, explained in 1967 that it took 'not more than 16 people' to extract the data needed for a pilot study on political news between 1896 and 1908.[18] A large project on Sweden during the Second World War had

a group of researchers investigating newspaper debates during the war. In order to render the newspaper material searchable, they coded 165,000 articles to create an index based on punch cards. The research team manually coded 138 variables for every article.[19] In the 1970s, a series of dissertations from Lund University used similar methods to process newspaper articles on various topics during the postwar era. Gunnel Rikardsson, who wrote about *The Middle East conflict in the Swedish press* (1978), did not elaborate on the manual tasks, but explained that six people had been involved in the process and that the 'coding work was experienced as exacting, mainly due to the high degree of concentration needed'. When the newspaper data was finally coded, however, the computer took over the workload: 'Manual processing had not been possible.'[20]

In his article from 1966, Andræ speculated on future research possibilities. Governmental agencies were already using computers to store and process data. Thus, for future historians who wanted to analyse the data, computer skills would be an absolute necessity. Most of the sources that historians worked with in the 1960s and 1970s were not 'born digital' though. The technologies (such as Optical Character Recognition, OCR) transferring analogue data to digital media showed promising results, but the majority of the research projects relied on manual labour. Millions of hours were spent on manual coding, punching and proofreading. The name of the research centre founded by Busa in the early 1960s was *Centro per L'Automazione dell'Analisi Letteraria*. Yet, and contrary to the automation emphasised in the name, photographs from the centre show what was often left unnoticed when research output was presented: rows of human operators, most of them young women.[21]

Struggling with Noisy Newspapers

The manual tasks needed today are of a different kind. The digitisation of sources is part of many research projects, but with scanners and software for OCR the digitisation of printed texts can be more or less automatised. Even handwritten texts can to some degree be digitised with the help of OCR technology. A significant difference, though, is that archives and libraries do much of this work for us. This is especially true for newspapers and books, parliamentary records and collections of audio-visual media, paintings and maps, and other museum artifacts. As long as the copyright allows for it, texts and images are made available online. In most cases, we do not need (and cannot afford) 35 assistants transferring data from one medium to another. Full-text search, topic modelling and tools for text analysis often make it unnecessary to code individual texts manually.

And yet, not all datasets are ready for processing out of the box; many of them can be very messy. As Carl Lagoze has pointed out, traditional archives and libraries used to guarantee the integrity of their records, at least in principle.

Control and curation were meant to facilitate the provenance and stability of data. The digitisation of collections and archival records has meant a fracturing of this control zone.[22] When millions and millions of pages are transferred (or translated) into digital formats, no one can guarantee the integrity of the data anymore. For large datasets of non-canonical texts in particular, libraries have spent less resources on curation, leaving researchers with much of the cleaning and preparation. Newspaper databases are notorious in this respect. Frequent OCR errors are well known, problems related to text segmentation less so, but both kinds of errors make it difficult to process the texts without manual interventions.

In one of my projects, I wanted to analyse discursive patterns in newspaper reports about the electrical telegraph in mid-19th-century Sweden.[23] From the National Library of Sweden, I was able to download a complete dataset covering one major newspaper from 1830 to 1862, about 10,000 pages. A systems developer helped me to penetrate the data (the first person who was asked refused to work with a dataset this noisy). Our first goal was to find every article containing the words 'electrical' and 'telegraph' ('elektrisk' and 'telegraf' in Swedish). Since we expected a high frequency of OCR errors, we used a Levenshtein distance to identify corrupted versions of our keywords, allowing three characters to be added, replaced or missing. In this way, we got 489 different hits for 'electrical' and 4,017 for telegraph. This was all done with a few simple commands, and the result came quickly.[24]

Not all of the hits had anything to do with the electrical telegraph though, and in order to filter out the false positives I had to go through the lists manually. That 'dialektisk' and 'apoplektisk' referred to something else was easy to figure out, but what about 'pelektriska' and 'elepris'? What about 'tograf', 'tfiesraf' and 'ttlefrnf'? Such combinations of characters can only be interpreted in the context of their appearances in the newspaper. In order to single out the proper keywords, I had to search the database and read the texts. It turned out that many of the incomprehensible words generated by the OCR actually referred to the electrical telegraph. My corpus would have been much smaller if I had not spent some time on this semi-manual step.

