Opportunities and Risks in Digital Humanities Research

Book Section

Mark Hall

Opportunities and Risks in Digital Humanities Research

1 Introduction

The term *Digital Humanities* (DH) has established itself as one of the main umbrella terms under which humanities research that incorporates digital aspects is organised. The various different terms and computational areas that had developed in the individual humanities disciplines have slowly been coalescing under this new term, bringing with them a wide variety of methods and data. This has created the necessary critical mass and exchange of ideas that has led to a rapid development of a wide range of new methods and tools for investigating humanities research questions. Over time these have become easier to use and the DH field has expanded outward to include researchers who are no longer directly interested in developing methods and tools, but who use the methods and tools in the pursuit of their own research questions.

While there has been much discussion about where the boundaries of DH lie,[1] in this article I will take a very broad definition, including any research that makes use of or plans to make use of digital tools or methods. The reason for this is that the use of any tool changes how we interact with the world, even if the tool is just a word-processor, but already much more so when it is a spreadsheet, and significantly more when algorithms are applied to data. Although it has been suggested that just working with digital materials is not sufficient to be counted into DH,[2] I believe that the true value of DH can only be achieved if the term is used in an extended definition that goes beyond those who are interested in developing the methods and algorithms to encompass

1 Matthew G. Kirschenbaum: What is Digital Humanities and What's It Doing in English Departments, in: Matthew K. Gold (ed.): Debates in the Digital humanities, Minneapolis 2012, pp. 3–11; Bethany Nowviskie: Digital Humanities in the Anthropocene, in: Digital Scholarship in the Humanities, 30.1 (2015), pp. i4–i15; John Unsworth: What is Humanities Computing and What is Not?, in: Melissa Terras, Julianne Nyhann, Edward Vanhoutte (eds.): Defining Digital Humanities, London/New York 2016, pp. 51–63.

2 Kathleen Fitzpatrick: The Humanities, Done Digitally, in: Gold: Debates, pp. 12–15.

ISBN Print: 9783847111771 – ISBN E-Lib: 9783737011778

data and tool users. For this wider group we need to make the case for why they should add digital aspects to their research, but more importantly they need to be aware of the major risks that adding computational methods to their repertoire of research methods present. An awareness of the pros and cons of the computational methodologies will enable them to correctly judge the impact of the digital tools they are employing on the outcomes that they observe and the conclusions they can draw.

At the same time, I will make some assumptions about the research context and particularly the kind of data used in the DH research. The assumption is that the data under investigation is of a historic nature, primarily text based, and from a time-period long enough ago, that a verification of any research results through other, independent sources is no longer possible.[3] The focus on text is due to the prevalence of text as the primary data type in DH projects and also in the wider humanities.[4] However, the main arguments about the advantages and risks of DH tools and algorithms apply to other data-types (images, sound, video, meta-data, etc.) equally and some of the examples will be drawn from that wider set to illustrate that. Similarly, while the critique applies also to current non-historic data or data where verification through other sources is possible, the degree to which it is relevant varies, in particular as for newer data it is often easier to mitigate the potential risks. However, where verification or triangulation is possible, it also has to be applied, otherwise the issues raised here are just as valid.

The aim of this article is both to encourage researchers to consider making use of the digital methods and tools that are out there, while at the same time reminding everybody of the risks that these introduce into their research. As a result the remainder of the article is structured as follows: first I will discuss the main opportunities provided by the DH methods and tools (if you already use these, you might skip this), then I will spend significant time analysing the risks posed by poor methodology and use of the tools, and finally I will discuss the potential dangers for DH as a field if the risks are not taken into account fully. In particular I will highlight the need for true collaboration between humanities and computer science researchers, which has the potential to deliver truly novel insights to both areas.

3 Patrik Svensson: Humanities Computing as Digital Humanities, in: Terras, Nyhann, Vanhoutte (eds.): Defining Digital Humanities, pp. 175–202.

4 Stefan Jänicke et al.: On Close and Distant Reading in Digital Humanities: A Survey and Future Challenges, in: Rita Borgo, Fabio Ganovelli, Ivan Viola: Eurographics Conference on Visualization 2015, pp. 83–103. DOI: https://doi.org/10.2312/eurovisstar.20151113 (last access: 07.02. 2020).

ISBN Print: 9783847111771 – ISBN E-Lib: 9783737011778

2 Opportunities in the Digital Humanities

DH offers a vast range of new tools and methods that could be used to augment non-computational research methods, which can be bewildering as it is often difficult to decide where to start. There have been various calls to humanities researchers to embrace digital tools and methods in their working routine, most of these have either been very high-level or alternatively contain a specific selection of tools and methods used to demonstrate the value of adding digital.[5] Here I will attempt to achieve a compromise between these two approaches. Identifying two major areas where adding digital methods have the potential to significantly alter the research process – data access and data volume – while illustrating the advantages provided by each area through a selection of tools or methods.

2.1 Data Access

The data sources for humanities research are primarily archives, libraries, and museums – with a few exceptions such as archaeology, which requires venturing out into the world. As a result non-computational[6] humanities research first requires identifying those physical archives that are of relevance to the research and then physically visiting them to discover the relevant objects. The interaction with the objects in the archive will generally be mediated through the archivist or librarian and rely on indexing tools such as card catalogues. Findings in the archive are noted and then at a later time, in the office the notes are analysed, and the research outputs generated. At this point in time if there is information missing in the notes, a repeat visit to the physical archive would be necessary. The physicality of this process and the time (and financial) costs of visiting the archive necessitate very careful planning and note-taking, but also place an intrinsic limit on the amount of material that can be sighted in the archive within the given time-span, meaning that careful sampling is the only viable way to get representative data.

