A Convergence of Methodologies: Notes on Data-Intensive Humanities Research

Nina Tahmasebi, Niclas Hagen, Daniel Brodén, Mats Malm

University of Gothenburg, Sweden
{nina.tahmasebi, niclas.hagen, daniel.broden, mats.malm}@gu.se

Abstract. In this paper, we discuss a data-intensive research methodology for the digital humanities. We highlight the differences and commonalities between quantitative and qualitative research methodologies in relation to a data-intensive research process. We argue that issues of representativeness and reduction must be in focus for all phases of the process; from the status of texts as such, over their digitization to pre-processing and methodological exploration.

1 Introduction

It is common among scholars in the field of the humanities to emphasize the inherent differences between the methodologies of humanistic and natural-scientific research. Although the differences between the humanities and the natural sciences have been conceptualized in various ways – whether it has to do with different fields of interest, ontologies, etcetera – commentators have tended to warn that the humanities should not model itself after the methods of natural science (see Kjørup 2009: 75–91 [11]).

However, the merging of traditional humanities with quantitative approaches has opened new methodological venues that point towards the natural sciences rather than the interpretive, largely qualitative methodology associated with humanities. Digital humanities is to some extent a merging of two fields, the humanities and data science[1]. The aim of data science is to enable a high-level overview of our data and give us the possibility to discern patterns that could otherwise not be seen, and to grasp quantities and time spans of data that that we otherwise would have no hope to cover in a single lifetime. While traditional humanities has made use of relatively small amounts of text, data science has, in theory, the possibility to handle infinite amounts[2]. Still, data science develop methods that are limited to small scopes and single questions, while the qualitative humanities can see beyond individual texts and times, to answer multi-faceted questions related to cultural contexts, societal and identity conditions and change. Qualitative humanities can potentially capture the suggestive, sensory, existential, contradictory, ambivalent and ambiguous aspects stemming

[1] While data science uses scientific methods and processes to extract knowledge and insights from data of all kinds, in this paper, we consider only textual data.

[2] The limiting factor is most often the amount of available, relevant texts.

from the human imagination and experience. This is the domain expertise that the humanities bring into the meeting with data science.

The convergence between the humanities and data science brings with it clear benefits, including the possibilities to base conclusions on large and representative samples of text. This paper focuses on methodological convergence. Commentators have previously noted that, apart from offering alternative methodological and epistemological venues for traditional humanities, the expansion of digital humanities has also given rise to great number of questions and issues that need to be carefully considered from an epistemological perspective (see e.g. Berry 2012 [5]; Kitchin 2014 [10]). While these issues certainly have been treated within the digital humanities, and even more within the social sciences, there is need for clarifying the conditions of data-intensive research in the humanities more systematically.

The overall purpose of the paper is to present a starting-point for addressing some specific epistemological issues induced by the methodological convergence of humanities and data science. We discuss a number of interrelated issues and concepts concerning representativity and reduction, as they apply to method, data and results. Our ambition is not to be comprehensive, and each projects will inevitably be concerned with these issues to different degrees. We highlight some key features of the data-intensive humanities research process, in the hope to further an important discussion in, not only the digital humanities community, but also within the traditional humanities.

The disposition of the paper is as follows: We begin by discussing the data-intensive research process. A hypothesis-driven process is natural in data science, while the humanities are more often driven by a research question. In this section we connect the two, in order to clarify how research questions and hypotheses can work in data-intensive humanities research. Second, we discuss use of models as one instance of reduction. Third, we discuss the validation of results from the data-intensive process. Fourth, we discuss how the rise of digital humanities and quantitative methodologies places traditional humanistic approaches in another light. We conclude by emphasizing the current challenge for the digital humanities community in exploring, interpreting, validating, evaluating data science methods and, not least, discussing its further implications for the humanities.

