CA Journal of Cultural Analytics

Bischof, Andreas, and Konstantin Freybe. "Grounding Theory in Digital Data: A Methodological Approach for a Reflective Procedural Framework." *Journal of Cultural Analytics*, vol. 7, no. 4, Dec. 2022, https://doi.org/10.22148/001c.57197.

ARTICLE

Grounding Theory in Digital Data: A Methodological Approach for a Reflective Procedural Framework

Andreas Bischof[1], Konstantin Freybe[2]

[1] Chemnitz University of Technology, [2] Leipzig University

Keywords: DH theory, methodology, digital data, case study

https://doi.org/10.22148/001c.57197

Journal of Cultural Analytics

Vol. 7, Issue 4, 2022

Instead of looking for new paradigms for Digital Humanities (DH), we present Grounded Theory Methodology (GTM) as a methodological approach to frame digital research practices more reflectively. By turning to the epistemological and practical implications of digital tools like Topic Modeling and digital data sources like YouTube comments, we highlight the theoretical assumptions that are already in the game—and call for more explicitness and methodical monitoring. To explain the procedures of GTM and the proposed worth for DH, we present an example of a qualitative research project using machine learning techniques to narrow down a large scale of data to human interpretable resample. The methodically monitored resampling process provided valuable means to validly minimize the amount of data without losing a qualitative trajectory of the process itself. Defining and tracing *relevant* content in our original data set enabled us to find related comments and textual conversations to be analyzed further. We discuss the example iteration in two ways: Our prototype and procedure show on the one hand, how qualitative research and computational methods can be better intertwined without compromising their epistemological foundations. On the other hand, we argue for an understanding of DH as research practice, that should follow an abductive research agenda in order to ground its theories in data.

1. Introduction

With the advent of statistical data collection and mathematical analysis of social phenomena in the early 19th century—long before computers and the internet—a discourse began in the social sciences and humanities about the discovery of theoretical knowledge from data (Jahoda). This discourse is not uniform and rigorous but rather recurs in waves, especially when new procedures or data, or even social crises fuel it. The availability of voluntarily or involuntarily generated digital data on almost every area of human activity is a "crisis" of this kind, and it poses fundamental challenges for the sciences. Manovich's concept of cultural analytics has developed an ambitious agenda in response to this challenge. This agenda operates at the intersection of diverging scientific cultures that require new ways of dealing with digital data and research tools. In this article, we want to focus on a key aspect of this agenda from an epistemological and methodological perspective: "How can

we combine computational analysis and visualization of large cultural data with qualitative methods, including 'close reading'?" (Manovich, "Science of Culture" 2).

This historical contextualization raises the suspicion that an assumed "end of theory" cannot help us to address questions of the epistemic foundations and theorizing capacity of digital humanities (DH). Therefore, we want to ground the discourse of theorylessness *methodologically* in this article. This means we aim to understand digital research practices in DH as a complex epistemic interplay of humans and technology and different logical operations, namely induction, deduction, and abduction.

The notion of the "end of theory" is not an adequate account of the theoretical landscape because it obscures the theory and decisions that are already there. The talk of discovering theories in data always presupposes theories, incorporated in technical tools and the digital data themselves. The idea of successfully understanding and modeling digital traces of human behavior without a priori theories is itself a presuppositional—and methodologically myopic—theory. We illustrate this interplay by demonstrating the epistemic implications of two key sources of digital research practices: the implicit theories in technology and data (2.).

From this point of view, the notion of theorylessness becomes rather a question of *how to* deal with the multiple and conflicting theoretical implications that digital research practices already bring to the table. Our core argument is that we do not need a new epistemology to deal with this challenge. Instead of new theories, we argue for a reflective and methodologically probing theory-generating endeavor that is grounded empirically. The grounded theory methodology (GTM) offers a reflective procedural framework to carefully inspect the entanglement between research practice, methods, and data. GTM acknowledges that building models from data is a long, iterative journey and seeks to guide researchers along the way. By turning to GTM, we want to show that existing strategies to deal with messy data and researchers' implicit decisions are valuable for understanding and avoiding a lack of theory when conducting digital research in DH (3.1).

In order to ground our argument in practical research experience, we provide a short example that allows us to highlight the opportunities and limitations of our approach. We present a section of a larger study in which a large quantity of digital data was rendered usable for the purposes of qualitative research via machine learning tools. For this study, we used existing open software tools to conduct computational analyzes of a text corpus of 1880 automatically generated captions from YouTube videos. More specifically, we explain how computational topic modeling of natural language was used to identify *relevant* data for our research question. Our example focuses on how human researchers can use computational methods in a methodologically reflective way to identify relevance in natural language data. We argue that

this method of *making sense* of such data is no less rigorous and applicable than computational methods—and accordingly must also become part of the documentation and evaluation (3.2).

Finally, we discuss our findings on grounded theorizing in digital research practices in two respects. On the one hand, we frame our approach as hand-over between qualitative research tradition and new computational methods. We argue that those handover situations are specifically critical for the methodological quality and overall soundness of digital research practices as scientific endeavors (4.1). On the other hand, we turn to the epistemic implications of GTM for researching digital phenomena: In contrast to understandings where qualitative and quantitative modes of knowledge are presented as mutually exclusive, we show how both can be justified within GTM as part of an abductive attitude leading to the generation of abstract theory from data (4.2).

Our article consequently shows that a turn to the *practice* of digital research productively addresses two problematic aspects in the context of theory(lessness) in DH: First, looking at the implications of digital-research technologies and data, it is clear that many assumptions are already embedded in a DH project before it has even begun. Second, GTM gives us a procedural framework that renders these implications practically manageable and even puts epistemic operations typically presented as opposites into a procedural framework. We believe that this perspective can bridge the discourse on DH theory (or the lack thereof) and DH methods and tools, informed by the tradition of qualitative research methodologies (5.).

2. The Implicit Theories of Digital Methods & Data

From a methodological point of view, the notion of theorylessness is hard to accept. Any act of positioning in a research field, any data practice, and any question asked of a data set involves theory. Of course, this does not have to be an elaborated cultural theory every time, but any form of human access to a digital research tool or digitized data implies an interest in or a reason for doing so: We look into the data to discover something. In this section, we want to move on from such misleading contrasts between theory and data.

To question and explore the notion of "theorylessness", we want to highlight two sources of implicit theoretical assumptions within digital research practices: technology and data. First, we address DH as a socio-technical trading zone. We thereby highlight the fact that the different paradigms and theories within DH unite around digital research practices that rely on technologies—which come with their own implications (2.1). Second, we problematize the reductionist juxtaposition of theory and empirical data by revisiting a less discussed recent wave of the "end of theory" discourse in the social sciences. It becomes clear that data is neither raw nor objective but rather relative to its conditions of production and use (2.2).

Both findings show that instead of a lack of theory, there is a surfeit of theoretical assumptions within digital research practices before the actual research has even started. Our conclusion is that reliable research practices in DH have to acknowledge and apply methodologies to account for such influences while still grounding their theories in digital data.

2.1. Digital Humanities as Socio-Technical Trading Zone

The DH cannot be understood as a unified scientific discipline; it is rather an epistemic landscape (Svensson, "Landscape"). DH has been influenced by perspectives from the fields of library and information science, which are typical hosts for DH projects, by humanities subjects like history or cultural studies, and by computer science—which cannot be understood as an epistemically unified field either (e.g., Harrison et al.).

At the end of his series of articles on the DH landscape ten years ago, Svensson (Svensson, "Humanities Project") turned to a concept from the history of science to describe the epistemological implications of this multitude of metatheories and scientific cultures. In analyzing the different paradigms within physics, Galison used the term "trading zones" to explain how scientists can communicate and collaborate on a local level, even if they come from globally conflicting paradigms (e.g., experimentalists versus theorists). By turning to concrete objects, like building and maintaining high-energy test facilities, physicists and their collaborators were able to maintain disciplinary depth and overcome incommensurabilities at the same time.