5 Kathleen Fitzpatrick: The Humanities, Done digitally, in: The Chronicle of Higher Education. https://www.chronicle.com/article/The-Humanities-Done-Digitally/127382 (last access: 07.02.2020); Ted Underwood: Dear Humanists: Fear Not the Digital Revolution, in: The Chronicle of Higher Education 2019. https://www.chronicle.com/article/Dear-Humanists-Fear-Not-the/245987/ (last access: 07.02.2020).

6 One of the difficulties in comparing DH work with humanities research that does not use digital tools or methods is how to label that kind of research without judgement. For this reason I will refer to that kind of research as non-computational, to distinguish it from DH research that uses digital or computational methods.

ISBN Print: 9783847111771 – ISBN E-Lib: 9783737011778

The digitisation of archives' holdings over the last few decades and their availability via online portals represents a step-change in how archival research can be undertaken.[7] Depending on the type of materials involved, this might include photographs of the objects and associated meta-data or, in the case of text documents such as books or letters, full text generated through *Optical Character Recognition* (OCR) or manual transcription. Search engines are then used to index these data and the research can now enter one or more search keywords and retrieve all objects that share these keywords. The researcher can then download the results to their own computer for future analysis.

This represents a number of major changes to the humanities research flow. First, the requirement to physically visit the source archives is reduced, as much content is now available digitally. Second, it is now possible to refer to a digital representation of the actual source material during the analysis and write-up, rather than being restricted to the notes made at the archive. Third, if during the analysis it becomes clear that additional data are needed, it is now significantly easier to acquire this. At the same time relying purely on digital archives does introduce new sources of data uncertainty that will be discussed later.

2.1.1 Technological underpinnings

The technology underpinning digital archives is *Information Retrieval* (IR).[8] IR focuses on quickly finding things that satisfy the information need of a user and has four core concepts: the documents that can be searched for, the keywords used to index the documents, the search keywords provided by the user, and the relevance of each document with respect to the search keywords. Important to note is that the concept of a document is used to describe any kind of thing that is indexed by the IR system, regardless of whether this is full-text, images, or just meta-data.

To achieve high retrieval performance the data-set within which the user can search is first pre-processed and each document is indexed with one or more keywords that describe the document. After indexing the user can formulate their information need as one or more search keywords. The information retrieval system then uses the index keywords to retrieve those documents that best match the search keywords.

7 Kimberly Barata: Archives in the digital age, in: Journal of the Society of Archivists 25.1 (2004), pp. 63–70; Bob Nicholson: The Digital Turn: Exploring the methodological possibilities of digital newspaper archives, in: Media History 19.1 (2013), pp. 59–73.

8 Ricardo Baeza-Yates, Berthier Ribeiro-Neto: Modern Information Retrieval, 2nd ed., Harlow et. al. 2011.

ISBN Print: 9783847111771 – ISBN E-Lib: 9783737011778

In this process there are two main areas that are of relevance to the humanities researcher, as they influence what the researcher is shown for their search keywords. First, how the input data are pre-processed to generate the keywords describing each document and second, the way in which the 'relevance' of a document with respect to the search keywords is calculated.

In pre-processing the document is split into individual keywords using a language model in order to determine where the boundaries between keywords are. For example such a model knows that 'keyword-search' is a compound noun, which in the index should not be split into two. Because the search system uses exact matching[9] between search and document keywords, to ensure performance, the tokens are generally further processed in the index to increase the likelihood of the two matching.[10] Stemming or lemmatisation are common techniques, which reduce variants of the same word to a single variant. For example 'house', 'houses', and 'housing' would all be reduced to 'house'. The same is applied to the search keywords, with the effect that a search for 'house' also returns documents that refer to the plural or to the verb form. This increases the number of documents returned (also known as the recall), but also increases the number of documents that are found that are not relevant (what is known as a loss of precision). However, it is important to note that a further manual inspection of the results is necessary to ensure that the retrieved documents actually match the specific variant of the word that the researcher was looking for, as the results will contain all variants, regardless of how the user spelled the word in the query.

When the user then searches, the systems matches the query to document keywords and most current IR systems will also attempt to rank the matching documents by how relevant they are to the query keywords. To achieve this, each keyword is given a score for each document it appears in. Then, when searching the scores for each query keywords are combined per document and then the documents ranked by that score. One of the most successful scoring formulas is BM25,[11] which has been extended in the past to achieve the quality seen in modern search systems[12], but the fundamental principle as covered here remains the same. Documents are ranked by measuring how often a se-

9 There are of course also fuzzy approaches to matching query and document keywords, but those are deployed and used infrequently.

10 Safaa I. Hajeer, et al.: A New Stemming Algorithm for Efficient Information Retrieval Systems and Web Search Engines, in: Aboul E. Hassanien, et al. (eds.): Multimedia Forensics and Security. Foundations, Innovations, and Applications, Cham 2017, pp. 117–135.

11 Stephen E. Robertson et al.: Okapi at TREC-3, in: Proceedings of the Third Text REtrieval Conference, Gaithersburg 1995, pp. 109–126.

12 Sergio Jimenez et al.: BM25-CTF: Improving TF and IDF factors in BM25 by using collection term frequencies, in: Journal of Intelligent & Fuzzy Systems 34.1 (2018), pp. 1–13.

ISBN Print: 9783847111771 – ISBN E-Lib: 9783737011778

arch keyword appears in a single document and also how often it appears across all documents. A relevant keyword for a document is a keyword that appears frequently in that document and only in few other documents. It is relevant, because it distinguishes the document from all other documents. The list of documents that match the user's query is then sorted by these relevance values and the user is shown 1–10 of 10,000 documents, but again it is often necessary to page through all the results, as this technical 'relevance' might not match the semantic relevance desired by the researcher.

2.2 Volume of Data

The second major aspect in which DH research distinguishes itself from non-computational humanities research is the amount of data that can be used as part of the research.[13] Non-computational, manual analysis imposes a temporal upper bound as to how much time can be invested and thus how much data can be analysed. The use of digital tools does not alter the amount of time that can be invested in the manual analysis. However, the DH methods allow a much larger data-set to be automatically processed, the results of which can then be analysed manually.