2 Data-Intensive Digital Humanities Research

To a great extent, digital humanities employs data science methods to gain insights into large scale, often diachronic, collections of digital text. While this has been done on other materials and for other purposes in data science, the aim in digital humanities is to generate humanities knowledge. This collaboration should, in the optimal case, offer the possibility for the humanities to base their conclusions on larger and representative samples of text, as well as offer possibilities to ask other kinds of research questions. This alternative methodology should generate knowledge that is sustainable and robust against time, and additional scrutiny of the same or additional sources. In addition, it should

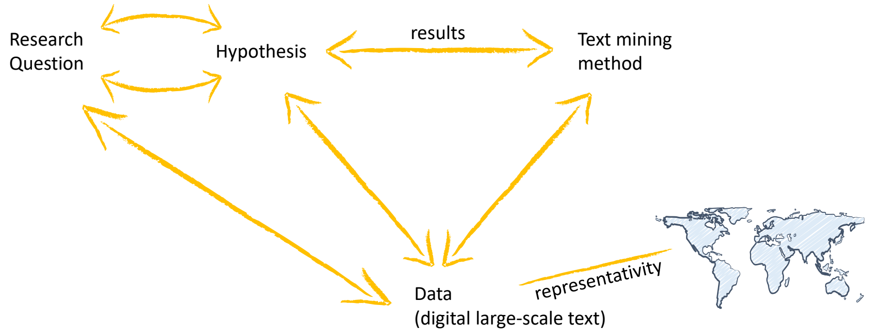

Fig. 1. A schematic model of the research process in data-intensive humanities.

offer the data science new, broader kinds of problems to target their methods towards.

However, typical digital humanities projects are conducted with either a strong data science or humanities bias. The data science projects, on the one hand, are often conducted with a computer science, math, or language technology perspective where the interpretation and the understanding of the research questions are sacrificed at the expense of mining techniques and large quantities of data. The humanities projects, on the other hand, are often conducted on smaller scale data using methods that may not be the best suited for the problem, or data, at hand.

2.1 The Data-Intensive Research Process

If we concentrate on the methodologies of systematic data-intensive research, it typically has a clear process and several important components. There is data, a text mining method and results. Motivating this are research questions and hypotheses. In the process of data-intensive research, there are two main methods for making use of large scale text. First, in an **exploratory fashion** to find and formulate interesting hypotheses, that is, work departs from a general research question. Alternatively, one starts with a well-defined hypothesis and employs large scale text to find evidence to support or reject the hypothesis in a **validating fashion**. For both of the above, there is a research question involved that can correspond to one or several different hypotheses.

The process can be schematically illustrated as in Fig. 1. Both the exploratory and the validation paths follow the same process, but start at different points. The exploratory path moves from the research question to hypothesis via data and text mining method. The validation path has already been boiled down to one or several clearly defined hypotheses and starts from there. Then data and methods can better be chosen with respect to the research question at hand. The exploratory path can be said to aim at discovering patterns, while the validation path is aimed at demonstrating or proving patterns.

In both paths, a text mining method is employed to generate (directly interpretable or aggregated) results from the text.

2.2 Research Questions and Hypotheses

One of the great challenges of digital humanities is to reason about how results from text mining (or other kinds of data science) can be used to corroborate or reject a hypothesis, and as an extension, can contribute to the wider research question. This amounts to interpreting the results and "translating" them into conclusions about the original research question. Here is where the humanities' in-depth domain knowledge comes into play. However, let us compare three different starting points for a research process when it comes to the relationship between research question and hypothesis.

1. *One research question and one hypothesis:* A researcher is interested in how the general sentiment with regards to a concept, like a trade or technology, has changed over time. The research question focuses on "how", and data and method are designed so as to follow the exploratory path. If this results in a hypothesis about more precisely how notions changed, then this hypothesis "that" can be corroborated or refuted through the validation path with adjusted data and method.
2. *One research question and several hypotheses:* A researcher is interested in how a certain technology, means of transportation or communication, has affected society. This research question needs to be broken down into several, and a number of them must be used to answer the question in full. *Which sentiments were there? Which new behaviors were the result? Which facets of life were affected?* Suitable data and method need to be devised and by following the exploratory path, these questions can be reformulated as propositions: hypotheses, which are tested using the validating path.
3. *Data and text mining method but no research question:* We can envision the case when there is an interesting source of data but no clear research questions (for example, the digitized letters of an influential author). A text mining method can be used to find interesting patterns and signals to explore further. That is, we follow the exploratory path to find a rewarding hypothesis. The focus is the data and the text mining method. Often times, a method like topic modeling is used as a way of getting an overview of different themes around a concept of interest. These topics can be explored and good hypotheses formulated in a more informed fashion.