When applying Falison's concept to DH, the relevant question to ask is: What are the local practices and objects DH scholars care about when seeking to bridge their epistemic differences? Svensson's analysis points to the role of technology and digital(ized) data in DH as a zone of trade, similar to a "boundary object" (Leigh Star) of intersectional inquiry. The "local" practices around which collaborations in DH are built, engage with digital research tools and digital data. Digital technology in DH becomes a "tool, object of inquiry, medium of expression, activist venue and more" (Svensson, "Landscape"). Interestingly, Svensson did not offer a deeper analysis of technology in the epistemic landscape of DH. But his analysis of digital research practices and technology as the campfire around which DH scholars gather has a strong implication for the question of theory and alternative accounts of theory in DH: Theory is already there, incorporated in digital technologies and their use.

Science and technology studies (STS) have long shown how technologies are shaped by cultural influences and entail compelling social consequences—meaning they are themselves social (e.g. Winner). Especially advanced digital technology like information retrieval, natural language processing, or image analysis has a lot of theoretical implications. These are not only of a mathematical nature, like probabilistic models that calculate

the likelihood of following events. Digital research tools, like Voyant Tools, connect input sources to output modalities (like visualizations) in a targeted way to solve problems outside the software—defined by developers and scientific researchers. Furthermore, the input sources and training data sets used by advanced models contain information of (inter-)subjective nature, like annotated data or user generated input. Finally, the seemingly purely statistically generated results of computational methods are then interpreted by humans—following cultural patterns and social dependencies.

Acknowledging how entangled the human practices around digital technologies are with each other, it becomes obvious that we cannot speak of "technology" as a solitary entity but rather of socio-technical systems. Scholars from media and software studies have shown how the typical sources of large-scale digital data on human behavior, like YouTube or other social media platforms, not only incorporate theories about social reality but also powerful mechanisms to influence the part of social reality at which they aim. Rieder has shown how the evaluative metrics behind Google's PageRank are based on and reproduce a notion of authority (Rieder). In a study on the YouTube (YT) ranking algorithm, which often thrives on controversy and dissent, Rieder and colleagues showed how ranking algorithms lead to "ranking cultures" embedded in the meshes of mutually constitutive agencies between computational procedures, user generated content, and patterns of consumption (Rieder et al.). The influence of the computational ranking of YT videos and the consumption of recommended media is a striking example of the power and interdependence of technologies like neural networks, which are trained with user data, and subsequent behavior: Recommendations account for the majority of all video clicks from the YT home page (Davidson et al. 296).

This excursus through STS, media studies, and the implications of technology may not be new or surprising to DH scholars. Yet, our argument is that such implications of digital technologies also apply—on a smaller scale—to the practices of DH research itself. The implications of digital technology often become invisible when they are blackboxed as "tools," especially in research settings (Latour). When we use digital research tools to *discover and fix* results for phenomena yet to be understood, we need to understand and reflect on the epistemic implications of these digital technologies. Hence, DH is not only a trading zone between diverging scientific epistemic cultures; it is also a socio-technical epistemic trading zone between human researchers and their digital technologies.

2.2. Digital Data as Cultural Artifact

A second important reason for questioning of the notion of theorylessness concerns the epistemic qualities of digital data. In order to explain and explore this claim, we want to revisit a recent "end of theory" discourse from sociology. The trigger for this discourse was the availability of "big data,"

coupled with new data analytic methods, which led some authors to conclude (or rather call for) paradigm shifts across multiple disciplines. Although this "end of theory" discourse starts with the disruptive potential of technologies and data, it is more specifically related to empirical social research than Anderson's WIRED article. We date it back to Burrows & Savage's article "The Coming Crisis of Empirical Sociology", which was published one year before Anderson's article and roughly in parallel to Manovich's work problematizing the role and legitimacy of the social researcher (886):

> "[I]n the early 21st century social data is now so routinely gathered and disseminated, and in such myriad ways, that the role of sociologists in generating data is now unclear."

The paper was a polemic—an intervention aimed to alert colleagues (Burrows and Savage). How can sociology get back into the driver's seat and gain a holistic understanding of society when most of the traceable actions disappear onto private servers and into the hands of corporate data scientists? Beside the questions related to social science as a profession—namely the means of gathering data and the access to it—the methodological focus of this discourse is interesting: What is contested here is the role and possibility of explanatory social science itself regarding the different "modality" of digital data. The challenge posed was the proliferation of what the authors called "'social' transactional data"—such as credit card billing information or user data from the web—which were then already routinely collected and combined with public data, such as census data or electoral rolls data and so on, in order to produce highly sophisticated socio-spatial maps that were processed and analyzed by a wide variety of private and public institutions. Whereas most sociological methods rely on accounts of actions, these data might enable what Quetelet was already searching for in the 19th century: a scientific model of human behavior relying on "objective" data itself (Jahoda).

In contrast, digital traces of action—like "big data"—offer quite mundane and routine reporting of numbers, events, places, times, and so on. The question "Where have you been the last 48 hours?" can be answered more accurately by using the GPS data from our smartphones than by asking the persons themselves. But, like the (verbal) accounts of an event given by humans, which give an additional layer of information that the interpretative social sciences are specialized in reconstructing, digital data itself is never raw (Gitelman). Of course, data can be viewed as an abstract entity or tool—but it also has to be understood in terms of its conditions of production and the way that it is used. Data is—even when automatically gathered—an artifact, with its own history, materiality, and purposes. Taken out of context, digital data risks losing its meaning and thereby its explanatory power. As Boyd and Crawford (671) pointed out, there is value to analyzing data abstractions, although context remains critical. Context is hard to interpret at scale and

even harder to maintain when data are reduced to fit into a model. Managing "context" in light of digital research practices is the key methodological challenge here.

To summarize: the availability of digital research tools and digital data has provoked debates about epistemologies and methodologies in almost all fields of science. DH and cultural analytics have emerged as fields that seek to combine the strengths of classical humanities and social science research practices with computational methods. By highlighting the cultural and social implications of both digital technology and data, we argue that digital research practices need to critically reflect on and check the implicit theories and assumptions of tools, methods, and data. We do not consider this analytical ambition to be new to DH or cultural analytics. We rather argue for more explicitness in dealing with this challenge.

While the Anderson discourse on the "end of theory" fueled epistemologically naive promises—for instance research without any a priori assumptions (Kitchin 4)—the discourse following Burrows and Savage calls for a less radical but still data-driven epistemology that modifies the existing scientific method by combining aspects of abduction, induction, and deduction (Kitchin 10). This speaks to a long-standing tradition in qualitative inquiry, namely GTM.

2.3. Grounded Theory Methodology as Procedural Framework to Make Sense of Digital Data

Grounded theory is part of a long-standing tradition within qualitative social research. Although it was formulated 55 years ago, the legacy of grounded theory is relevant to the challenges of the digital research practice presented here in two respects: its understanding of "data" and its goal of data-based theorizing. Glaser and Strauss developed the idea of a grounded theory in the 1960s as a response to sociological grand theories such as Parsons' structural functionalism (Glaser and Strauss; Charmaz). In particular, they criticized that data were most likely to serve as a *test of theory* and instead called for *theories to be discovered from data*. Furthermore, Glaser argued, explanations of social reality must consider data sources beyond statistical data and qualitative interview data. The founders of grounded theory stressed early on that anything can be used as data, including involuntarily provided data, newspaper articles, or information on the role of objects in social interactions—the decisive factor was whether it fits the continuously evolving model of the field structure (Glaser 196).

