Digital History: towards new methodologies

Serge ter Braake[1], Antske Fokkens[2], Niels Ockeloen[3], and Chantal van Son[2]

[1] Media Studies, University of Amsterdam, Netherlands sergeterbraake@gmail.com
[2] Computational Linguistics, VU University Amsterdam, Netherlands
antske.fokkens@vu.nl, c.m.van.son@vu.nl
[3] Computer Science, Web and Media, VU University Amsterdam, Netherlands
niels.ockeloen@vu.nl

Abstract. The field of Digital Humanities is changing the way historians do their research. Historians use tools to query larger data sets and they apply a different methodology to tackle certain research questions. In this paper we will discuss two propositions on the necessity of adapting to and taking advantage of the technological changes: 1) Digital Humanities tools are not the enemy of the historian, but they need to be used in a proper way. This requires historians to make 'tool criticism' part of their methodological toolkit; 2) Digital Humanities tools allow for a more data-driven and bottom-up approach to historical research. This eliminates some of the historian's preconceptions that are inevitably part of more traditional historical research.

Keywords: Digital History · Tool Criticism · Data Driven Research

1 Introduction

In 1984 the famous Dutch mathematician Edsger Dijkstra[4] wrote his closing speech for a summer school in computer science. He notes that new sciences, like computer science, often are confronted with high expectations, especially if they are not understood very well. 'We all know, how computing is now expected to cure all ills of the world and more, and how, as far as these expectations are concerned, even the sky is no longer accepted as the limit.'[5] More than thirty years later, Dijkstra's observation seems to fit the excitement surrounding Digital Humanities well. The field is still being defined, often misunderstood and while some think it will solve all humanities problems, it is also met with critique and skepticism [9] [18] [13, p. 82] [1] [10]. Even though some humanists would argue that the computer is just a tool [13, p. 74], forerunners of the field have always stated that the use of computers is not just increasing the scale of the research, or making life easier for researchers, but it also entails a new way of doing research, a new methodology. The 'Founding Father' of digital humanities, father Robert Busa, stated in 1980 that 'the use of computers in the humanities has as its principal aim the enhancement of the quality, depth and extension of research

[4] https://en.wikipedia.org/wiki/Edsger_W._Dijkstra

[5] https://www.cs.utexas.edu/users/EWD/ewd08xx/EWD896.PDF

and not merely the lessening of human effort and time. It has remained relatively unexplored, however, how doing history has changed after the digital turn [20, p. 4] [13, p. 76], or why and how those "new methodologies" are beneficial to humanities research at all [1].'

Twelve days before the deadline of this paper Allington et al. wrote a quite hostile essay against digital humanities, which according to them 'was born from disdain and at times outright contempt, not just for humanities scholarship, but for the standards, procedures, and claims of leading literary scholars.' Part of the authors' grumbles seems to stem from their observation that a lot of funding goes to digital humanities research, that is 'promoting methodologies that, until the emergence of these funding sources, had little support within the fields themselves.' [1] We find this statement surprising because academic progress is by definition achieved through improving old or advocating new methodologies. Investing in something new and promising therefore seems to be a wise thing to do, leaving digital humanists with the task to prove that what they are doing indeed is promising.

As is the case with any new methodology, tool or technique, it should only be applied where and when this makes sense, and in an academically sound manner that respects and takes into account the long tradition of research that has already been done in that field. In this paper we will discuss what is needed to 1) conscientiously and successfully apply digital humanities technology for historical research and 2) apply a new, more data driven methodology to old research questions.

2 Tools and data

'Algorithms are inherently fascistic, because they give the comforting illusion of an alterity to human affairs.' [10]

This quote by Stephen Marche from his influential essay *Literature is not data* illustrates that humanities researchers are not used to work with algorithms and computational tools in their research. The current section deals with the sometimes complicated relationship between historians and digital humanities tools. Some historians classify these tools either as 'fascistic', or use them without really knowing how they work. Historians cannot be blamed for the latter, since algorithm reading usually is not part of their education (yet). Even some forerunners of the digital humanities field readily admit that they cannot code themselves [1]. Instead, and rightly so, history students are trained to criticize the sources they work with from the start of their curriculum. The academic value of historical research is doubtful without a sound source criticism addressing questions such as: What is the context in which this source was written? For what purpose? What can be said about the author and their possible beliefs and preconceptions? This holds for both primary sources (diaries, accounts, letters) and secondary sources (history works in which primary sources are analyzed and brought to a synthesis).