2.3 Interpretation of Results

In traditional humanities, the researcher is the bridge between results and interpretation. In data-intensive humanities research, the situation is slightly different. The typical result of a text mining method is not necessarily directly interpretable for the hypothesis, nor need the hypotheses be directly interpretable with respect to the research question. The process of moving between results and the research question is in itself a result and in need of evaluation.

To exemplify a "model of interpretation" (see more on models in the context of humanities in section 3), we go outside of the digital humanities where data-intensive research is tied to societal impact (SCB, [19]). To measure *severe material poverty* in the EU, a set of nine different criteria are measured. These include the possibility to pay for unforeseen expenses, have a color TV, phone or car, and have sufficient heating in the home. If someone cannot fulfill at least four of the criteria, they are considered to live in severe material poverty.

Each individual criterion is measured by asking a sample of a countries' residents.To measure the criteria (hypotheses in our case), each person is asked a set of questions (corresponding to the text-mining methods). These questions are aimed at capturing the phenomena that we are interested in. One example question is *Can you afford at least two pair of shoes?* The answers from each participant are then weighted according to group belonging, and groups with lower answering rates are weighted up to get a fair representation. This weighting is a part of the interpretation; different weightings provide different results and should be subject to evaluation and discussion. In addition, the number of criteria needed to define material poverty is also an essential part of the interpretation of results. All of these acts go into a model of interpretation. In this example, parts of the model are clearly defined and can be subject to discussion and evaluation. We argue that all digital humanities-projects should make their model of interpretation clear, and preferably evaluate with respect to alternative models.

2.4 Reduction as a part of the Data-Intensive Process

The process of formulating appropriate hypotheses to answer a research question is almost always reductive: narrowing focus to particular aspects, sorting out in order to clarify. In all cases, data is central. Data-intensive research can only provide results for that which is represented in the data. Already at this point, it is important to reason about representativity. Text is always a reduction of the world, and only representative for a part of it. In historical text, like the Google books corpus, men are almost ten times more likely to be mentioned than women, until the beginning of the 20th century, when the two begin moving towards the middle and finally meet somewhere in the 1980's ([8]). Other socioeconomic factors also play a role, both for modern and historical texts. Different genres obviously represent society in different ways and with different restrictions. Therefore, when using text to study cultural or social factors, it is important to remember who is present in the text (and who is not).

These conditions of text as data are general: to them come methodological reductions. Typical digital humanities applications have restricted access to text corpora, or reduce them for their specific purposes and make use of collections of digitized newspapers, literature, or social media, to answer research questions. Reduction can be seen as the other side of the coin of representativity; typically, the more representativity, the less reduction. At this point, a first methodological reduction is taking place. Then come other kinds of methodological reduction, as part of the method used. The model as such provides an illustrative example of methodological reduction and Fig. 1 is one example. Consequently, the usage of scientific models within data-intensive approaches constitute another case of reduction. In the next section, we discuss the issue of scientific models, and their function and implication for the digital humanities.

3 Scientific Models as an Instance of Reduction

Models and their usage within science has been an object of debate within philosophy and theory of science for a long time, yet these discussions have almost exclusively been devoted to the usage and importance of models within the natural sciences and not within the humanities (e.g. Bailer-Jones 2009, 2003; Giere 2004; Morgan and Morrison 1999). Yet, several aspects within these discussions have implications for the purpose of this paper as the issue of models and modeling are an intrinsic reductive feature within the methods of data-intensive humanities.