Methods for dealing with large text data-sets are commonly described as 'distant reading'[14] and that term has slowly expanded to sometimes being used to label any kind of large-scale, quantitative analysis of digital humanities data – irrespective of whether that is an appropriate extension of the term.[15] Before briefly looking at some of these methods, it needs to be said that while the term distant reading is frequently used, it is misleading as none of these methods actually 'read' the data they are provided with. Instead all these techniques take the source data, apply a variety of processing steps which result in a series of quantitative data, and from these statistical measures of the data can be produced that can then be interpreted. This does not invalidate the methods, it is just important to not take the label to literally.

Techniques for dealing with large scale textual data tend to be based on word counts, either just counting individual words or counting co-occurrences. In both cases the text is initially split into words, using a language

13 Christoph Schöch: Big? Smart? Clean? Messy? Data in the Humanities, in: Journal of Digital Humanities 2.3 (2013), pp. 2–13.

14 Franco Moretti: Distant Reading, London 2013. Other terms such as 'zoom' / 'zooming' have also been used, but this is the one that has become most widespread.

15 Ted Underwood: A Genealogy of Distant Reading, in: DHQ: Digital Humanities Quarterly 11.2 (2017). http://www.digitalhumanities.org/dhq/vol/11/2/000317/000317.html (last access: 07.02.2020).

ISBN Print: 9783847111771 – ISBN E-Lib: 9783737011778

model as described earlier. Then for individual word counts, these are simply calculated. For the co-occurrence models, whether two words co-occur is then calculated by moving a window over the text.[16] Two words are then calculated to co-occur if they appear within the window, in other words, if they appear within a certain distance of each other in the text. While the exact size of the window – the number of words within which words are defined to co-occur – varies depending on the context, ten words is commonly used as a window size. The individual counts and co-occurrence values form the basis for three commonly used methods: topic modelling, vector-space models, and more general machine-learning methods.

2.2.1 Topic Modelling

Amongst the various topic modelling algorithms, *Latent Dirichlet Allocation* (LDA)[17] is one of the most commonly used techniques, but most other algorithms take a similar approach. The fundamental idea behind the topic models is that topics are made up of words, with each word having a likelihood attached that defines how important it is for the topic. Next, a document is seen as a collection of topics in the same way, again with each topic having a likelihood attached to define how much of the document it represents. Finally, the words we see in the text are assumed to have been selected through a random selection of topics and then randomly selecting words from the topics, the random selection being weighted by the topic and word likelihoods.

As it is unknown what exactly the topics are and how they appear in the document, the topic modelling algorithms use the observed co-occurrence data to automatically infer the topic to word and document to topic likelihoods. The result is generally a list of topics and then for each topic the researcher is shown the most important words for the topic. These results can then be interpreted manually to analyse the topics that appear in the data-set.[18] An important aspect of modern topic modelling algorithms is that they generally assign more than one topic to a document, making it possible to model the nuances of different topics co-existing in a document.

16 Romain Vuillemot et al.: What's Being Said Near "Martha"? Exploring Name Entities in Literary Text Collections, in: IEEE Symposium on Visual Analytics Science and Technology, Atlantic City 2009, pp. 107–114.

17 David M. Blei, Andrew Y. Ng, Michael I. Jordan: Latent dirichlet allocation, in: The Journal of Machine Learning Research 3 (2003), pp. 993–1022.

18 Jonathan Chang et al.: Reading Tea Leaves: How Humans Interpret Topic Models, in: Advances in Neural Information Processing Systems 22, Vancouver 2009, pp. 288–296.

ISBN Print: 9783847111771 – ISBN E-Lib: 9783737011778

2.2.2 Vector-space models

Vector-space models come from the information retrieval world, but in the DH field are mostly used to describe word semantics models like *word2vec*[19]. The idea behind these models is that the meaning of a word can be defined through the words that it co-occurs with.[20] The underlying co-occurrence data is generally represented as a matrix where each cell in the matrix indicates how often two specific words co-occur. The problem with these matrices is that they are generally very sparse, meaning that most cells in the matrix are zero (as those two words never co-occur). What models like *word2vec* do is use machine learning algorithms to reduce the number of dimensions in the matrix from many thousands to a few hundred. In the resulting lower dimension models similar words tend to cluster together, thus by comparing the distances between pairs of words in the model, it is possible to determine which words are semantically more or less related.

2.2.3 Machine learning

The third approach to dealing with the volume of data is more general, namely *artificial neural network* (ANN) algorithms, which have become very popular in DH. In ANNs, the starting point is a data-set consisting of pairs of input data and output label. The aim of the ANN training is then to learn how to transform the input data into the output labels. To this end ANNs have three layers. An input layer that represents the data to learn and an output layer that represents the labels you want the ANN to output for the input layer. In between these you have at least one hidden layer that connects the input to the output layer. Each connection between two nodes in the ANN is weighted and each node in the ANN contains a discontinuous function that defines what level of signal has to come over all conceptions from the predecessor nodes in order for this node to activate. The aim of the learning process is then to learn the connection weights and activation functions based on the training data that is provided.

To achieve this, ANNs generally need very large sets of labelled data in order to learn the model. The reason for this is that in the learning process the ANN basically classifies each data-point in the training set and compares the generated classification to the manual labelling. This will include errors, and

19 Tomas Mikolov et al.: Distributed Representations of Words and Phrases and their Compositionality, in: Advances in Neural Information Processing Systems 26, Lake Tahoe 2013, pp. 3111–3119.

20 John R. Firth: A Synopsis of Linguistic Theory, in: Studies in Linguistic Analysis, Oxford 1957, pp. 1–32.

ISBN Print: 9783847111771 – ISBN E-Lib: 9783737011778

where the model makes mistakes, corrections for the mistake are applied starting from the output layer, moving through the hidden layers, until we get to the input layer, a process known as back-propagation. This learning process is repeated multiple times until the error rate stabilises.