With an edited list of keywords, it was possible to locate every 'textblock' in the XML-files where 'electrical' and 'telegraph' co-occurred. A textblock is a unit of text identified as a coherent text by the text segmentation tool used in the digitisation process. However, nineteenth-century newspapers are difficult to process for the tool. The small print, the lack of headlines and the packed columns give few graphical clues on where one text finishes and another one starts. Human eyes can see it quite easily, while digital tools make several mistakes. Many libraries send the auto-segmented newspaper pages to private firms with outsourced divisions in Eastern Europe, Cambodia and India. The job of the staff is to correct the segmentation where it has failed.[25] The National Library of Sweden have skipped this crucial step in the process, however. I had to do the job myself.

We soon discovered that the textblocks generated by the tool had little to do with the texts as they were printed in the paper. Short news items from the same column were regularly merged into one single textblock, and longer texts chopped up into shorter pieces. The only way to single out the texts I wanted was to read through the whole corpus of identified textblocks and delete the unrelated parts. I also deleted text lines and combinations too difficult to decipher, such as 'lPlApfos2kOS2viKfSbmNAl' and 'rilet4R12bin1dPRRmo-8botoFrfutmfsOMMFgpgFvf'. I did not read the texts as carefully as I would have done if close reading was my main research method. But still, I had to read them.

With a somewhat clean dataset we could finally start to explore what the texts had to say about the electrical telegraph. We used a fairly simple and transparent method to identify semantic patterns. We looked at words co-occurring in a sliding window, and used the network analysis tool Gephi to find clusters of frequently co-occurring words. We still had some problems with noise though. Our method identified co-occurring words no matter the quality of the OCR, but for the final analysis we wanted to merge corrupted versions with the uncorrupted (for example, 'oeanen' and 'oceanen' (the ocean), 'Mo«se' and 'Morse'). Once again, we used a Levenshtein distance to pick out the most likely candidates to be merged, but I went through the lists to confirm the results manually.

In the end, we came up with some new and fascinating results. Many of the ideas we frequently associate with the electrical telegraph were more or less absent in the newspaper reports. Very few mentioned anything about the utopian potential of the new medium, it was not seen as an immaterial way of communicating and the idea that it 'freed communication from the constraints of geography' must be contextualised.[26] A bureaucratic discourse on regulation was much more prominent than a utopian on liberation, many of the articles described the material components of the new network instead of immaterial flows of electrical signals, and the geographical prerequisites (such as ocean floors and mountains) that determined where cables could be laid out were described in detail. I recognised much of this already when I read the texts in order to delete the noise, but I believe the quantitative analysis made the conclusions more convincing.

Scholars writing about computational text analysis usually emphasise the need to combine distant and close reading.[27] You need to switch between different perspectives to get an understanding of general patterns, as well as individual cases. In my own research, I already had to read the texts more or less closely in order to clean and prepare the corpus. When I reviewed the lists and graphs of frequently co-occurring words, I had an in-depth knowledge about the dataset on which they were based, making it easier to interpret the output. The time I spent reading and editing turned out to be well invested, but the process was very different from what I had imagined when I started the project.