Although—or because of?—this adequacy to the challenges of DH, GTM has often been misunderstood both in its scope and its theoretical fundamentals. The most common misunderstanding is to use GTM first and only as a data analysis method. In this respect, GT (in that case without M) has gained a dubious reputation as a 'simple' qualitative evaluation

method, which is clearly inferior to the complexity of the actual coding procedure (Bischof and Wohlrab-Sahr). A more advanced misunderstanding is to understand GTM—or parts of it—as a standardized procedure to structure qualitative data. GTM does indeed provide a procedural framework, but even if the initial data and computational analysis methods remain the same, *this framework must be modified and specified* for each project depending on the interest of the researchers and their research questions.

Before we discuss this in our case study more graphically (cf. 3.), we would like to briefly discuss this non-trivial methodological problem. GT(M) has been discussed several times in the large body of DH work, for example in Computational GT (Nelson) proposed adaptation for GT with machine learning and natural language processing. Furthermore, a lot of methodological and conceptual DH projects, like Benardou et al. (2010) or more recently topic modeling projects, like Liu et al. (2017), refer to GTM explicitly and use it, e.g. to build categories within data sets. In the course of this, GTM applications in DH have used process steps similar in detail to those we want to present—but have missed the real strength of using GTM to make sense of digital data of human behavior. GTM is not first and foremost a method of analysis, but a *procedure* anchored in data, reflecting its theoretical preliminaries, that informs and structures an entire DH project—from the definition of the task, through the selection of data, their (partially) automated evaluation, further collection of data, their analysis, etc.

Grounded theory is a methodological practice; hence, we prefer to call it grounded theory *methodology*. GTM was developed to deal with the complex *interplay of processes* of gathering and analyzing data with a view to grounding theories on empirical data. To achieve this, GTM offers both concrete procedural steps for researchers and overarching principles to inform decisions and next steps. GTM is thus not a rigid grid but a *methodological tool* in the truest sense of the word. GTM helps the researcher to always connect the adequacy of the theories to be developed—but also the data to be chosen, the definition of cases, and the choice of evaluation methods as well as generalization schemes—back to the original research interest and the data obtained from social reality.

The epistemological core of GTM as a procedural framework is a recursive balance of moments of inductive, deductive, and abductive reasoning, which is reflected in the guiding principles and procedural steps. The balance between inductive and deductive reasoning is mainly achieved through an interplay of generating hypotheses from the material and further testing and refinement of these hypotheses through several steps of coding. The goal is to test these hypotheses rigorously by contrasting cases. Therefore, it is inaccurate—and indeed infuriating—when critics claim that research

using GTM does not require prior knowledge or even explicitly excludes it. GTM requires exactly the opposite: It requires at least one empirically and theoretically supported assumption—an "educated guess" (Peirce 241–56)—which is to be shaped and tested in the process of research.

Two main principles of GTM are especially useful for empirically grounding open-ended research questions with digital data, as in DH: the principle of constant comparison (which aims to guide the sampling process), and the iterative and constant alternating between the analysis itself and the (re-)conceptualization of data gathering. These two principles help to guide the researchers from their initial interest to empirically grounded decisions on how to narrow down the data in terms of its scale and messiness. A grounded research agenda as a framework thus features two key elements:

- An open **data sampling strategy** that has the capacity to unsettle the assumptions and interests of the researchers—and those implicated in different digital data sources and analytical tools—by continuously contrasting cases and units of analysis (principle of constant comparison).
- A **continuous sequence of inductive and deductive steps**, whereby data collection and hypothesis generation (inductive) is alternated with new, theory-driven data collection based on these hypotheses (deductive) and corresponding testing and elaboration of theoretical concepts. Thus, there is the need for a constant (re)integration of new knowledge within the initial research interest and project goals that allows for adjustments and explicit reflection on its implications.

GTM constantly relates three elements we have developed as core sources for implicit theoretical assumptions in digital research practices in this section. Accordingly, we propose that digital research practice answers these questions explicitly and probes the answers within its procedures and reports with an appropriate methodology.

- We look into the data to discover something—what is our interest?
- We use digital tools to create order in messy data—what are the methodological implications of our methods?
- We rely on digital data from human practices—what is the context and meaning of those data?

To address these crucial questions, DH does in our view *not* need a new epistemology but rather a reflective procedural framework to probe the entanglement between research practice, technology, and data, which GTM provides.

3. Grounding Theory in Digital Data – An Example Iteration

To discuss a process of grounding theory in digital data with the appropriate methodological checks and the iterative character of this process, we turn to an example from our own research practice that aims to explain the interactions of gaming influencers and gamers on YT. At its core, the example presented here concerns the interplay between selecting data sets and using machine-learning technique topic modeling to identify and trace relevant data in our original research data set.

In presenting our case study, we highlight the constellations of handovers and interplay between the qualitative and computational methods. Our example originates from a PhD project that used GTM to investigate the cultural practices of gaming influencers with regard to the tension between authentic self-fulfillment, self-commodification, and consumerism. To explore these topics in the empirical data, we conducted a field assessment in which we reconstructed a *model of the field structure*. This model encompassed relevant types of actors, relations, and practices within the phenomenon to be explained. This model was then used to identify potential cases and data sources, which prompted the decision to use YT as the primary source (3.1).

To cope with the vast amounts of data, we needed to develop relevance criteria that guided the composition of our data sample. We also needed to identify adequacy criteria to ensure that we had a sample that was fit for computational methods—and still captured the context and meaning of the data within the interactions between gamers and gaming influencers, which were the subject of the research. Our sample structure reflected this: We defined criteria for cases to create a resample for further investigation (3.2). Before running a topic modeling analysis on this subset of data, we furthermore created a simple data model to incorporate the criteria derived from the model of the field structure (3.3). As a result, we listed possibly meaningful connections between terms that needed further human interpretation in order to prepare for an actual qualitative analysis, such as close reading or qualitative coding (3.4). To describe this specific iterative process of accessing the field, creating a model for the field structure, deciding on a data sample, and undertaking a computationally aided search for relevant cases within the sample, we propose the term "resampling".

In the context of the overall research project, our example serves several purposes. It helps us to identify relevant data in a given corpus of text and to narrow down the scope of resampled data. We therefore finally place the work on the technical prototype and the iterative research practice for this resampling in the context of the overall project (3.5).