One of the biggest risks with training ANN models is 'overfitting'.[21] Overfitting means that the algorithm learns a model that fits the training data very precisely, but which if given data that it has not seen before, makes huge mistakes. To address this, the learning process is generally given a second data-set, the so-called 'test' data, which after a number of iterations of the learning process is used to evaluate the model. This allows the learning process to determine when it is no longer improving the quality of the model and is in risk of overfitting. It does this by looking at whether the classification accuracy of the model decreases as the model is trained. If the accuracy no longer decreases, then the model has been overfitted and training should be stopped and the previous model used as the final model.

All these techniques open up humanities research to working with amounts of data that exceed what can be handled manually, but by increasing the distance between the researcher and the source material, they introduce new error sources into the research process, which will be discussed below.

3 Risks in the Digital Humanities

As illustrated above, digital tools offer new opportunities, but they also come with significant new risks. The rise in the amount of data and range of new methods that can be used in DH has been accompanied by a significant amount of work focused on a critical analysis of DH itself. Initially much of this focused on the question of balance between theory and practice.[22] To that the discussion then added a focus on issues around cultural criticism aspects of

21 Tom Dietterich: Overfitting and Undercomputing in Machine Learning, in: ACM Computing Surveys 27.3 (1995), pp. 326–327; Nitish Srivastava et al.: Dropout: A Simple Way to Prevent Neural Networks from Overfitting, in: The Journal of Machine Learning Research 15.1 (2014), pp. 1929–1958.

22 Andrew Prescott: Making the Digital Human: Anxieties, Possibilities, Challenges, in: Digital Riffs 2012. http://digitalriffs.blogspot.com/2012/07/making-digital-human-anxieties.html (last access: 21.06.2019); Tibor Koltay: Library and information science and the digital humanities: Perceived and real strengths and weaknesses, in: Journal of Documentation 72.4 (2016), pp. 781–792; Bethany Nowviskie: On the Origin of "Hack" and "Yack", in: Matthew K. Gold, Lauren F. Klein (eds.): Debates in the Digital Humanities, Minneapolis 2016, pp. 66–70; Tanya E. Clement, Daniel Carter: Connecting theory and practice in digital humanities information work, in: Journal of the Association for Information Science and Technology 68.6 (2017), pp. 1385–1396.

ISBN Print: 9783847111771 – ISBN E-Lib: 9783737011778

DH research.[23] All of these make very good points and particularly the issues around cultural criticism in DH research remain largely unaddressed, but the one thing all of these overlook are risks to DH research caused by methodologically poor use of the available data, tools, and methods. However, unlike work such as Da's critical analysis of computational literature studies[24], I intend to give a broader overview over methodological issues, rather than analysing a specific area in depth.

The risks presented here are indicative of how the digital aspects are being used by researchers and are not inherent in the digital aspects themselves. Thus, their existence does not imply that the DH field should be abandoned in whole or even in parts, only that the field needs to become more aware of these issues and incorporate them into its research practices. As such any examples selected in this section are to be treated as exemplars of a general trend, not specifically failing projects.

3.1 Data-driven Weaknesses

The first and most common methodological weakness that I want to address is missing criticism or critical analysis of the data-sets based on which the research is conducted. As discussed above, in non-computational humanities research a manual selection of a sub-set of the available data is always necessary to allow the research to be completed in a realistic time-frame. At the same time, it is also known that the selection processes are influenced by social and economic pressures, as the gender and colonial-studies research show. Due to this a critical analysis of the sources and the selection process have historically been a core aspect of humanities research and work in recent times on gender and colonial aspects have pushed this ever more into the foreground.[25]

23 Alan Liu: Where is Cultural Criticism in the Digital Humanities?, in: Gold: Debates, pp. 490–509; Gerben Zaagsma: On Digital History, in: BMGN – Low Countries Historical Review 128.4 (2013), pp. 3–29; David Berry et al.: No Signal without Symbol: Decoding the Digital Humanities, in: Matthew K. Gold, Lauren F. Klein (eds.): Debates in the Digital Humanities, Minneapolis 2019. https://dhdebates.gc.cuny.edu/projects/debates-in-the-digital-humanities-2019 (last access: 07.02.2020).

24 Nan Z. Da: The Computational Case against Computational Literary Studies, in: Critical Inquiry 45.3 (2019), pp. 601–639.

25 Jenny Bergenmar, Katarina Leppänen: Gender and Vernaculars in Digital Humanities and World Literature, in: NORA – Nordic Journal of Feminist and Gender Research 25.4 (2017), pp. 232–246; Roopika Risam: Decolonizing Digital Humanities in Theory and Practice, in: Jentery Sayers (ed.): The Routledge Companion to Media Studies and Digital Humanities, New York 2018, pp. 98–106.

ISBN Print: 9783847111771 – ISBN E-Lib: 9783737011778

In the DH this critical analysis of the data-sources and their potential biases has taken a back-seat. As DH algorithms allow the analysis of very large collections of objects, there is a perception that it is no longer a necessity to select a sub-set, instead 'everything' that is available can be fed into the algorithm. It seems that as a consequence the question of how the collection that is being analysed has been created is barely considered or completely ignored.

All the texts in the corpus must, at the time of their initial collection, have been judged to be sufficiently worthy to be included in a library or archive somewhere. Then, the texts will have to have survived two world wars, a period of active book-burning, water damage, theft, and a wide range of other physical risks. Then the texts will have to have been digitised, where a further two hurdles appear. First the texts will have had to have been selected for digitisation. Digitisation is both a time and financially expensive process, it is thus generally impossible to digitise the whole collection in one go and selections have to be made. In the ideal case, the selection is done topically or temporally to minimise bias, but at the beginning of the process or where third-party funding is used for the digitisation, the focus is generally on the 'important' or 'valuable' works, as these make it easier to argue that the digitisation is worthwhile.