Recognising Named Entities

What media technology we consider to be the first one ever invented depends on our definition of media. One common definition emphasises that a medium is a technology for the storage and/or transfer of information.[28] In that case, the tally stick might be the oldest media technology in human history. A tally stick keeps track of things you want to count (days, people, objects, etc.) and makes it possible to save the counts for later and to transport them from one place to another. The oldest one found, a bone from a baboon with carved markings, is at least 40,000 years old. 'Although our ancestors could not have known it, their invention of the notched stick has turned out to be amongst the most permanent of human discoveries.'[29] That my colleague Erik is using their invention to keep track of an imprecise digital tool in 2018 would definitely be beyond their imagination. Erik counts on paper though, not a bone from a baboon.

Tools for NER make it possible to identify and extract names of persons (even mythological creatures), organisations and places in digitised texts, as well as expressions of time (1857, 'next week'), monetary values and so on. The extracted data can be used for geographical visualisations, for network analysis, in timelines and as building blocks in other kinds of text analysis. HFST-SweNER, a language-processing technology developed to extract named entities from Swedish texts, is based on a dictionary as well as rules for identifying entities not in the dictionary, but likely candidates based on their contexts.[30] Tests have shown that it works fairly well for a curated corpus of texts from the 1990s, but will it work for 19th-century newspaper texts?

Erik Edoff is a media historian interested in geography. In one of his projects, he tries to figure out how new communication technologies in the 19th century reorganised the notion of space.[31] Was the world getting smaller when telegraphs, railroads, canals and steamships made it possible to communicate across space in a shorter time or in no time at all? Did far-away places come closer as a result of a time-space compression? One way to examine this (but certainly not the only one) is to identify and count place names in newspapers before and after the introduction of the new technologies (Erik selected papers from 1850 and 1890). Were names of distant locations printed more frequently when news travelled faster? The first results generated with NER indicated that places in the local region were in fact getting relatively more attention when new connections made communication faster, compared to places outside of the region. These were exciting results, since they seemed to show that the impact of the new technologies was different from what is usually believed. The question then became whether these numbers could be trusted. Did the tool find all the place names printed in the papers? If not, was it biased towards local Swedish place names?

In order to calculate the precision and recall, Erik chose a few newspaper issues for every title and year in the corpus. He read through the NER-tagged text files manually, and kept track of valid hits and false negatives in two

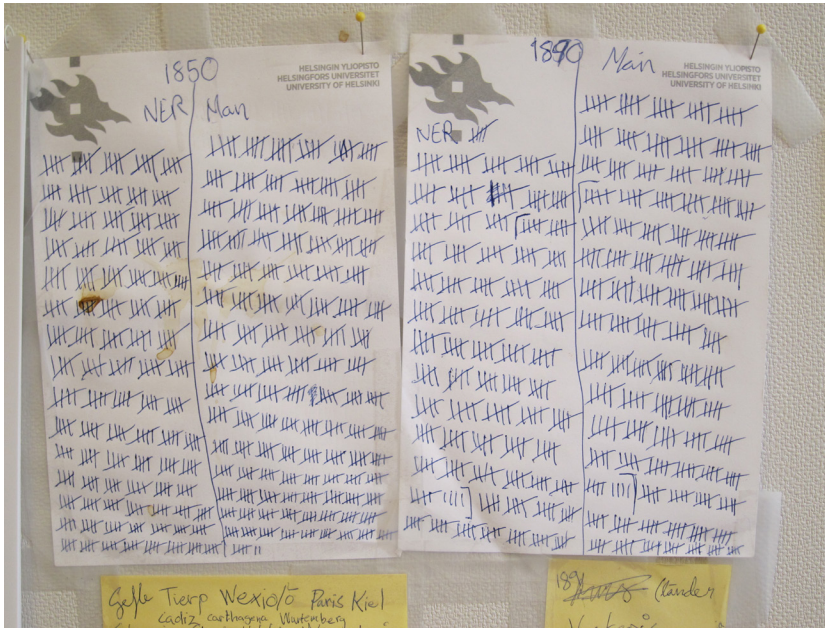

Figure 7.1: How many named entities (locations) did NER find, and how many more did Erik find? New locations not tagged by the tool were recorded on post-it notes. Source: Author.

columns on a couple of paper sheets. The method of counting was basically the same as the one used by our distant ancestors making notches on a bone: one mark for every word counted (see Figure 7.1). The brackets enclosing some of the counts separate place names mentioned in advertisements and lists, such as weather reports, stock market prices, etc. Those entities were more difficult to identify for the digital tool. There are other and perhaps more sophisticated ways to count occurrences of place names. But pen and paper are often efficient tools for minor tasks. No downloading or installing is required, and no special training. The interface makes the paper easy to use, and it is highly flexible.