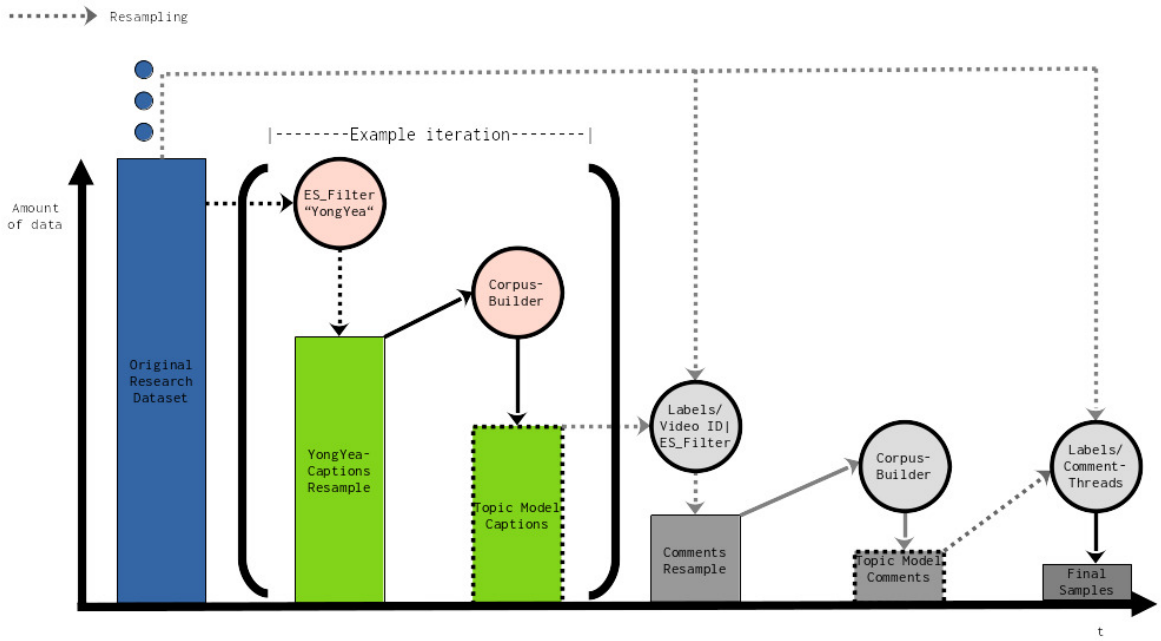

Figure 1. Resampling Strategy in Context of Overall Data Gathering. (own graphic)

3.1. Example Iteration: Identifying Relevant Data

The main challenge of the iteration we present here in detail was to identify data that was relevant to our research—that is, to reduce the quantity without losing the quality of the data. We built a working prototype of a digital research tool for topic modeling and explored the compatibility between its use and the procedural framework of GTM.

Our example pertains to an individual iteration of continuous inductive and deductive steps within a larger research process (figure 1). We will refer to our efforts to establish a procedural interconnection between the overall research interest and the requirements of individual substeps as *resampling*. The basic idea is to delegate specific tasks to adequate computational tools and use customized derivatives of the original research data set. This gave us sufficient flexibility to meet the requirements of various research procedures, both computational and interpretational. The main objective was to narrow down the volume of resampled data until we reached levels where qualitative research tools can reasonably be applied. In this case, the task was to prove a concept: that topic modeling guided by GTM criteria could help us identify relevant data for further resampling. The following sections will describe the methodological steps we followed in order to structure (3.2.2), assemble (3.2.3), and analyze the data (3.2.4).

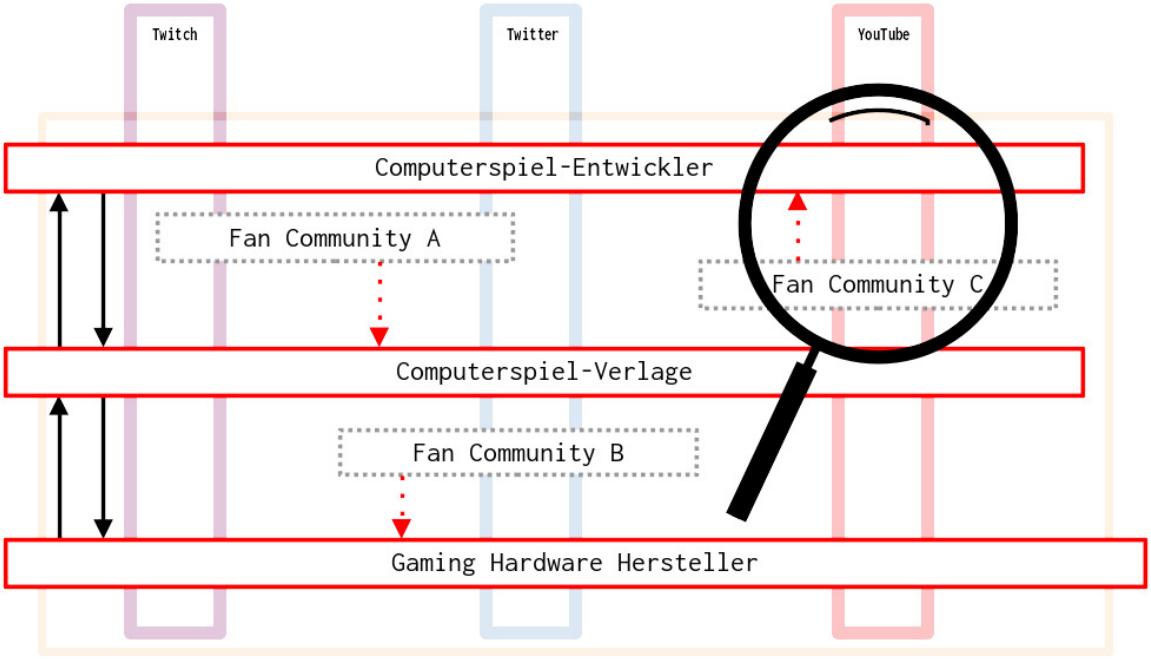

Figure 2. Generalized Field Structure Model of Gaming Industry, Gaming Influencers & Fan Communities (own graphic)

MODELING FIELD STRUCTURE

Field assessment was a crucial and foundational step for the whole research project and for the example we provide here. Within GTM, field assessment served to reconstruct a model of the field structure. This model encompassed relevant types of actors, relations, and practices within the phenomenon to be explained (*figure 2*). This model was then used to identify potential cases and data sources and prompted us to use YouTube as the primary source. At this stage, the model was the result of our intellectual rather than computational efforts. It started at YT's front end as it appears to human actors in social reality. Until this point, the initial field assessment resembled ethnographies of online behavior (Kozinets).

From there, we began preparing to approach the process from the back end of these online practices. To cope with the vast amounts of data—videos from YT and Twitch, comments, tweets, and message board threads, etc.—we further needed to develop qualitative relevance criteria that guided the composition of our data sample. We also needed to identify adequacy criteria to ensure that our sample was fit for computational methods and still captured the context and meaning of the data within the gaming influencer-gamer interactions we wished to research.

GTM's iterative design promotes frequent evaluation and re-evaluation during the research process. Informed by our theoretical model of the field structure, our strategy was to create data snapshots of selected channels, ranging from computer games publishers, developers, journalist media, and

fans, and assemble an original research data set. Our original data set was 5.4 gigabytes in size, consisted of data from 49 different channels, and covered a total of 24,016,086 comments linked to a total of 71,508 videos. The data for this original data set was queried from YT's API in March 2019. The data was stored in an ElasticSearch instance that was run locally in a Docker container. We received data from YT in JSON format which was ingested into ElasticSearch by running a custom script. ElasticSearch in conjunction with Kibana, a browser-based interface, allowed for convenient exploration, visualization, and exporting of our data.

Three core criteria guided our choice to resample data on the channel "YongYea" from our original data set. The channel:

- was conducted by a computer games enthusiast, indicated by its brand communication, e.g. through profile picture or video backgrounds, that make strong references to games;
- developed a sense of mission. Content programming was diverse and shifted from gameplay oriented videos to critical news reporting on the gaming industry;
- produced sufficient amounts of data by frequently uploading content and through the active community of commenting viewers.

In short: We were certain that this is the kind of influencer we were looking for and had access to sufficient amounts of data for this iteration. However, before we were able to assemble data, we needed to check our resample for the distribution of our targeted data type: automatically generated captions of the video audios.

3.2. Negotiating Resample Structure

Choosing a data source for this iteration was guided by relevance criteria developed during the field assessment. We needed to discover cases in the data that would help us to answer our research question. Here, we were most interested in text analysis and therefore we targeted data properties that referred to this data type. We found captions to be the most expressive accounts of spoken YT content.

We realized two things when modeling resampled data for this iteration. First, the task of generating machine-readable data pushed us to the outer edges of our domain-specific expertise. Second, the topic modeling tools by themselves were not concerned with our broader research contexts and did not maintain context and links to other data for us. Therefore, it was our responsibility to develop a slim, yet expressive data model.