The final step is the transformation of the scanned images into text (where appropriate). Due to the scale of the task, automated processes such as *Optical Character Recognition* have to be employed.[26] As the OCR output is heavily dependent on the quality of the print, the typesetting, and the age of the text[27], texts that can be processed successfully tend to be processed first.

Due to all these reasons the digital source is already biased in a way that can make it quite hard to pin down exactly how strong the bias is, before we even look at finding the objects that make up the analysis data-set. Consider the German text corpus available via the *Deutsches Textarchiv*.[28] Unlike other online corpora they provide some information online about the content of the archive and how it was curated, giving some indication of potential biases. However, even there the level of detail is limited ('selected bibliographies', 'specialist recommendations') and they explicitly state that the initial focus was

26 Ahmad P. Tafti et al.: OCR as a Service: An Experimental Evaluation of Google Docs OCR, Tesseract, ABBYY FineReader, and Transym, in: George Bebis et al. (eds.): Advances in Visual Computing, Cham 2016, pp. 735–746.

27 Florian Fink, Klaus U. Schulz, Uwe Springmann: Profiling of OCR'ed Historical Texts Revisited, in: Proceedings of the 2nd International Conference on Digital Access to Textual Cultural Heritage, Göttingen 2017, pp. 61–66.

28 http://www.deutschestextarchiv.de/, selected as an example because they actually provide some information on its creation (last access: 07.02.2020).

ISBN Print: 9783847111771 – ISBN E-Lib: 9783737011778

on canonical works, with selected additions.[29] What was selected and why is not clear, but focusing on the canonical works will introduce significant bias in the data, regardless of what was added.

All of these issues should not really create an issue for humanities research, as the selection of sources and the critical analysis of these sources and the selection process is a core research concept. However, in the DH area, the whole process is frequently summarised as "a query for the topic empire was used and a total of 843 texts returned, which were downloaded in TEI and form the basis of the analysis" or "we acquired the corpus of letters sent by X". While there might have been critical engagement with the archive and a number of actual query keywords might have been used to get good coverage over the given topic, no details are given on this process, even though they might significantly skew the distribution of texts in the analysis corpus.

To undertake methodologically sound DH research, these methodological aspects need to be reported in detail. Where the required information is not available within the archive, an analysis of potential biases in the corpus should be undertaken and reported. A positive example is the work by Müller, where the digital source is analysed in detail to understand its specific attributes, these and the methods used are reported in detail, and where necessary reference is made back to the original physical source.[30] Without this detail, any conclusion that goes beyond "in the specific set of texts we analysed" cannot be relied upon, not because the conclusion is right or wrong, but because we have no idea of how the specifics of the data skew the analysis.

3.2 Algorithms are opinionated[31]

Computer Science's roots lie in mathematics and in particular in the field of boolean logic. Boolean logic holds that for any given question there can only be two possible answers: true or false. At the technical level, a computer is a collection of transistors which implement exactly this logic decision and where the combination of millions of these transistors allow us to implement the kind of complex logic flows that make DH happen.

29 Alexander Geyken et al.: Das Deutsche Textarchiv: Vom historischen Korpus zum aktiven Archiv, in: Silke Schomburg et al. (eds.): Digitale Wissenschaft. Stand und Entwicklung digital vernetzter Forschung in Deutschland, Köln 2011, pp. 157–161.

30 Andreas Müller: Vom Konversationslexikon zur Enzyklopädie, in: Das Achtzehnte Jahrhundert 43.1 (2019), pp. 73–89.

31 For an early treatment of this see David Sculley, Bradley M. Pasanek: Meaning and Mining: the Impact of Implicit Assumptions in Data Mining for the Humanities, in: Literary and Linguistic Computing 23.4 (2008), pp. 409–424.

ISBN Print: 9783847111771 – ISBN E-Lib: 9783737011778

This fundamental binary view of the world extends into all algorithms, as even where they are designed to support uncertainty, this is always simulated and fundamentally reduced to a binary representation. The effect of this is that algorithms generally are incapable of modelling when they cannot answer a question. An algorithm will always provide an 'answer' to a question, fundamentally splitting the set of input data elements into those for which the question is true and those for which it is not. This happens even if the data that is fed into the algorithm is not of the appropriate type or fulfils the constraints set be the algorithm.

The most common example of these kinds of mistakes is the use of fundamental statistical measures (which are also algorithms). Where quantitative results, for example frequencies of word co-occurrences, are calculated, the results for different conditions, for example words that co-occur with male vs. female characters, are often compared using Student's t-tests to prove that they are significantly different. However, the t-test requires that the measurements being compared are independent and normally distributed. Neither of these two are likely to be the case for the results of text analysis. Unlike for example the height of trees in two separate forests, the frequencies of words in a single text are possibly to be linked together quite strongly, breaking the independence assumption. Secondly, word distributions are generally more likely to follow a Zipf distribution,[32] breaking the normal distribution assumption. Now this might not be the case for a specific piece of research, say because the word counts have been drawn from unrelated texts and they happen to be normally distributed, but this needs to be reported. Otherwise any statistical significance is meaningless and any research conclusions drawn from it are as likely to be false as true, fundamentally invalidating any conclusions.

In addition to this fundamental bias towards always providing an answer, algorithms can also have specific opinions that may not be immediately visible. This is particularly the case for many of the machine-learning models that have become very popular in the recent years. As explained above machine-learning models take large amounts of manually annotated data and try to learn how to map the input to the desired annotation. While the results are often very promising, the difficulty is that it is unclear what has been learned. One of the classic examples are neural networks trained on the *ImageNet* data-set to recognise a wide range of objects. They work very well, but what researchers then found was that by manipulating small aspects of the objects in the images, they could get the neural network to completely misidentify the images. For example by making minute changes to the texture of an image of a turtle,

32 George K. Zipf: Human Behavior and the Principle of Least Effort, Cambridge, Mass. 1949, p. 24.

ISBN Print: 9783847111771 – ISBN E-Lib: 9783737011778

changes so small they are barely noticeable to a human, they could induce the neural network to recognise the image as an automatic rifle.[33] The problem arises because the neural network has actually learned to distinguish different texture patterns and image features that are often barely visible to the human eye. While for the data it was trained on this works, with unknown or slightly different data it quickly produces erroneous results.