The use of digital humanities tools for historical research is complicated for a variety of reasons. Stephen Marche is right in saying that digital humanities tools generally treat texts badly. While historians use texts as the core of their studies, such tools often cut these texts up into 'data' and use complex algorithms to analyze them. This disintegration of text into data is necessary for a computer to be able to do its computations. The potential complexity of the computations often depends again on the format in which the data is stored. It is a valid question however, to ask what the transition from text to data means for humanities research. Marche worries, for example, about the decontextualization of the texts: 'The algorithmic analysis of novels and of newspaper articles is necessarily at the limit of reductivism. The process of turning literature into data removes distinction itself. It removes taste. It removes all the refinement from criticism. It removes the history of the reception of works.'

Besides this 'reductivism' and what it does to the source awareness of the historian, the use of software for historic research, as access portals to the sources, also introduces a new layer between the historian and his/her sources. In 2012 Rieder and Röhle stressed that digital tools rely on assumptions made by humans and should not be treated as being objective. To grasp the finer nuances of a tool however, one may need a bachelor's degree in Computer Science [16, p. 76]. It is not realistic to expect that historians will reach this level any time soon. What we propose instead is a close collaboration between 'tool literate' historians and computer scientists in developing tools for research, which are documented in such a way that historians can understand the basics of how they work and what happens in the process between query and result. It is for example easy to count words with tools like the Google Ngram viewer, but the results should always be interpreted with the way the tool was built in mind, including the estimated reliability, the sources it uses and how it selects elements within these sources [6]. In other words: historians should get proper training in 'tool criticism', which will allow them to make use of the newest technology in an academically sound manner and to connect traditional source criticism and other historical best practices better.[6] Of course the same applies to academics in other humanities fields, such as media studies [12].

Tool criticism especially comes into play when the computer has to make choices that are difficult to interpret automatically. It is important to realize that even the most logical choices could lead to wrong results. For example: geographical locations in text can be identified using resources such as GeoNames.[7] If we would want to know the place of death of a group of people from the Netherlands, then it would be a logical assumption to interpret an ambiguous place name as the one in or nearest to the Netherlands.[8] This could, however,

[6] In May 2015 a small workshop on tool criticism was organized in Amsterdam

[7] http://www.geonames.org/. Though of course we have to be cautious here, since it is difficult to 'translate' historical place names to modern names, and problematic because the geographical region covered by a town will be smaller or bigger throughout history.

[8] A similar example can be found in Ockeloen et al. [14]

lead to a distortion of the results if we deal with historic place names like Batavia (present day Jakarta). If the software prefers to assign locations that are in or nearest to the Netherlands, it will point to Passau in Germany, or to places in the United States or South America, rather than to the town in Indonesia we are looking for. Automatic disambiguation of geographical locations in text can thus be useful, but the researcher needs to take into account the context in which it is applied.

Another example is related to the sophistication of technology. The OCR quality of digitized texts can lead to misleading results. A part of the Google Books corpus can be queried with an N-gram viewer, which provides the relative frequency of a word over time. Any possible imprecision usually is not detrimental, because of the sheer size of the corpus [11]. In the past, however, the letter 's' in some of the texts with a so-called 'gothic' font would be misread as an 'f'. If a researcher looked, for example, for *Amsterdam* in the eighteenth century, he/she would get almost no hits. This is not because the town is never mentioned, but because the OCR had transcribed the spelling of the town as 'Amfterdam'. This inaccuracy has been fixed, but it leaves the researcher wondering what else needs fixing and how one can estimate the 'uncertainty' of tools such as the N-gram viewer [19].

It should go without saying that the potential usefulness of a tool relies for a big part on the data it runs on. Questions which need to be addressed are: How representative is this corpus to answer my research question? What sources does the tool not use that could (potentially) also provide an answer to my questions? Why was this data used in the first place? These issues are important for traditional research as well, but are especially relevant considering that only a very small part of the archives are digitized [5]. It is furthermore important to realize that even if a source is digitized it does not mean it is automatically available for advanced digital text analysis. There is a whole spectrum of forms - e.g. digital photos, OCR-ed text, XML files - in which text can be digitized, which allows digital humanities tools to do a finer or rougher degree of analysis. Hence, historians should be aware of what we might call, the 'granularity' of the data. Certain data sets can be created for a particular purpose with its 'granularity' fit accordingly. Re-using that data set for a different purpose that requires another level of 'granularity' is rarely without complications.