Daniela Bailer-Jones defines a scientific model as a “description of a phenomenon that facilitates access to that phenomenon” (2009: 1 [3]). Margaret Morrison’s and Mary S. Morgan’s analysis of the functions and usage of models present a similar viewpoint, where they mean that models not only function as a means of access but also as a means of representation as they represent either some aspect of the world, or some aspect of our theories about the world, or both at once. Hence a model’s representative power allows it to function not just as a way to gain access, but to teach us something about the phenomena it represents (Morrison and Morgan 1999: 11–12 [15]). However, as part of their function as a means of representation and as a tool for intervention, models also carry with them a reductive power. Models practically never aim to describe or represent a phenomenon in its entirety, scientific models always involve certain simplifications and approximations (Bailer-Jones 2003: 66 [2]; Morrison and Morgan 1999: 16 [15]). Models then simplify the phenomena in question and through that reduction they are intended to capture the essence of something while leaving out less essential (from the perspective at hand) details. So, a model within the natural sciences involves a process of reduction wherein the phenomena or problem areas that stand in focus are reduced or simplified in order for researchers to be able to produce scientific knowledge about them.

Within the humanities, at least within mainstream humanities, the usage of models is still rare (McCarty 2004: 255 [12]). Nevertheless, it can be argued that models and the practice of modeling are at the very heart of digital humanities (Rockwell and Sinclair 2016). The various phenomena that stand in focus for digital humanities research not only have to be represented through a model, but even the process whereby the researcher investigates the phenomena in question needs to be translated into a model. Approaching the phenomena with methods and tools developed within digital humanities, the models give research a leverage. Moreover, models can also be adapted and tested by others in a way that more “verbal” theories cannot do (Rockwell and Sinclair 2016 164 [18]). A text, then, is modeled through digitization that transforms the original text into a digital model of the text. That model, if sufficiently rich, can serve as a surrogate for the text. The digital model can then be processed using implementations of hermeneutical processes. The text and the tools are thus both implementations of theoretical models of the phenomenon of interpretation. One is a usable model of the interpretation as edition; the other a reusable model of interpretation as process (Rockwell and Sinclair 2016: 164 [18]). We may add that the model of the text reduces the original text, while the model of the interpretation reduces the first model, that of the text.

In relation to the function and usage of models within natural science, models within digital humanities seem to attain the same kind of functions and usage. For example, Willard McCarty points up on how models offer both possibilities for an increased access and manipulation, as well as a way for representation of phenomena: 'By "modeling" I mean the heuristic process of constructing and manipulating models; a "model" I take to be either a representation of models within something for purposes of study, or a design for realizing something new' (McCarty 2004: 255 [12]). Both the possibility for representation and manipulation of phenomena are pointed upon by McCarty in his outline of the function and usage of models within digital humanities. Moreover, what also reoccurs in McCarty's discussion is that models and modeling within the field, as in the natural sciences, is a form of reduction: 'a model is by nature a simplified and therefore fictional or idealized representation, often taking quite a rough-and-ready form' (McCarty 2004:255 [12]).

We can say that models and modeling within digital humanities carry a reductive potential; that is, at the same time as models - by simplification - offer a number of advantages (they both represent and offer possibilities for manipulation of phenomena of interest), they reduce the phenomena that we want to investigate. Consequently, models do carry an intrinsic reductive power that we have to be aware of; *of what is the model representative*?

4 Representativity and Reduction

In all phases of the data-intensive humanities research process, reductions take place. The first, basic step of reduction is that our texts cannot be fully representative for all phenomena that we want to study. Next, our selection of the texts further reduces our base. The digitization of the text is, as explained above, a model of the original text and thus a reduction in itself. Depending on the quality of the digitization, more or less of the textual information is intact. Finally, the pre-processing and the text mining methods additionally reduce the original text in several ways.