This illustrates the susceptibility of neural networks nicely, but what is primarily relevant for DH is the factor that it is not clear what the neural networks are learning, thus it is not clear what their 'opinion' of the data is. For example the same neural network for image classification can be used to cluster historic photographs.[34] The results look very promising, but because we don't know what the model has learned, we don't know whether the results represent what we, as humans, see in the data (a clustering by image type: portrait, group photo, architecture, etc.) or some underlying bias. For example, because the technology used in photography changes over time, the neural network might actually be clustering the images based on technical aspects of the photographs, because it is picking up on variations caused by different cameras, film materials, or development processes. While this might coincide with different subject types, making the results look appropriate, when further images are added, that use, for example, the same film material in a different setting, it is possible that the algorithm would completely misclassify the images. As in the case of the data biases, it is not so much the existence of the bias that is the issue, but the fact that it is not clear whether there is any and what kind of bias is included.

3.3 Evaluation

Unlike non-computational humanities research, where the argument is the research output, in digital humanities research there is an intermediate step of data-generation and the research output is derived through interpretation from the data. Because this intermediate data-generation step is undertaken using an algorithm, it is necessary to evaluate the accuracy and correctness of the algorithm, before and independently from the conclusions drawn from its output. The biggest methodological weakness around evaluation is that in most pieces of digital humanities research it is missing. Two reasons for this are that

33 Anish Athalye et al.: Synthesizing Robust Adversarial Examples. arXiv preprint arXiv:1707.07397 (2017).

34 Peter Leonard: Lina Jonns Efterträdare: Machine Vision and Lund's Photographic History, in: Book of abstracts 4th Conference of The Association Digital Humanities in the Nordic Countries, Copenhagen 2019. https://cst.dk/DHN2019Pro/DHN2019BookofAbstracts.pdf (last access: 07.02.2020).

ISBN Print: 9783847111771 – ISBN E-Lib: 9783737011778

evaluating takes time and that assessing the algorithm's output with respect to the research question is conceptually mistaken for the act of evaluating the algorithm.

An area where this type of mistake is often seen is where topic modelling is used. In many cases only a selection of topics is actually reported on, detailing how they support the original hypothesis. However, if we generate random lists of words that frequently co-occur, there is likely to be at least a few lists which can be interpreted relative to the research question. This does not mean that the model as a whole is valid, as is never the case for a random sample. If the five or six reported topics are the only understandable ones out of a model with 30 topics, then this is not a good model and no conclusions should be drawn based on such a small selection. This, as Da points out, also goes for negative results. Only if it is clear that the negative results are not just random noise can any conclusions be drawn.

Correctly evaluating the algorithm means testing it with a range of input data and then assessing whether the output produced by the algorithm is correct for the given input, irrespective of the research question. In addition to not taking the research question into consideration, two basic rules need to be followed.

First, the algorithm must never be evaluated using the data it was trained on. This does not just mean splitting the data-set into a learning and an evaluation data-set, it means choosing an evaluation data-set that is drawn from a different source than the learning data-set. This is necessary, as even if the original data-set is randomly split into a learning and an evaluation data-set, there are likely to be underlying characteristics of the data-set that distinguish the data-set from all others and which introduce bias into the algorithm that limits its generalisability. By choosing an independent evaluation data-set, it is possible to fully test the degree to which the algorithm generalises.

During the development and evaluation of the algorithm the learning data-set is itself split into a training and a test data-set. When testing with the test data, the generated output can be compared against the expected outcome and can be used to improve the algorithm. However, when evaluating using the evaluation data, only accuracy or correctness metrics may be generated. The results of evaluation run must not be used to further improve the algorithm, as in that case the evaluation data is no longer separate from the training data, negating the benefits of the separate evaluation data-set.

The second basic rule is that the evaluation must be designed to counteract any known biases in the data. If, for example, the aim is to develop an algorithm that can classify text for the whole 19th century, then the evaluation data must be selected to provide even coverage over the target time period, even if

ISBN Print: 9783847111771 – ISBN E-Lib: 9783737011778

the training set has a bias towards texts later in the period. This is necessary, as otherwise the evaluation results do not provide a realistic assessment of the algorithm quality.

As part of this we should also become more open to reporting negative results. While this is something most disciplines struggle with, the humanities with their predisposition for critical self-analysis should really be primed for leading on this.

4 The Spectre of Techno-Positivism

In the sections above I have discussed what I see as both the opportunities and main risks facing the Digital Humanities at this point in time. However, the biggest risk I see in the future of the Digital Humanities is that as a discipline we drift into what I term 'techno-positivism'. We build more and more digital archives, digital editions, and many other kinds of data-sets. We apply digital methods to the data-sets, we visualise and then describe the results. What we don't do on a large scale is critically engage with the data or the results. Our algorithms are not grounded in humanities theories, our results are barely contextualised based on existing knowledge. We are undertaking essentially positivist research, but hide that truth under large amounts of data, interactive interfaces, and fancy visualisations.

If, as a discipline, we do not address these issues, then we are complicit in dragging research back into the past, where gender, colonialism, canonisation, and related issues are ignored, where the Digital Humanities are the study of dead, white dudes[35]. This risk is significant, as this kind of research is sufficiently simplistic that it can be supported with current tools and algorithms. However, it also means that significant amounts of current Digital Humanities research should be labelled as interesting, but the results do not allow generalise to anything that was not in the analysed data, and no general conclusions can be drawn.