Historians still have to get used to the transition from text to data and the decontextualization that comes with this process. Decontextualization can be partly remedied by a detailed provenance plan and by providing access to the original texts [14]. This way tools facilitate traditional source criticism. The question remains if this solution also works in practice and, if not, what else could be done to avoid historians drifting away from their source material [20, p. 27]. Historians get access to a wealth of digitized sources, but not in the direct and contextualized way they are used to. Computer scientists (and the historians working with them) need to document their work in such a way that all choices are accounted for. In order to work with digital tools in an academic environment and apply good 'tool criticism', historians need to be able to judge:

- What layers or barriers a tool introduces between them and their sources, and whether these barriers can be overcome;
- Whether results produced using the tool are verifiable, either within the tool itself or externally using the provided data;
- Whether any assumptions were made in the tool's logical decision making of which understanding must be gained to interpret results;
- What possible mistakes could be made by a lack of sufficiently sophisticated technology;
- What other data could have been used to answer this question that was not used by this tool;
- What the 'granularity' of the data used by the tool is.

Knowing this is one thing, but making sure that researchers and students gain the skills to do all this is another. Humanist scholars already need to have knowledge on a wide variety of fields. Historians, for example, often need to be able to read many (old) languages, need to know know economical, sociological or psychological principles, or need to get acquainted with law, philosophy, statistics and theology. Historians are thus able and used to gain knowledge of other fields, but the question remains if they can and are willing to spend their time learning something new when they already have to know so many things. To persuade them to get some basic tool understanding should therefore start by making them see the benefits of digital humanities research.

It would be more efficient to ensure that history students already have a basic grasp of tool criticism when they leave their universities with a diploma. From our own experience at two Amsterdam universities and from contacts with other European institutions we can tell, however, that humanities students do not eagerly flock to 'Digital Humanities' courses. A recent blog post from Ryan Cordell [4], in which he also makes some interesting suggestions about how (not) to teach Digital Humanities, depicts a similar situation in Wisconsin: humanities students do not choose history or literature to learn how to program. The non-technical part of Digital Humanities often is too reflective, too much on a meta-level, for students who are still learning the finer nuances of their field. Instead, Cordell tries to integrate tools and technology into his courses on 'traditional' humanities topics. What we have defined here as 'tool criticism' and the six points which should be judged are quite similar to what is required for proper source criticism. Tool criticism could therefore be relatively easily integrated in the present humanities curricula as a self evident and necessary extension of the traditional source criticism.

3 Top-Down and Bottom-Up Approaches

The previous section dealt with the sometimes complicated relation between historians and digital tools. In this section we will discuss how digital humanities technology fundamentally changes the way historians approach, or should approach, topics in their field. Without claiming this is the only, or even most important methodological change, we will argue that digital humanities enables historical research to become more data-driven and bottom-up. By taking the available data itself as the starting point for analysis rather than preexisting models and interpretations, some of the historian's preconceptions can be eliminated. We will provide two examples close to our own research of traditional research questions tackled by digital tools. These examples are by no means exhaustive, but will illustrate our point. First we will deal with the theme of canonization of people in history and second with studying of shifting concepts through time.

3.1 Canonization of People

Who becomes famous in history and why are interesting questions for historians [17]. Traditional historical research would likely start with looking at the currently famous historical people and trace their fame through history by going back in time. This is a labor intensive, but usually fruitful way of approaching such a topic. Many female heroines, whose fame had been debunked in the nineteenth century of chauvinistic academic history writing, are being restored this way [8]). Until recently however, it was nearly impossible to trace fame for larger groups of people. The Google Books team tried to quantify fame between 1800-2000 by searching for names in the corpus of Google Books [11]. One of the authors of that paper, Adrian Veres, also created a science Hall of Fame, in which he ranks the scientists in milli-darwins, based on the number of mentions in the Google Books texts.[9] Ironically, these computer scientists still used a classic historical approach to study fame. They used existing lists of famous people and mapped their fame through time, running the risk of missing out on people that were famous in their own time but did not make it to the modern canon (and also missing possible instances of 'Charlef Darwin').

Digital Humanities tools make it possible to approach the topic differently. In recent research, we proposed a methodology to get around this top-down bias by extracting *any* names from digitized text and disambiguating them. We tracked the names by searching for strings of capitalized words (like 'Johan Cruijff', but also 'Johan de Witt' and 'Joan Derk van der Capellen tot den Poll'). Even though this results in lists of names with a lot of noise, it is easy to manually filter out strings that do not refer to people (like 'Den Haag'). We disambiguated the names mainly by using time stamps to separate people with the same name. When running tests on the data of the Biography Portal of the Netherlands, we traced a preference for the house of Orange, foreign rulers and Jesus Christ as being the most frequently mentioned individuals in other people's biographies [2]. The availability of more machine-readable texts and tools to analyze them therefore makes a completely data-driven and bottom-up approach to the study of canonization in history possible. By approaching the topic bottom-up the narrow selection of predefined canons can be avoided, for as far as the digitized material allows this.