The manual control revealed that the tool had left several place names untagged. For some reason, it did not recognise locations such as Paris, Kiel or Swinemünde, nor the Swedish towns Gävle (in the 19th century: Gefle), Växjö (Wexiö) and many minor towns and villages. One explanation might be the old spelling, but in some cases (when the spelling changed between 1850 and 1890), the tool recognised the old spelling, but missed the new. And the spelling does not explain the case of Paris. One geo-administrative category was left untagged almost completely: the parish. Today, it is hardly used outside of the Swedish church, but in the 19th century it was one of the most common ways in which Swedish locations were identified. Apart from these place names, Erik

found several locations untagged because of OCR errors. All of the entities identified manually were fed back into the system in order to make the final hit list more complete.

It turned out that the trend indicated by the first results was even more prominent once the false negatives were included. The relative frequency of places in far-away countries did not increase with the introduction of new communication technologies. Rather, locations close to the towns where the newspapers were published got more attention in 1890 compared to 1850. Erik's close reading of the sample issues provided him with some possible explanations. New places were put on the map thanks to new communication technologies: railway intersections, telegraph stations, bridges where steamships picked up passengers and goods, locks connecting canals and lakes. The places most frequently mentioned were those in the region, such as neighbouring towns and villages connected by railway, harbours close to home and regional centres nearby where telegrams were sent. New communications brought neighbours together. What was already close came even closer, while distant places were as far away as they were before. The repetitive task of recording place names on paper paid off in an interesting and convincing analysis. NER was a helpful tool, but it needed human assistance.

Troubleshooting Black Boxes

Digital models and tools will continue to improve. In the future there will, hopefully, be no need to carry out many of the manual tasks described in this text. OCR is getting more accurate every year; for some languages, NER seems to work fine already. On the other hand, as digital research practices are becoming more widespread, researchers will try to use the methods for new kinds of materials and in new areas—even areas where they will not run as smoothly. If we limited our research to clean datasets, very little would be accomplished. Many of the manual tasks carried out by research assistants and undergraduates in the 1960s are automatised today. New tools can achieve things unthinkable 50 years ago, but not always without human interventions. New problems seem to arise as old ones are taken care of.

The long history of information management can be seen as a series of new solutions generating old problems. In a fascinating article about the paper technologies used by Carl Linneus, in his big data-project on the natural system, Staffan Müller-Wille and Isabelle Charmantier note a 'curious dynamic' in the attempts to master information overload. 'The many technologies that were designed to contain information actually fuelled its further production, partly by providing platforms for more efficient data accumulation, partly by bringing to the fore new structural relations and patterns within the material collected.'[32] The result of technologies, developed to create order, overview and searchability, is often a new information overload. The digital media of today have other

capabilities than Linneus' paper slips and lists, but their operations are not as precise and clean as we might think. Rotten data, spam and noise thrives in a digital habitat (an interesting research topic in itself).[33] As shown by libraries' digitisation efforts, new technologies are far from perfect and human assistance is sometimes needed to keep them on track.

To edit, clean and validate large datasets manually or semi-manually may seem highly ineffective. In many cases, however, these procedures can be quite effective. Reading, counting, deleting and merging texts and other kinds of data in a manual or semi-manual fashion is a way to bridge distant and close reading. Insights from such encounters with data can be fruitful in the final analysis. It might also be a way to dig deeper into the inner workings of the digital tool on which the researcher is relying, to figure out how a specific dataset was processed and why the output turned out as it did. Troubleshooting is a good way to start if we want to examine what is inside the black boxes.