However, unlike other critical commentators I do not argue that the Digital Humanities should be abandoned. Quite apart from the fact that the popularity of DH with funding bodies makes this a ludicrous idea, I believe that addressing these things head-on will actually allow DH research to fully achieve its potential. The current, techno-positivist approach allows the humanities researcher to create the data, which is then handed to the computer scientist for processing, and the results are interpreted by the humanities researcher. While

35 Lisa Marie Rhody: Why I Dig: Feminist Approaches to Text Analysis, in: Gold, Klein: Debates, pp. 536–539; Kim Gallon: Making a Case for the Black Digital Humanities, in: ibid., pp. 42–49.

ISBN Print: 9783847111771 – ISBN E-Lib: 9783737011778

this represents basic working together, it is not true collaboration. True collaboration requires an in-depth understanding of what the other does in order to truly benefit from their input and in order to provide responses back that allow the other to fully understand the value and nuances of the own work.

This level of collaboration will require computer scientists to develop algorithms that are grounded in complex theoretical humanities frameworks developed by humanities researchers. That see the world as more than a collection of word frequencies, collocations, and pixel colours. It will also require humanities scholars to fully understand the power and limitations of these algorithms, so that they can properly contextualise them and place the results in larger frameworks of understanding. This has significant benefits to both sides.

First, for computer scientists it will drive the development of truly novel algorithms that are able to model the complexities of reality, rather than just reducing reality until it fits into simple models. This will address the sometimes voiced question of why a computer scientist is needed on a project. Second, for humanities researchers it will produce truly novel methods, rather than simplistically scaled-up versions of the counting methods that were introduced by and popular in positivist research. This will allow humanities researchers to pose new research questions, rather than just technically rehashing basic questions. It might mean that results are possibly produced not quite as quickly, but it will produce results that have true value.

This can be achieved and we are seeing the slow drip-drip-drip of research taking these issues into account, but to truly achieve the aim of DH being a transformative new approach, we need to stop fooling ourselves that current computational methods achieve even close to the depth of non-computational humanities research methods. Counting words is not reading text. Applying machine learning is not understanding. The focus needs to move from what is technically possible to how we can transfer at least parts of humans' deeper understanding of these issues to the algorithmic world. Then DH will be truly transformational.

> "The first principle is that you must not fool yourself—and you are the easiest person to fool. So you have to be very careful about that. After you've not fooled yourself, it's easy not to fool other scientists. You just have to be honest in a conventional way after that."
> Richard P. Feynmann, Caltech Commencement Address, 1974

ISBN Print: 9783847111771 – ISBN E-Lib: 9783737011778

Acknowledgements

I would like to thank my colleagues at the Digital Humanities Meetup in Halle for the many constructive discussions that helped focus my thinking around these methodological issues. I would like to particularly thank Andreas Müller and Thorsten Roeder for the constructive feedback that has significantly improved this article.

References

Anish Athalye et al.: Synthesizing Robust Adversarial Examples. arXiv preprint arXiv:1707.07397 (2017).

Ricardo Baeza-Yates, Berthier Ribeiro-Neto: Modern Information Retrieval, 2nd ed., Harlow et. al. 2011.

Kimberly Barata: Archives in the digital age, in: Journal of the Society of Archivists 25.1 (2004), pp. 63–70.

Jenny Bergenmar, Katarina Leppänen: Gender and Vernaculars in Digital Humanities and World Literature, in: NORA – Nordic Journal of Feminist and Gender Research 25.4 (2017), pp. 232–246.

David Berry et al.: No Signal without Symbol: Decoding the Digital Humanities, in: Matthew K. Gold, Lauren F. Klein (eds.): Debates in the DigitalHumanities, Minneapolis 2019. https://dhdebates.gc.cuny.edu/projects/debates-in-the-digital-humanities-2019.

David M. Blei, Andrew Y. Ng, Michael I. Jordan: Latent dirichlet allocation, in: The Journal of Machine Learning Research 3 (2003), pp. 993–1022.

Jonathan Chang et al.: Reading Tea Leaves: How Humans Interpret Topic Models, in: Advances in Neural Information Processing Systems 22, Vancouver 2009, pp. 288–296.

Tanya E. Clement, Daniel Carter: Connecting theory and practice in digital humanities information work, in: Journal of the Association for Information Science and Technology 68.6 (2017), pp. 1385–1396.

Nan Z. Da: The Computational Case against Computational Literary Studies, in: Critical Inquiry 45.3 (2019), pp. 601–639.

Tom Dietterich: Overfitting and Undercomputing in Machine Learning, in: ACM Computing Surveys 27.3 (1995), pp. 326–327.

Florian Fink, Klaus U. Schulz, Uwe Springmann: Profiling of OCR'ed Historical Texts Revisited, in: Proceedings of the 2nd International Conference on Digital Access to Textual Cultural Heritage, Göttingen 2017, pp. 61–66.

John R. Firth: A Synopsis of Linguistic Theory, in: Studies in Linguistic Analysis, Oxford 1957, pp. 1–32.

Kathleen Fitzpatrick: The Humanities, Done Digitally, in: Matthew K. Gold (ed.): Debates in the Digital humanities, Minneapolis 2012, pp. 12–15.

Kathleen Fitzpatrick: The Humanities, Done digitally, in: The Chronicle of Higher Education. https://www.chronicle.com/article/The-Humanities-Done-Digitally/127382.

ISBN Print: 9783847111771 – ISBN E-Lib: 9783737011778

Kim Gallon: Making a Case for the Black Digital Humanities, in: Matthew K. Gold, Lauren F. Klein (eds.): Debates in the Digital Humanities, Minneapolis 2016, pp. 42–49.

Alexander Geyken et al.: Das Deutsche Textarchiv: Vom historischen Korpus zum aktiven Archiv, in: Silke Schomburg et al. (eds.): Digitale Wissenschaft. Stand und Entwicklung digital vernetzter Forschung in Deutschland, Köln 2011, pp. 157–161.