[9] http://www.sciencemag.org/site/feature/misc/webfeat/gonzoscientist/episode14/index.xhtml

3.2 Tracing Concepts through Time

'Concepts' and how they change over time have a decade-long history of being studied by historians. A concept is a notion or an idea, referred to by one or several words, and which has certain attributes that can change over time. Especially German historians spent a lot of time and resources to get a grasp on concepts. In the *Geistliche Grundbegriffe* project a German team of scholars studied state formation and the course of history in general between 1750-1850 for decades, by seeing how certain concepts, or words related to concepts, changed meaning [15]. In sociology some speak of contested and contestable concepts, like 'democracy' and 'freedom', on which people never seem to agree [3]. Contested concepts are considered to be of great interest to study in order to see how the political climate changes in a certain period of time.

One thing that these studies have in common is that the researchers determine in advance what concepts are worth investigating, and that they look into how these concepts fit into an already conceived model. Of course historians have already read a significant number of texts on which they base their ideas and hypotheses, before studying anything more closely. Still, their approach is mainly top-down. Once a certain topic is singled out, it usually has to be brought to completion, simply because it is too labor intensive to easily abandon. Similar to the research on the canonization of people in history, one runs the risk of interesting concepts not being found because of this necessarily subjective approach to the selection of a topic.

Fortunately, concepts are a fruitful topic for digital humanities research, as was showcased recently on a workshop in Helsinki.[10] When the necessary digital material is available, a few simple, exploratory exercises could already help determining the most sensible direction to answer a research question. The concept of 'nation' for example was still very much in flux in the nineteenth century, and possibly referred to by many different words. It is difficult to determine in advance then, what terms are worth investigating. For our research we have text files available of all published volumes of the Dutch liberal and intellectual monthly journal *De Gids*, in which nation ('natie') is a frequent topic. We can search for terms like 'natie', read the key texts and determine what related terms are worth investigating. An alternative and possibly a more neutral approach is to count all words in a certain period of time and look closely at what is discussed and how this is related. In a corpus of 396,963 words from the first year of *De Gids* in 1837, we find only 20 instances of the word 'natie', but 114 of 'vaderland' ('fatherland') and 106 of 'volk' ('people'). Tracing the shifts in those terms and similar words will be a good starting point for further investigations.

One of the most neutral more sophisticated approaches is that of Tom Kenter et al. [7], who trace concepts through time by starting with one or two 'seed words' and use vector coordinates to see what terms are associated with it over time. Words that occur frequently in the same sentence together get vector coordinates close to each other and are likely to be connected somehow. In our

[10] http://www.helsinki.fi/collegium/events/conceptual_change/index.html

case above regarding the concept of ‘nation’, it would be interesting to determine if frequently occurring words like king (‘koning), foreigner (‘vreemdeling’), language (‘taal’), freedom (‘vrijheid’) and religion (‘godsdienst’) are closely related or if they mostly occur in different contexts. The chosen seed words will still be predefined, but they can be based on the findings of simple preparatory text analysis. This methodology also allows for a more flexible way to approach research on concepts and may result in leads that would otherwise remain hidden. By tracing the concepts underlying these seed words for a longer period of time, it is eventually even possible that the original seed words will disappear. To use the example of Kenter et al.: the concept of portable music player may be behind the seed word I-Pod, but of course ‘I-Pod’ will not be present in texts from the twentieth century.

4 Conclusions

Historians and other humanists can be skeptical about technology and ask what kind of true innovations digital humanities technology brings to the field of history. Part of this skepticism has to do with a too narrow focus on the tools and how they work and too little on the way they alter existing research methods. This paper provided some reflections on how the field of history can, or maybe should, adapt to the changes brought by the field of digital humanities and can progress thanks to it.

Tools will become increasingly more reliable and interesting to use. Historians should be aware of the barriers a tool erects between them and the sources, whether the results are verifiable, what sources could and could not be used, what biases the tool has built-in and what its technical limitations are. Such ‘tool criticism’ is necessary to see what tools do with historical sources, to determine what this means for the interpretation of the output, and to see how this can lead to innovations on how history should be studied. This criticism could also be an essential part of a criticism-improvement feedback loop between historians and tool creators.