Notes

[1] The research presented here is part of the project 'Digital Models: Techno-historical collections, digital humanities & narratives of industrialisation', funded by the Royal Swedish Academy of Letters, History and Antiquities.
[2] Star 2002: 109.
[3] On the role of marginal (and yet central) figures, actions and technologies in the history of science, see Becker & Clark 2001 and Krajewski 2018.
[4] Thylstrup 2018: 42–43. See also Price & Thurschwell 2005.
[5] Fyfe 2016.
[6] Dasu & Johnson 2003: ix.
[7] Parikka 2012: 111.
[8] Wickham 2014: 2.
[9] See, e.g., Jockers 2013; Graham, Milligan & Weingart 2016; Rockwell & Sinclair 2016.
[10] Rieder & Röhle 2012.
[11] Krajewski 2018.
[12] Bush 1945: 104.
[13] Robert Busa quoted in Burton 1981: 1.
[14] Krajewski 2018: 308.
[15] Burton 1981: 3.
[16] Andræ 1966: 96.
[17] Jarlbrink 2015.
[18] Hadenius 1968: 68.
[19] The coding manual is now available online. See Åmark 2013.
[20] Rikardsson 1978: 59–60.
[21] Jones 2018.
[22] Lagoze 2014.

[23] Jarlbrink 2018.
[24] The newspaper noise is further explored in Jarlbrink & Snickars 2017.
[25] Fyfe 2016: 565.
[26] Carey 2008: 157.
[27] Jockers 2013: 26; Blevins 2014: 126; Hitchcock & Turkel 2016: 953.
[28] Mitchell 2017.
[29] Ifrah 2000: 64.
[30] Kokkinakis et al. 2014.
[31] Edoff, forthcoming.
[32] Müller-Wille & Charmantier 2012: 4.
[33] See Parikka & Sampson 2009; Eriksson 2016.

References

Andræ, C. G. (1966). Clio inför automationen. *Historisk Tidskrift, 86*(1), 47–79.

Becker, P., & **Clark, W.** (Eds.) (2001). *Little tools of knowledge: historical essays on academic and bureaucratic practices*. Ann Arbor, MI: University of Michigan Press.

Blevins, C. (2014). Space, nation, and the triumph of region: a view of the world from Houston. *Journal of American History, 101*(1), 122–147.

Burton, D. M. (1981). Automated concordances and word indexes: the fifties. *Computers and the Humanities, 15*(1), 1–14.

Bush, V. (1945). As we may think. *The Atlantic Monthly*, July, 101–108.

Carey, J. W. (2008). *Communication as culture: essays on media and society*. London and New York, NY: Routledge.

Dasu, T., & **Johnson, T.** (2003). *Exploratory data mining and data cleaning*. Hoboken: John Wiley.

Edoff, E. (forthcoming). Revolutions in communication? Digital methods and nineteenth century Swedish press.

Eriksson, M. (2016). Close reading big data: the echo nest and the production of (rotten) music metadata. *First Monday, 21*(7). DOI: https://doi.org/10.5210/fm.v21i7.6303

Fyfe, P. (2016). An archaeology of Victorian newspapers. *Victorian Periodicals Review, 49*(4), 546–577.

Graham, S., Milligan, I., & **Weingart, S.** (2016). *Exploring big historical data: the historian's macroscope*. London: Imperial College Press.

Hadenius, S. (1968). En kvantitativ innehållsanalys av dagspressen: teknik och användning i modern historisk forskning. In *Opinion och opinions bildning som historiska forskningsobjekt: Föredrag vid Nordiska fackkonferensen för historisk metodlära på Hässelby slott 4–6 maj 1967*. Oslo: Universitetsforlag.