Large scale texts cannot be studied by taking all aspects and words into account. It resembles the situation with creating a dictionary of language usage. Not all words, nor every single usage of a word, can be included in the dictionary, without making the dictionary as large as the text itself, and thus rendering it useless. Instead, a dictionary generalizes the usage. In text mining, the generalization is done by focusing on certain aspects of a text, or certain parts, or both. A typical method for focusing in large scale text mining is to keep only the most frequent words (typically ranging from the 10 000 to the 250 000 most frequent word forms), or to keep words of certain part-of-speech (nouns, noun phrases, verbs, adjectives etc). Stopwords and function words (words that carry little meaning but are very frequent) are often filtered out to speed up the process and to increase the chances of the methods to find relevant information. In each filtering step, a reduction is made.

In the example: "... Cecilia was a very clever woman, and a most skillful counter-plotter to adversity." we keep "Cecilia was / clever woman / skillful counter-plotter / adversity"after filtering function words and stopwords. Only "Cecilia", "woman" , and "counter-plotter" are kept if we focus on nouns and only "was" if we keep verbs. Additional words might be removed if we filter on frequency as well.

The original information is thus reduced at the preprocessing step. In addition, the text-mining method itself also performs a reduction. To see patterns of different kinds, we cannot view all of the words at the same time. Methods like topic modeling, clustering and word representations of different kinds lead to a set of results. These results need not include all text. For example, topic modeling can result in a set of topics that are not representative for each text on which the topic modeling was performed. Again, a reduction is performed. Additionally, the accounts need not include all results: often, the number is predetermined. Finally, the results are interpreted in support or refutation of one or several hypotheses, or as indications pertaining to a wider research question.

From the results of text mining, we have to draw conclusions that contribute to the knowledge of one or another aspect of the world. At this stage, we need to reason about representativity as the opposite to reduction. We began by discussing how texts are representative of a part of the world, and different texts have different representativeness. Social media texts for example, have a much higher representativeness of western young people, than of middle east elderly or women. Next, the reduction of the text mining method affects the representativity of the result. How much of the original text are the results valid for? Finally, the same question should be asked once we reach the conclusion step; these conclusions that we draw, for which part of the world are they valid; men or women, rich or poor, young or old? – the parameters are numerous. The limit is set by the representativity of data; results can not be considered, with sufficiently high confidence, to be representative for more than the original text was; if there were few women in the original text, then the results cannot fully reflect women.

Generally, we can see a trade-off between reduction and representativity which we illustrate in Fig. 2. We rarely start with full representativity, and we rarely end with full reduction (or we would not be left with anything at all). The relation between the two need not be linear; representativity does not have to decline at the same rate as the amount of information is reduced. Good text mining methods attempt to reduce information with a lower loss, similar to compression methods of images or videos; we can reduce the amount of information

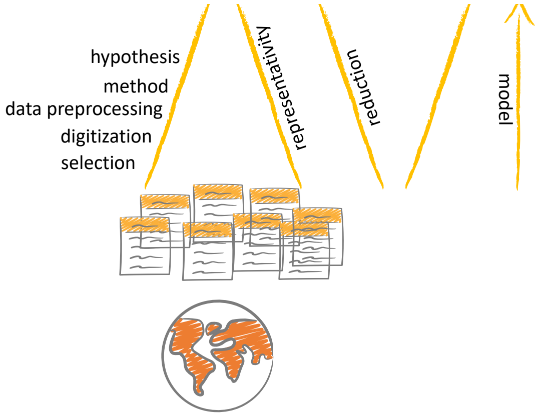

Fig. 2. A schematic model of the relation between representativity and reduction in data-intensive humanities research.

that is being sent, while keeping a fairly high quality of color and movement. While improvements are made at a quick pace, current data science techniques are almost always associated with loss of representativity.

This loss of representativity means that our results are a small window out of which we view our large-scale and possibly long-term text. Our window gives us access to incomplete pictures of the view, and different positioning of the window will result in different views. The image that can be viewed corresponds to the text(s) that we have chosen as our basis. The different positioning of windows corresponds to the method and the preprocessing that has been chosen. Creating topic models using nouns or verbs, using 10 or 100 topics, looking at the first 10 or 500 most likely words, are all choices that result in very different windows out of which we view the same text that affect the conclusions that we draw.