Safaa I. Hajeer et al.: A New Stemming Algorithm for Efficient Information Retrieval Systems and Web Search Engines, in: Aboul E.Hassanien et al. (eds.): Multimedia Forensics and Security. Foundations, Innovations, and Applications, Cham 2017, pp. 117–135.

Stefan Jänicke et al.: On Close and Distant Reading in Digital Humanities: A Survey and Future Challenges, in: Rita Borgo, Fabio Ganovelli, Ivan Viola: Eurographics Conference on Visualization 2015, pp. 83–103. DOI: https://doi.org/10.2312/eurovisstar.20151113.

Sergio Jimenez et al.: BM25-CTF: Improving TF and IDF factors in BM25 by using collection term frequencies, in: Journal of Intelligent & Fuzzy Systems 34.1 (2018), pp. 1–13.

Matthew G. Kirschenbaum: What is Digital Humanities and What's It Doing in English Departments, in: Matthew K. Gold (ed.): Debates in the Digital humanities, Minneapolis 2012, pp. 3–11.

Tibor Koltay: Library and information science and the digital humanities: Perceived and real strengths and weaknesses, in: Journal of Documentation 72.4 (2016), pp. 781–792.

Peter Leonard: Lina Jonns Efterträdare: Machine Vision and Lund's Photographic History, in: Book of abstracts 4th Conference of The Association Digital Humanities in the Nordic Countries, Copenhagen 2019. https://cst.dk/DHN2019Pro/DHN2019BookofAbstracts.pdf.

Alan Liu: Where is Cultural Criticism in the Digital Humanities?, in: Matthew K. Gold (ed.): Debates in the Digital humanities, Minneapolis 2012, pp. 490–509.

Tomas Mikolov et al.: Distributed Representations of Words and Phrases and their Compositionality, in: Advances in Neural Information Processing Systems 26, Lake Tahoe 2013, pp. 3111–3119.

Franco Moretti: Distant Reading, London 2013.

Andreas Müller: Vom Konversationslexikon zur Enzyklopädie, in: Das Achtzehnte Jahrhundert 43.1 (2019), pp. 73–89.

Bob Nicholson: The Digital Turn: Exploring the methodological possibilities of digital newspaper archives, in: Media History 19.1 (2013), pp. 59–73.

Bethany Nowviskie: Digital Humanities in the Anthropocene, in: Digital Scholarship in the Humanities 30.1 (2015), pp. i4–i15.

Bethany Nowviskie: On the Origin of "Hack" and "Yack", in: Matthew K. Gold, Lauren F. Klein (eds.): Debates in the Digital Humanities, Minneapolis 2016, pp. 66–70.

Andrew Prescott: Making the Digital Human: Anxieties, Possibilities, Challenges, in: Digital Riffs 2012. http://digitalriffs.blogspot.com/2012/07/making-digital-human-anxieties.

Lisa Marie Rhody: Why I Dig: Feminist Approaches to Text Analysis, in: Matthew K. Gold, Lauren F. Klein (eds.): Debates in the Digital Humanities, Minneapolis 2016, pp. 536–539.

Roopika Risam: Decolonizing Digital Humanities in Theory and Practice, in: Jentery Sayers (ed.): The Routledge Companion to Media Studies and Digital Humanities, New York 2018, pp. 98–106.

ISBN Print: 9783847111771 – ISBN E-Lib: 9783737011778

Stephen E. Robertson et al.: Okapi at TREC-3, in: Proceedings of the Third Text REtrieval Conference, Gaithersburg 1995, pp. 109–126.

Christoph Schöch: Big? Smart? Clean? Messy? Data in the Humanities, in: Journal of Digital Humanities 2.3 (2013), pp. 2–13.

David Sculley, Bradley M. Pasanek: Meaning and Mining: the Impact of Implicit Assumptions in Data Mining for the Humanities, in: Literary and Linguistic Computing 23.4 (2008), pp. 409–424.

Nitish Srivastava et al.: Dropout: A Simple Way to Prevent Neural Networks from Overfitting, in: The Journal of Machine Learning Research 15.1 (2014), pp. 1929–1958.

Patrik Svensson: Humanities Computing as Digital Humanities, in: Melissa Terras, Julianne Nyhann, Edward Vanhoutte (eds.): Defining Digital Humanities, London/New York 2016, pp. 175–202.

Ahmad P. Tafti et al.: OCR as a Service: An Experimental Evaluation of Google Docs OCR, Tesseract, ABBYY FineReader, and Transym, in: George Bebis et al. (eds.): Advances in Visual Computing, Cham 2016, pp. 735–746.

Ted Underwood: A Genealogy of Distant Reading, in: DHQ: Digital Humanities Quarterly 11.2 (2017). http://www.digitalhumanities.org/dhq/vol/11/2/000317/000317.html.

Ted Underwood: Dear Humanists: Fear Not the Digital Revolution, in: The Chronicle of Higher Education 2019. https://www.chronicle.com/article/Dear-Humanists-Fear-Not-the/245987/.

John Unsworth: What is Humanities Computing and What is Not?, in: Melissa Terras, Julianne Nyhann, Edward Vanhoutte (eds.): Defining Digital Humanities, London/New York 2016, pp. 51–63.

Romain Vuillemot et al.: What's Being Said Near "Martha"? Exploring Name Entities in Literary Text Collections, in: IEEE Symposium on Visual Analytics Science and Technology, Atlantic City 2009, pp. 107–114.

Gerben Zaagsma: On Digital History, in: BMGN – Low Countries Historical Review 128.4 (2013), pp. 3–29.

George K. Zipf: Human Behavior and the Principle of Least Effort, Cambridge, Mass. 1949.

ISBN Print: 9783847111771 – ISBN E-Lib: 9783737011778