We decided to include two metadata properties in addition to the caption text: First, we added the identifier of the respective video to make sure that we were able trace this ID by computational means. Second, we included the title of the video to provide human-readable data. This served as an opportunity for us to—at least preliminarily—classify the video content. After we negotiated the sample structure, we moved on to assemble the data for resampling into a form that topic modeling tools could digest.

3.3. Resampling Composition

Our model of the field structure informed the composition of our original data set. In order to create an excerpt from our source material for this example, we had to balance the demands of maintaining the human readability of the text and creating machine-readable data that caters to requirements of topic modeling. After creating a model for our sample data, we needed to apply the model to our actual resample for this iteration. We examined the distribution of the sample caption data to test the suitability of this data for topic modeling.

The primary language present in the caption texts was English. From our target channel, we had data on 2,276 videos, but only 1,880 of these videos had automatically generated captions attached to them. In most cases, we found automatically generated captions that showed some degree of normalization already. All characters were lowercase, and often there was no punctuation. We chose to define all caption text that belonged to an individual video ID as a document. Without punctuation in the caption text, we needed to work on a word-level basis because sentences could not be detected automatically.

When preparing the topic modeling, we decided to perform minimal preprocessing on the text data. We found that a too rigid cleaning of the raw text may do more harm than good because we needed to account for the indexicality of fan communication. After we removed special characters and transposed words in our corpus to (computable) tokens, we decided to use only a basic stopword list to remove highly frequent terms (“a”, “the”, etc.) from the corpus. We needed to balance text normalization and preservation of specific traits of fan communication. Our aim was to preserve as much of the fan-specific language as possible in our data. This meant that we kept frequent terms like “game” or “play” for their significance in gaming culture. Terms like “snake”, “ocelot”, “wolf”, and more are likely not references to animals but to fictional characters from the games. Next, we chose to lemmatize (spaCy) the words contained in our caption documents. This procedure resets the terms to their base form, which further promoted text normalization. In contrast to lemmatization, stemming our terms, which basically refers to normalizing word forms by cropping them, seemed too rough a procedure. Lemmatization concluded our preparations for topic modeling.

Next, we handed over our corpus to the tool prototype we built. This constituted a critical handover constellation in our research process. On the one hand, we had to ensure that our example fit our research interest. On the other hand, we had to conform to technological standards to successfully apply topic modeling. Both processes by themselves are quite complex and delicate. When seeking to generate interpretable results that matched the ambition to generate more abstract theory from them, it was important to think of them as conjoined rather than mutually exclusive.

3.4. Evaluating the Results and Preparing the Next Iteration

It was challenging to evaluate the output from our topic modeling because we had to prepare data that conformed to both our qualitative research requirements and the technical standards. We handed over a processed resample from our original research data set. Since we built our custom software in Python, the Gensim library provided us with the means to produce an LDA-based language model based on our data. This model determined topics probabilistically and clustered words into topics according to their frequency of occurrence in the corpus (Blei et al.). The results of these computations were then transposed back to words. The raw output took the form of eight lists of terms plus a numerical value for each term. The number of words included was set to 10. Each list of 10 term-value pairs constituted a topic found by our language model:

- Topic 0: look (0.022), let (0.020), thing (0.017), really (0.017), guy (0.017), kind (0.012), talk (0.011), time (0.010), play (0.010), well (0.010)
- Topic 1: snake (0.025), music (0.025), mission (0.020), enemy (0.014), weapon (0.012), shot (0.011), attack (0.010), move (0.009), find (0.008), leave (0.008)
- Topic 2: aggression (0.006), 않으면 (0.004), wuld (0.001), 시간 (0.001), 블레이드단의 (0.001), slen (0.001), 정보를 (0.001), 리프튼에 (0.001), 도와드릴까요 (0.001), 어딘가로 (0.001)
- Topic 3: trailer (0.022), character (0.020), new (0.015), show (0.014), feature (0.011), world (0.011), look (0.011), time (0.010), young (0.010), information (0.008)
- Topic 4: switch (0.012), game (0.009), money (0.009), way (0.007), sell (0.007), console (0.007), content (0.006), thing (0.006), community (0.005), pay (0.005)
- Topic 5: player (0.122), system (0.032), game (0.029), item (0.025), destiny (0.021), level (0.020), mod (0.014), feature (0.013), weapon (0.013), shader (0.012)

- Topic 6: wuld (0.000), 블레이드단의 (0.000), 시간 (0.000), 到着時刻は (0.000), 도와드릴까요 (0.000), 알두인은 (0.000), 어딘가로 (0.000), 하지 (0.000), 리프튼에 (0.000), 정보를 0.000
- Topic 7: game (0.085), time (0.015), new (0.012), year (0.011), work (0.010), much (0.008), young (0.008), release (0.008), company (0.007), people (0.007)

We quickly realized that the ability to make sense of this kind of output depended on its presentation—and the basic output was simply too cumbersome to handle. After being handed the results, we needed to decide on how to visualize the output. We produced separate image files over the course of our exemplary research iteration. For future research, it seems helpful to organize such images into a dashboard-style overview.

We visualized general information—like the distribution of word counts over the documents—in a histogram. We found that modeling eight topics was most expressive—again, this choice was impacted and informed by knowledge from the field assessment. We used histograms to visualize how many documents containing how much text shared a dominant topic. This helped us contextualize topics according to the share of data that was labeled with a topic. To make sense of the terms contained in each topic, we used word clouds that displayed the 10 words (from each topic) that were most frequent in the topic. LDAvis was the last and maybe most complex visualization at our disposal. However, it may also have been one of the most convenient to set up.

LDAvis generates dynamic HTML files from the topic model. It presents two interactive visualizations. On the left, there is a two-dimensional coordinate system on which each topic is represented as a circle. A circle's diameter represents how dominant the topic in question is in the data. The distance between the center points of the circles indicates how similar the topics are (from a computational point of view). On the right, we are presented with a stacked bar chart that represents the terms that belong to a topic and lists the terms according to their local and global frequency and a mathematical relevance measure introduced by Sievert and Shirley.

The last step was to label the data so that we knew which topic was dominant in which document (caption from video ID). We used K-means clustering to label our documents. Due to the field knowledge we had acquired, we were able to assign the following labels to our eight topics (see above):

- label_name0: game & play[1]

1 The digits indicate the corresponding topic number.

- label_name1: MGS
- label_name2: junk topic1
- label_name3: anticipating game releases
- label_name4: community-industry relations
- label_name5: game-player relations
- label_name6: junk topic 2
- label_name7: industry controversies

We identified Topic 1 as the most relevant for us. A total of 399 documents were labeled with the topic "MGS"[2]. Therefore, we achieved our main objective, namely, to narrow down the sample size. These materials enabled us to sufficiently identify several clusters that matched our relevance criteria.

We labeled topics 2 and 6 as junk topics. This is, in part, a workaround to deal with the existence of multiple languages in our data. As our sample presentation of the topics found (see above), these topics contain Korean and Japanese characters. Junk topics became a helpful—albeit *quick and dirty*—way to contain the issue of multilinguality as it enabled us to maintain our research focus.

Our example concluded with a csv file containing URLs, IDs, labels, and titles of all videos that belonged to the "MGS" cluster. This file was the result of a methodologically controlled sequence of steps that began with an assessment of our field of research, which led us to a model of the field structure. This model informed the composition of our original research data set from which we generated an excerpt that was customized for topic modeling. This resample had to be modeled differently from our original data set before we could feed it into our topic modeling setup. Visualizing the output was necessary to enable us to qualify topics within our superordinate research project and constitutes another critical handover situation. We investigated whether there were indexical terms contained in the topic that we already encountered during the field assessment. In other words: we used our knowledge on the field to semantically recharge the terms that form topics and recontextualize them.