In the mathematical world, this insight is well known. Assume that we want to find the highest point on (or the value that maximizes) a curve, for example corresponding to profit. The curve has two peaks, the right one is higher than the left one, and in between there is a low point. If we start at the leftmost end of the curve, and walk upwards as long as there is an upwards, we will end up in the peak that is the lowest of the two (a local optimum). If our criterion for continuing is that we have to keep moving upwards, this criterion is met at the peak, and we will stop walking. With a limited window size, we will not be able to see the other, higher peak. If we instead start in the right end of the curve, using the exact same methodology, we will end up in the higher of the two peaks (a global optimum). Again, without knowing that we have found the highest peak, because of our limited view out of the window. If we look through the first window, we see the local optimum and might draw one set of conclusions. If we instead look through the second window, we might draw completely different conclusions (for example, that the benefits are high enough to continue with the development of a product, or not).

5 Reduction and Validation

Any scholarly approach demands reduction in order to be precise, and in the humanities, assumptions about society and mankind traditionally govern these perspectives in more conspicuous ways than in the hard sciences. This has considerable implications when new kinds and volumes of data are introduced and historical transformations are modeled. In particular, the question of representativity and validation is affected by the question of quantitative and qualitative methods.

For a critical example of methodological issues inherent in a traditional humanities approach, we can turn to the study of film genres. In recent years there has been a revisionist trend in film genre studies. Commentators have argued that many traditional accounts of popular film genres are inaccurate and that the research has been driven primarily by critical and theoretical agendas, rather than by a commitment to wide historical analysis. Not only have scholars all too often limited their interest to a handful canonized classics or works by well-known filmmakers, but they have also tended to substitute "assumptions and generalizations for detailed empirical research" (Neale 2000: [16]). Many traditional accounts of the historical transformation of film genres have not been grounded in evidence drawn from representative sampling, but either on "bald assertions or too invidious comparisons between a couple of titles [...] selected specifically to illustrate the assertion" (Gallagher 2012: 299–300 [7]). Although the example is drawn from a specific field, film genre scholars have hardly been alone in their treatment of a limited number of "classics" as representative for, for example, broad historical transformations, when the texts matched more or less arbitrary paradigms.

A similar situation can be found within literature studies. Here, Franco Moretti's proclamation about distant reading as a way of reinterpreting literary history can serve as an illustration of both the enrichments and the complications induced by methods in which the empirical material is analysed through various quantitative methods (summarized in Moretti 2013 [14]). Moretti's point that distant reading serves as a way to substitute literary historiography based on a narrow canon with broader understandings of a very wide range of books has been rightly influential: we very much need to free ourselves from previous understandings. Still, the concept of machine-aided reading needs developing, as do the practices of it. Distant reading is per definition quantitative, and while that offers many new possibilities, there are distinct limitations to it. Generally, to assess quantitative results, the data they are produced from have to be distinctly representative. If they are not, or their representativeness is not clearly defined, they risk being misleading.

In the same vein, Matthew Jockers, a forceful advocate of quantitative literary analysis consequently delimits studies to the quantitative in his Macroanalysis (Jockers 2013 [9]). This approach is fruitful, but has attracted criticism not least as the data has been little representative with regards to such aspects as gender, race, etc. (Nowviskie 2012 [17]; Bode 2017 [6]; Bergenmar and Leppänen 2018 [4]). Pointing at a number of problems concerning selection of and accounting for material, Katherine Bode concludes that this way of modeling literary history is reductive and a-historical (2017, 79). Bode's major point is that this kind of distant reading actually repeats and reinforces the old, close reading notion of texts as static and "the source of all meaning", neglecting the dynamics of texts and contexts in evolving meaning, circumstances such as different interpretations, changes of wording and understanding of the text over time. Instead, departing from Jerome McGann (2014 [13]), Bode suggests a way toward big literary data inspired by the scholarly, contextualizing and in a number of ways non-reductive edition (91–2). The transferring of models from editorial projects certainly is of the essence (cf. Leonard 2016), but the question remains how to handle really large data sets without curating them to the extent that the project becomes untenable.