Hitchcock, T., & **Turkel, W. J.** (2016). The *Old Bailey proceedings, 1674–1913*: text mining for evidence of court behavior. *Law and History Review, 34*(4), 929–955.

Ifrah, G. (2000). *The universal history of numbers: from prehistory to the invention of the computer*. New York, NY: John Wiley.

Jarlbrink, J. (2015). Historievetenskapens mediehantering. In M. Hyvönen, P. Snickars & P. Vesterlund (Eds.), *Massmedieproblem: mediestudiets formering*. Lund: Lunds universitet.

Jarlbrink, J. (2018). Telegrafen från distans: ett digitalt metodexperiment. *Scandia, 84*(1), 9–35.

Jarlbrink, J., & **Snickars, P.** (2017). Cultural heritage as digital noise: nineteenth century newspapers in the digital archive. *Journal of Documentation, 77*(6), 1228–1243.

Jockers, M. (2013). *Macroanalysis: digital methods & literary history*. Urbana, Chicago and Springfield, IL: University of Illinois Press.

Jones, S. (2018). Reverse engineering the first humanities computing center. *Digital Humanities Quarterly, 12*(2).

Kokkinakis, D., Niemi, J., Hardwick, S., Lindén, K., & **Borin, L.** (2014). HFST-SweNER: a new NER resource for Swedish. In N. Calzolari, K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, S. Piperidis (Eds.), *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. No. 391. Reykjavik, Iceland: European Language Resources Association.

Krajewski, M. (2018). *The server: a media history from the present to the Baroque*. New Haven, CT: Yale University Press.

Lagoze, C. (2014). Big data, data integrity, and the fracturing of the control zone. *Big Data & Society, 1*(2), 1–11. DOI: https://doi.org/10.1177/2053951714558281

Mitchell, W. J. T. (2017). Counting media: some rules of thumb. *Media Theory, 1*(1), 12–16.

Müller-Wille, S.. & **Charmantier, I.** (2012). Natural history and information overload: the case of Linnaeus. *Studies in History and Philosophy of Biological and Biomedical Sciences, 43*(1), 4–15. DOI: https://doi.org/10.1016/j.shpsc.2011.10.021

Parikka, J. (2012). *What is media archaeology?* Cambridge and Malden, MA: Polity Press.

Parikka, J., & **Sampson, T. D.** (Eds.) (2009). *The spam book: on viruses, porn, and other anomalies from the dark side of digital culture*. Cresskill, NJ: Hampton Press.

Price, L., & **Thurschwell, P.** (2005). Invisible hands. In L. Price & P. Thurschwell (Eds.), *Literary secretaries/secretarial culture*. Aldershot and Burlington, VT: Routledge.

Rieder, B., & **Röhle, T.** (2012). Digital methods: five challenges. In D. M. Berry (Ed.), *Understanding digital humanities*. Houndmills: Palgrave Macmillan.

Rikardsson, G. (1978). *The Middle East conflict in the Swedish press: a content analysis of editorials in three daily newspapers 1948–1973*. Stockholm: Esselte studium.

Rockwell, G., & **Sinclair, S.** (2016). *Hermeneutica: computer-assisted interpretation in the humanities*. London and Cambridge, MA: MIT Press.

Star, S. L. (2002). Infrastructure and ethnographic practice: working on the fringes. *Scandinavian Journal of Information Systems, 14*(2), 107–122.

Thylstrup, N. B. (2018). *The politics of mass digitization*. London and Cambridge, MA: MIT Press.

Wickham, H. (2014). Tidy data. *Journal of Statistical Software, 59*(10), 1–23. DOI: https://doi.org/10.18637/jss.v059.i10

Åmark, K. (2013). *Sverige under andra världskriget: pressregister 1938–1945*. Stockholms universitet, Historiska institutionen: Svensk nationell data tjänst (SND). Retrieved from snd.gu.se/catalogue/file/3386