Topic modeling enabled us to compose a list of video content relevant for our research question. This list became a valuable asset as it was designed to be reused in subsequent iterations. By feeding our list of relevant video IDs into

2 MGS stands for *Metal Gear Solid* which is the name of a popular Japanese computer game franchise, owned and published by Konami Digital Entertainment and created by game designer Hideo Kojima. This series is at the core of the superordinate PhD project. Yet, we chose not to go into the details in order not to divert attention from our methodological argument.

a script that generates a query in Kibana Query Language format we expect to be able to filter for relevant comment data in our original research data set. This was an important step towards an analysis of conversations in comment sections associated with relevant video content.

3.5. Resampling as Part of an Overarching Research Process

In the context of our overarching research project, this example of resampling provided valuable means to validly minimize the amount of data without losing a qualitative trajectory of the process itself. Defining and tracing *relevant* video content in our original data set enabled us to find related comments and textual conversations among YT users in the comment sections below the videos.

Defining and tracing relevant video content was an important step towards generating *cases* that can be analyzed and compared, as GTM requires. The principle of theoretical sampling (see 3.1) means cases are selected that are assumed to be similar to the case already evaluated (minimum contrast). Finer differences in the concepts found are to be worked out and illuminated more closely. The selection of cases that are in maximum contrast to the previous ones, on the other hand, aims more at the expansion of found concepts and the questioning of previous findings.

Our experiences with YT data from "YongYea" enabled us to create a template for the computational analysis of other channel data in our original data set—to find cases for comparison. We may not be able to retrieve caption texts from every channel due to technical limitations (e.g. no caption text available). Our technological setup, however, seems reliable enough to be able to deal with all sorts of—properly prepared—text data. For our research project presented here, comment data was most interesting, since we aimed to trace and analyze influencer-community interactions. YouTube distinguishes between top-level comments and responses to the top-level, which is machine-readably expressed in the identifier of the comment. Therefore, *conversations* will become an important datatype to be modeled by us. Our preliminary findings show that within fan communities there is awareness of their function as consumers in the gaming industry. Rejecting the notion of passive consumption, they tend to valorize their status and attempt to apply pressure by threatening to boycott or other forms of refusing to spend money.

The methodological procedure of combining a computational method like topic modeling with the selection criteria of GTM has proven to enable qualitative researchers to tackle massive amounts of digital data while also staying in pursuit of their research questions. In our example, we managed to reduce the scale of our example data and keep the quality for in-depth analysis by carefully balancing steps of theory infused decision criteria (deduction), like creating a data model for the computational analysis, and applying

criteria derived from empirical insights (induction), like our model of the field structure. Here, we argue, lies the core for a productive, but above all methodologically controlled and epistemologically reflected style of research, that is truly able to ground theories on cultural and social phenomena in digital data. But to make it productive, such research practices do not just need carefully adapted tools like our working prototype for a topic modeling system, but also proper documentation and report in the respective publications. We argue that *making sense* of digital data is no less stringent and applicable as a method than computational methods—and accordingly must also become part of the documentation and evaluation.

4. Discussion: Digital Methods for Qualitative Questions

The research presented in this article originates from a specific context and a specific research question, that is interested in the role of consumers in the computer games industry. Its subject is social interaction and not e.g. morphology in literature. We need a mindset for this research that differs from Moretti's *Distant Reading*, since we cannot rely on a canonical body of works that constitute the genre of YT-comments, and therefore would have little to gain by pursuing a distant reading in the stricter sense. Approaching the empirical field openly, as proposed by Straussian GT, we cannot afford to prescribe a fixed research design and had to decide on the mix of methods according to emergent, albeit preliminary findings in an evolving set of data.

We made human interpretation visible in our research processes by highlighting 'handover' constellations. Our exemplary research project serves to support two arguments: First, we propose to use grounded theory as a description language for research. Second, we argue in favor of conceptually grasping human interpretation as a stringent method.

We discuss the value of our approach along two substantial findings. We want to show how to deal with the problematic notion of theorylessness by adopting GTM. We found that there is, in fact, little need for new a priori theories. Instead, we suggest deploying reflective and reconstructive procedures to generate theory grounded in empirical research.

- The interplay between computational and qualitative methods can be characterized by the relation between the reduction of dimensionality as we encountered it in topic modeling and the increase in abstraction which is the aim of GTM.
- If researchers seek to reduce the quantity of data without losing quality, we found an abductive research attitude to be best suited to develop a procedural research framework.

4.1. Interrelating Qualitative Research and Computational Methods better

We showed in our example iteration how we incorporated topic modeling methodologically in a qualitative research process. Our most foundational assumption is that *data* like automatically generated YT captions or user generated comments is not reserved for the realm of the digital. In the context of GTM, digital data is a type of data among others. GTM enables us to draw data from many sources and iteratively develop an empirically grounded theory on social and cultural phenomena that guides through the research process. None of this is, by itself, news because GTM as well as computational methods have widely proven themselves in practice. Our goal was to uncover implicit theoretical operations and assumptions that are already at play in many research processes. We propose to use GTM to organize, structure, and describe evolving research processes—especially between qualitative researchers and researchers who build and use computational methods like topic modeling.

The question of how to configure interdisciplinary research processes in DH has been raised frequently. We approached the outer edges of our expertise as qualitative researchers when we engaged with topic modeling in our example. We did not let ourselves be deterred from using this elaborate computational method because we had confidence in the work of our colleagues from the information and computer sciences.

According to Nikolenko et al. "novel fields of applications for topic modeling have begun to emerge" (88). The object of research in our case study can be characterized as "user-generated texts coming from the blogosphere or social networks" (ibid.) which already locates our case study in this area of application for topic modeling. The authors approach a problem that is similar to ours: How can topic modeling be made more accessible, expressive and reliable for qualitative researchers? Nikolenko's and colleagues' expertise lies in information science and it seems plausible for them to make an effort to propose viable extensions of topic modeling that may promote a healthy and fruitful relation between two very different domains. Our understanding from an epistemological and methodological point of view is that we share the ambitions of Nikolenko et al. We approach it, however, from the other end of the disciplinary spectrum. While they raise the question of how computer scientists could extend a helping hand to their qualitative colleagues, we propose GTM to make explicit what theoretical implications are often already present in interpretative research.

Our proposed theory-telling of theory grounded in data epistemologically implies two simultaneous but opposing movements: The **reduction of dimensionality** in data with the help of computational techniques on the one hand, enables an **increase in the abstraction** of theory on the other hand. This characterization adequately describes the research practice of the

example (3.2), and at the same time productively implements an epistemic tension that is usually presented as contradictory: the handling of complexity of social and cultural phenomena in computational processing.

We found that especially visualizations (3.2.4) are important tools to enable researchers to interpret computational results. This marks an important point of contact between reduction of dimensionality and increase of abstraction. In our example, we had to carefully consider how to visualize our results. Visualizations are, on the one hand, the product of computations performed in multidimensional theoretical spaces that are projected on a two-dimensional surface. On the other hand, it is critical for us as researchers that visualizations aid in increasing the level of abstraction so that our efforts can actually amount to an empirically grounded theory of the field of research.

While we share Nikolenko et al.'s ambition in the long run, our contribution—based on our expertise in qualitative researchers—seeks to encourage qualitative researchers to engage with computational methods like topic modeling, especially when being faced with overwhelming amounts of user-generated textual data from social media. It also extends the rigorous approach of GTM to continuously evolve, test and adapt implicit assumptions and theory to DH. In that regard, we see GTM as a helping hand to computer scientists in interdisciplinary contexts who wish to further their understanding of colleagues with a qualitative research background.