As Ted Underwood comments (2016 [21]), "It's okay to simplify the world in order to investigate a specific question. That's what smart qualitative scholars do themselves, when they're not busy giving impractical advice to their quantitative friends." Reduction is of the essence: the question is how to handle it.[3] One way of addressing that problem is very strict care and description concerning selection and representativity of empirical data – another way is to let the quantitative analysis pave the way for qualitative analysis without being overwhelmed by restrictions and deliberations. Using another kind of topic modeling than Jockers, Peter Leonard and Timothy Tangherlini have instead described a way to use quantitative methods in order to land at fruitful qualitative investigation. By starting with a "sub-corpus" the researcher is able to apply her/his domain expertise to govern the examination. The quantitative modeling and mapping of "topics", in this case, then provides the means for tracing patterns through large materials, and then going directly to the sources for detailed, qualitative analysis (Tangherlini & Leonard 2013 [20]).

There are more ways than this to achieve a productive convergence of quantitative and qualitative methods, but we may conclude that this kind of approach has at least two distinct benefits. On the one hand, it connects in a very natural way to traditional humanities, as it provides powerful ways to prepare qualitative studies with quantitative tools. On the other hand, it alleviates the demand for representativity, as the quantitative results need not claim to be exhaustive, but rather lead to careful explication of patterns, structures, lines of reasoning and the like – and as in traditional humanities, such explication can be made without strict demands of representativity. So, the proposition is that traditional micro levels cannot only be replaced by macro levels, but that meso levels are also needed, and above all, that it must be possible to freely move between the levels and between the quantitative and the qualitative: adaptive reading. It is here that the humanities can offer the sensibility of the vague and elusive cultural contextualization. Machine-aided reading also makes it possible to explore the dynamics of literature in other ways. Studying only the first edition of a work is fruitful in many ways, but we may now trace the changes of a work through all

[3] An entirely different, highly interesting exploration of reduction as a method is presented by Allison (2018) [1].

its editions: a wonderful opportunity to map the fluctuating character of the literary text. We may also, as the tools get more refined, trace the changes a work undergoes as it is translated into other languages: was it expanded, abbreviated, changed ideologically or aesthetically, or in other ways adapted for its new intended culture and readers? Even though there are a number of risks involved in digital humanities, as outlined above, it appears that these issues have the potential of truly contributing to the methodology of traditional humanities.

Somewhat paradoxically, questions of completeness and representativity are often (and rightly) leveled in criticism of digital humanities, but only rarely discussed in traditional studies. There is a distinct need to clarify how collection, preprocessing and exploration have been performed in digital humanities studies – but there is also a distinct need for the corresponding methodological accuracy in a great number of traditional studies. One reason that this has been neglected is the acknowledging that a traditional, qualitative study cannot be entirely representative when the material is too large – however, the traditional approach has also been that of course the study cannot claim to be entirely complete, but it can claim to show distinct developments and tendencies at the same time as there may be other examples that do not correspond. On the one hand, this is precisely the reason why qualitative digital humanities more easily can handle the problem of representativity and completeness. On the other hand, there is a need for concise clarification of how a study was performed, in order for it to be properly assessed. Here, we believe that digital humanities have the power to set fruitful examples and incite methodological development even in those parts of the humanities which do not focus on digital materials and methods.

6 Conclusions

With this paper, we hope to further a discussion of finding good ways of employing a data-intensive research methodology for the humanities. We have highlighted the differences and commonalities between the quantitative and qualitative in relation to a data-intensive research process.

Even though every project has its own conditions and characteristics, we argue that the data-intensive digital humanities must focus on representativeness and reduction in all phases of the process; from the status of texts as such, over their digitization to pre-processing and methodological exploration. Because the aim is to generate sustainable knowledge in the humanities, special attention must be given to the interpretation of hypotheses with respect to a research question. The merger of quantitative and qualitative methods emphasizes the need to validate on different levels of the research process and to maintain transparency. We conclude that the methodological convergence between the humanities and data science has the potential to raise methodological awareness in the more traditional and non-digital humanities, which only rarely or to a limited extent deals with questions of, among other things, reduction and representativity.