In our view one of the most exciting new opportunities digital humanities tools bring is a rigid studying of history bottom-up. Computers can read data far quicker than humans can, which allows for a fast and more extensive exploration of the sources before determining who or what to study more in depth. Instead of applying a ready-made model to one’s sources, current technologies increase the possibility of letting the data ‘speak’ for itself, experiment and try to find the interesting leads in it, which may eventually lead to new models.

5 Acknowledgements

This work was supported by the Amsterdam Academic Alliance Data Science (AAA-DS) Program Award to the UvA and VU Universities and by the BiographyNet project http://www.biographynet.nl (Nr. 660.011.308), funded by the Netherlands eScience Center (http://esciencecenter.nl/).

References

1. Allington, D., Brouilette, S., Golumbia, D.: Neoliberal Tools (and Archives): A Political History of Digital Humanities. LA Review of Books (may 2016)
2. ter Braake, S., Fokkens, A.: How to Make it in History. Working Towards a Methodology of Canon Research with Digital Methods. Proceedings of the First Conference on Biographical Data in a Digital World 2015, Amsterdam, The Netherlands, April 9, 2015 pp. 85–93 (2015)
3. Collier, D., Hidalgo, F.D., Maciuceanu, A.O.: Essentially Contested Concepts. Journal of Political Ideologies 11(3), 211–226 (2006)
4. Cordell, R.: How Not to Teach Digital Humanities, http://ryancordell.org/teaching/how-not-to-teach-digital-humanities/
5. Jeurgens, C.: The Scent of the Digital Archive: Dilemmas with Archive Digitisation. BMGN - Low Countries Historical Review 128, 30–54 (2013), http://www.bmgn-lchr.nl/index.php/bmgn/article/view/URN:NBN:NL:UI:10-1-110021/9783
6. Kelly, M.: Visualizing Millions of Words. In: Gold, M.K. (ed.) Debates in the Digital Humanities, pp. 402–403. University of Minnesota Press, Minneapolis, London (2012)
7. Kenter, T., Wevers, M., Huijnen, P., De Rijke, M.: Ad Hoc Monitoring of Vocabulary Shifts over Time. Proceedings of the 24th ACM International on Conference on Information and Knowledge Management pp. 1191–2000 (2015)
8. Kloek, E.: Kenau en Magdalena. Van Tilt, Arnhem (2014)
9. Liu, A.: The state of the digital humanities. A report and a critique. Arts and Humanities in Higher Education II(1:2), 8–41 (2012)
10. Marche, S.: Literature is not Data: Against Digital Humanities. LA Review of Books
11. Michel, J., Shen, Y.K., Aiden, A.P., Veres, A., Gray, M.K., Brockman, W., The Google Books Team Pickett, J.P., Hoiberg, D., Clancy, D., Norvig, P., Orwant, J., Pinker, S., Nowak, M.A., Aiden., E.L.: Quantitative Analysis of Culture Using Millions of Digitized Books. Science 131, 176–182 (2011)
12. Noordegraaf, J.: Computational Research in Media Studies: Methodological Implications. KWALON 61(21, 1, Special issue: Qualitative research in the digital humanities) (2016)
13. Nyhan, J., Flinn, A., Welsh, A.: Oral History and the Hidden Histories project: towards histories of computing in the hunmanities. Digital Scholarship in the Humanities 20(1), 71–85 (2015)
14. Ockeloen, N., Fokkens, A., ter Braake, S., Vossen, P., De Boer, V., Schreiber, G., Legêne, S.: BiographyNet : Managing Provenance at multiple levels and from different perspectives. Linked Science pp. 59–71 (2013)
15. Richter, M.: The History of Political and Social Concepts. A critical introduction. Oxford Univerity Press, New York and Oxford (1995)
16. Rieder, B., Röhle, T.: Digital Methods: Five Challenges. In: Berry, D.M. (ed.) Understanding Digital Humanities, pp. 67–84. Palgrave, Macmillan, Basingstoke (2012)
17. Smyth, J., Penman, M.: Reputations and national identity, or, what do our heroes say about us? Études écossaises 10 (2015)
18. Svensson, P.: Envisioning the Digital Humanities. Digital Humanities Quarterly 6(1) (2012)
19. Traub, M.C., van Ossenbruggen, J.: Estimating the Impact of OCR Quality on Research Tasks in the Digital Humanities. (2015), http://www.slideshare.net/ingeangevaare/06-traub

20. Zaagsma, G.: On Digital History. BMGN - Low Countries Historical Review 128, 3–29 (2013), http://www.bmgn-lchr.nl/index.php/bmgn/article/view/URN%3ANBN%3ANL%3AUI%3A10-1-110020