4.2. Grounding Theory as Abductive Research Attitude

The goal of GTM is to move towards higher-order (more abstract) theories from empirical data. Complexity reduction is explicitly necessary to condense data into a model that has explanatory power by revealing and relating the *relevant* aspects of the context to be explored. Our argument in this article, and that of GTM, is that these relevant aspects can neither be defined a priori analyzing data, nor obtained purely inductively from the data and their computational processing. Instead, an ***abductive research attitude*** is needed.

Abduction is necessary when something incomprehensible is discovered in the data, which does not fit any type or rule in the body of knowledge. Since no suitable rule can be found, a new one must be discovered to cover the case in relation to the existing theory. And this must be done by intellectual effort. The logical form of this operation is that a new rule is mentally designed and tested, and when it prevails, it becomes clear how the new case can be explained (Reichertz). Abduction is a thought process to help social researchers to be able to make new discoveries *in a logically and methodologically ordered way*.

Research, even with digital tools and digital data, is a constant problem-solving activity: Theories do not magically emerge from data, but through the researchers' access to it: what they want to know from it, how they

correlate the data, how they process it—and here in our example, how they make it processable and narrow it down in the first place. Our study does not start with selected data sets, but with an interest that could refer to a variety of possible data from social media interactions: Comments, numerical and metadata, machine-readable identifiers, text data from e.g. usernames, or automatically generated captions from the videos, etc. And similarly polyphonic as possible data sources for the analysis is the operationalization of possible factors that could be relevant for the phenomenon under investigation. The distinct ability of tools such as topic modeling is to reduce the dimensionality of this data while preserving relevant information or making it accessible in the first place. We have shown that defining relevance requires an iterative and reflective process (3.2, 4.1).

However, this process is not necessary because computational methods cannot in principle provide access to the complexity of social and cultural phenomena (as a prejudice in some communities of qualitative social research goes), but because it follows the 'nature' of the data-based research process itself. Complexity here is not an ontological question (e.g. about the number and simultaneity of variables in social reality), but an epistemological one about strategies with which the dimensionality of data—as is typical in DH—can be adequately reduced to the research interest and data in order to become or remain capable of action as a researcher. GTM directs the view to practices and processes of research rather than to states or even immovable qualities of the phenomena being researched. The GTM process (*figure 3*) is based on an abductive research attitude, that asks the researchers to continuously generate abstract and higher-order theories from data by going through loops of inductive and deductive inference.

Conclusion

Instead of a new paradigm or epistemological turn within DH or related disciplines, we argued to turn to qualitative methodology, namely GTM, to deal with the opportunities and challenges digital data and research tools provide. Our example iteration from an upcoming case study was able to show how both, machine learning techniques and qualitative data analysis, intertwine and are not mutually exclusive. Iterative resampling with the help of natural language processing can help to open up large amounts of data for qualitative evaluation in such a way that reliable statements via GTM can be worked out with justifiable effort. Here, sensitivity to where sampling criteria are defined and executed is important. We described the process of resampling therefore as alternating between human-generated assumptions and machine-generated. We argued that GTM provides descriptive language and methodological tools for such iterations.

Our article consequently shows that a turn to the practice of digital research productively addresses two problematic aspects in the context of theory(lessness) in DH: First, looking at the implications of digital-research

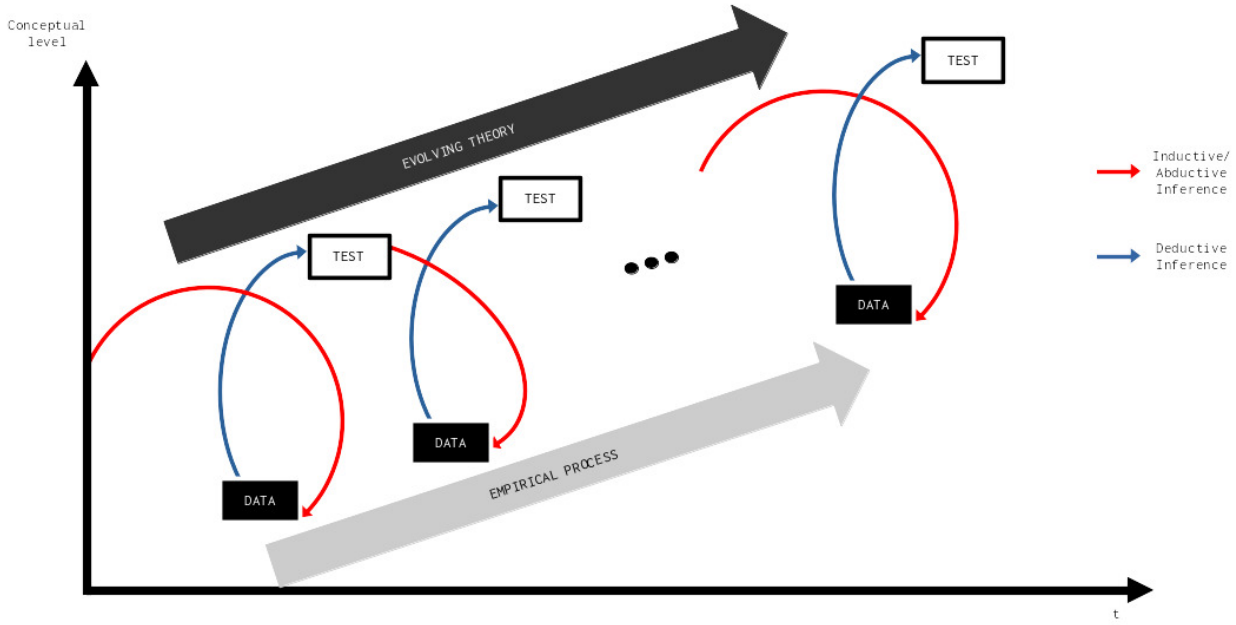

Figure 3. Abductive Research Attitude, adapted from Strübing (48).

technologies and data, it is clear that many assumptions are already embedded in a DH project before it has even begun. Second, GTM gives us a procedural framework that renders these implications practically manageable and even puts epistemic operations typically presented as opposites into a procedural framework. We believe that this perspective can not only inform the discourse on theory (or the lack thereof) in DH, but also the use of methods and tools to live up to the GTM criteria for grounding theory in data.

Submitted: June 15, 2022 EDT, Accepted: June 30, 2022 EDT

WORKS CITED

Anderson, Chris. "The End of Theory: The Data Deluge Makes the Scientific Method Obsolete." *Wired*, 23 June 2008, https://www.wired.com/2008/06/pb-theory/.

Benardou, Agiatis, et al. "Understanding the Information Requirements of Arts and Humanities Scholarship." *The International Journal of Digital Curation*, vol. 1, no. 5, 2010, pp. 19–33.

Bischof, Andreas, and Monika Wohlrab-Sahr. "Theorie-orientiertes Kodieren, kein Containern von Inhalten! Methodologische Überlegungen am Beispiel jugendlicher Facebook-Nutzung." *Praxis Grounded Theory. Theoriegenerierendes empirisches Forschen in medienbezogenen Lebenswelten. Ein Lehr- und Arbeitsbuch*, edited by Pentzold et al., Springer VS, 2018, pp. 73–101.

Blei, David M., et al. "Latent Dirichlet Allocation." *JMLR*, 2010.

Boyd, Danah, and Kate Crawford. "Critical Questions for Big Data: Provocations for a Cultural, Technological, and Scholarly Phenomenon." *Informacios Tarsadalom*, vol. 15, no. 2, Jan. 2012, pp. 662–79.