References

[1] Allison, S., Allison, S.: Reductive Reading: A Syntax of Victorian Moralizing. Johns Hopkins University Press (2018)

[2] Bailer-Jones, D.M.: When scientific models represent. International Studies in the Philosophy of Science **17**(1), 59–74 (2003)

[3] Bailer-Jones, D.M.: Scientific models in philosophy of science. University of Pittsburgh Pre (2009)

[4] Bergenmar, J., Leppänen, K.: Gender and vernaculars in digital humanities and world literature. NORA-Nordic Journal of Feminist and Gender Research **25**(4), 232–246 (2017)

[5] Berry, D.M.: Introduction: Understanding the digital humanities. In: Understanding digital humanities, pp. 1–20. Springer (2012)

[6] Bode, K.: The equivalence of "close" and "distant" reading; or, toward a new object for data-rich literary history. Modern Language Quarterly **78**(1), 77–106 (2017)

[7] Gallagher, T., Grant, B.: Shoot-out at the Genre Corral: Problems in the "evolution" of the Western

[8] Google: N-gram viewer, men and women. `https://books.google.com/ngrams/graph?content=woman_INF%2Cman_INF&year_start=1800&year_end=2000&corpus=15&smoothing=3&share=&direct_url=t3%3B%2Cwoman_INF%3B%2Cc0%3B%2Cs0%3B%3Bwomen%3B%2Cc0%3B%3Bwoman%3B%2Cc0%3B%3Bwomans%3B%2Cc0%3B%3Bwomaned%3B%2Cc0%3B%3Bwomaning%3B%2Cc0%3B.t3%3B%2Cman_INF%3B%2Cc0%3B%2Cs0%3B%3Bman%3B%2Cc0%3B%3Bmen%3B%2Cc0%3B%3Bmanned%3B%2Cc0%3B%3Bmans%3B%2Cc0%3B%3Bmanning%3B%2Cc0`, accessed: 2018-10-19

[9] Jockers, M.L.: Macroanalysis: Digital methods and literary history. University of Illinois Press (2013)

[10] Kitchin, R.: The Data Revolution: Big Data, Open Data, Data Infrastructures and Their Consequences. SAGE (2014)

[11] Kjørup, S.: Människovetenskaperna: problem och traditioner i humanioras vetenskapsteori. Studentlitteratur (1999)

[12] McCarty, W.: Modeling: a study in words and meanings. A companion to digital humanities pp. 254–270 (2004)

[13] McGann, J.: A new republic of letters. Harvard University Press (2014)

[14] Moretti, F.: Distant reading. Verso Books (2013)

[15] Morrison, M., Morgan, M.S.: Models as mediating instruments. Models as Mediators: Perspectives on Natural and Social Science

[16] Neale, S.: Genre and Hollywood. Genre and Hollywood, Routledge (2000)

[17] Nowviskie, B.: What do girls dig? Debates in the Digital Humanities: University of Minnesota Press (2012)

[18] Rockwell, G., Sinclair, S.: Hermeneutica: Computer-assisted interpretation in the humanities. MIT Press (2016)

[19] SCB: Undersökningarna av levnadsförhållanden (ulf/silc). `http://www.scb.se/hitta-statistik/statistik-efter-amne/levnadsforhallanden/levnadsforhallanden/undersokningarna-av-levnadsforhallanden-ulf-silc/`, accessed: 2018-10-19

[20] Tangherlini, T.R., Leonard, P.: Trawling in the sea of the great unread: Sub-corpus topic modeling and humanities research. Poetics **41**(6), 725–749 (2013)

[21] Underwood, T.: The real problem with distant reading. Blog post, https://tedunderwood.com/2016/05/29/the-real-problem-with-distant-reading (2016)