Burrows, Roger, and Mike Savage. "After the Crisis? Big Data and the Methodological Challenges of Empirical Sociology." *Big Data & Society*, vol. 1, no. 1, Apr. 2014, p. 205395171454028, https://doi.org/10.1177/2053951714540280.

Charmaz, Kathy C. *Constructing Grounded Theory: A Practical Guide through Qualitative Analysis*. Sage, 2006.

Clarin. *Tools*. https://www.clarin.eu/content/tools. Accessed 15 Nov. 2021.

Covington, Paul, et al. "Deep Neural Networks for Youtube Recommendations." *Proceedings of the 10th ACM Conference on Recommender Systems*, 2006.

Davidson, James, et al. "The YouTube Video Recommendation System." *Proceedings of the Fourth ACM Conference on Recommender Systems - RecSys '10*, 2010, pp. 293–96, https://doi.org/10.1145/1864708.1864770.

DockerHub. *Startingpage*. https://hub.docker.com. Accessed 15 Nov. 2021.

Elastic. "Docs." *Kibana Query Language*, https://www.elastic.co/guide/en/kibana/current/kuery-query.html. Accessed 15 Nov. 2021.

---. *Elasticsearch*. https://www.elastic.co/de/elasticsearch/. Accessed 15 Nov. 2021.

---. *Kibana: Explore, Visualize, Discover Data | Elastic*. https://www.elastic.co/kibana. Accessed 15 Nov. 2021.

Galison, Peter. *Image and Logic: A Material Culture of Microphysics*. University of Chicago Press, 1997, https://doi.org/10.1063/1.882027.

Gensim. *Gensim: Topic Modeling for Humans*. https://radimrehurek.com/gensim/. Accessed 15 Nov. 2021.

Gitelman, Lisa, editor. *Raw Data Is an Oxymoron*. MIT press, 2013.

Github. *Docker*. https://github.com/docker. Accessed 15 Nov. 2021.

---. "Elastic/Elasticsearch." *Public*, https://github.com/elastic/elasticsearch. Accessed 15 Nov. 2021.

---. *RaRe-Technologies/Gensim: Topic Modeling for Humans*. https://github.com/rare-technologies/gensim. Accessed 15 Nov. 2021.

---. "Sgsinclair/Voyant." *Public Archive*, https://github.com/sgsinclair/Voyant. Accessed 15 Nov. 2021.

Glaser, Barney G. *The Grounded Theory Perspective: Conceptualization Contrasted with Description*. Sociology Press, 2001.

Glaser, Barney G., and Anselm L. Strauss. *Discovery of Grounded Theory: Strategies for Qualitative Research*. Routledge, 2017, https://doi.org/10.4324/9780203793206.

Halevy, Alon. "Technical Perspective: Building Knowledge Bases from Messy Data." *Communications of the ACM*, vol. 60, no. 5, Apr. 2017, pp. 92–92, https://doi.org/10.1145/3060584.

Harrison, Steve, et al. "The Three Paradigms of HCI." *Alt. Chi. Session at the SIGCHI Conference on Human Factors in Computing Systems San Jose, California, USA*, 2007, pp. 1–18.

Jahoda, Gustav. "Quetelet and the Emergence of the Behavioral Sciences." *Springerplus*, vol. 4, no. 1, Sept. 2015, pp. 1–10, https://doi.org/10.1186/s40064-015-1261-7.

JSON. *Introducing JSON*. https://www.json.org/json-en.html. Accessed 15 Nov. 2021.

Kitchin, Rob. "Big Data, New Epistemologies and Paradigm Shifts." *Big Data & Society*, vol. 1, no. 1, Apr. 2014, p. 205395171452848, https://doi.org/10.1177/2053951714528481.

Kozinets, Robert V. "Netnography: The Essential Guide to Qualitative Social Media Research." *Netnography Unlimited*, Dec. 2020, pp. 3–23, https://doi.org/10.4324/9781003001430-2.

Latour, Bruno. *Science in Action: How to Follow Scientists and Engineers through Society*. Harvard University Press, 1987.

Leigh Star, Susan. "This Is Not a Boundary Object: Reflections on the Origin of a Concept." *Science, Technology, & Human Values*, vol. 35, no. 5, Aug. 2010, pp. 601–17, https://doi.org/10.1177/0162243910377624.

Liu, Alan, et al. "Open, Shareable, Reproducible Workflows for the Digital Humanities: The Case of the 4Humanities.Org WhatEvery1Says." *Project. In DH*, 2017.

Manovich, Lev. *Cultural Analytics: Analysis and Visualization of Large Cultural Data Sets*. Accessed 23 Nov. 2007.

---. "The Science of Culture? Social Computing, Digital Humanities and Cultural Analytics." *Journal of Cultural Analytics*, vol. 1, no. 1, 2016, p. 11060.

Mobygames. *Metal Gear Games*. https://www.mobygames.com/game-group/metal-gear-series. Accessed 15 Nov. 2021.

Moretti, Franco. *Distant Reading*. Verso, 2013.

Nelson, Laura K. "Computational Grounded Theory: A Methodological Framework." *Sociological Methods & Research*, vol. 49, no. 1, 2020, pp. 3–42, https://doi.org/10.1177/0049124117729703.

Nikolenko, Sergey I., et al. "Topic Modelling for Qualitative Studies." *Journal of Information Science*, vol. 43, no. 1, July 2016, pp. 88–102, https://doi.org/10.1177/0165551515617393.

Peirce, Charles S. "Pragmatism as a Principle and Method of Right Thinking." *The 1903 Harvard Lectures on Pragmatism*, State University of New York Press, 1997.

Reichertz, Jo. "Abduction: The Logic of Discovery of Grounded Theory." *The SAGE Handbook of Grounded Theory*, edited by Antony Bryant and Kathy Charmaz, Sage, 2007, pp. 214–28, https://doi.org/10.4135/9781848607941.n10.

Rieder, Bernhard, et al. "From Ranking Algorithms to 'Ranking Cultures' Investigating the Modulation of Visibility in YouTube Search Results." *Convergence*, vol. 24, no. 1, Jan. 2018, pp. 50–68, https://doi.org/10.1177/1354856517736982.

---. "What Is in PageRank? A Historical and Conceptual Investigation of a Recursive Status Index." *Computational Culture*, vol. 2, 2012.

Savage, Mike, and Roger Burrows. "The Coming Crisis of Empirical Sociology." *Sociology*, vol. 41, no. 5, Oct. 2007, pp. 885–99, https://doi.org/10.1177/0038038507080443.

Sievert, Carson, and Kenneth Shirley. "LDAvis: A Method for Visualizing and Interpreting Topics." *Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces. Baltimore, Maryland, USA*, 2014, pp. 63–70, https://doi.org/10.3115/v1/w14-3110.

spaCy. *Lemmatizer - spaCy API Documentation*. https://spacy.io/api/lemmatizer. Accessed 15 Nov. 2021.

Strübing, Jörg. *Grounded Theory. Zur sozialtheoretischen und epistemologischen Fundierung eines pragmatistischen Forschungsstils*. Springer, 2014.

Svensson, Patrik. "The Digital Humanities as a Humanities Project." *Arts and Humanities in Higher Education*, vol. 11, no. 1–2, Dec. 2011, pp. 42–60, https://doi.org/10.1177/1474022211427367.

---. "The Landscape of Digital Humanities." *Digital Humanities Quarterly*, vol. 4, no. 1, 2010.

Voyant. *See through Your Text*. https://voyant-tools.org/. Accessed 15 Nov. 2021.

Winner, Langdon. "Do Artifacts Have Politics?" *Daedalus*, 1980, pp. 121–36.

YouTube. *YongYea*. https://www.youtube.com/yongyea. Accessed 5 Nov. 2011.