Uncovering the Challenges in Collection, Sharing and Documentation: The Hidden Data of Social Media Research?

Katrin Weller
GESIS Leibniz Institute for the Social Sciences
katrin.weller@gesis.org

Katharina E. Kinder-Kurlanda
GESIS Leibniz Institute for the Social Sciences
katharina.kinder-kurlanda@gesis.org

Abstract

This paper offers insights into social media researchers' everyday practices in dealing with constraints and challenges in the areas of data access, in terms of data collection and sharing, and data publication. We believe that such insights need to be taken into account when discussing methodology and epistemology of social media data to ensure that strategies employed to achieve validity are appropriate. A qualitative, ethnographic research approach (based on expert interviews, observation and a qualitative questionnaire) is used to capture the practices and discussions of a variety of social media researchers. The paper concludes that due to the current situation in data sharing and publication opportunities, both the actual research data and much of the technical knowledge in social media research remain hidden unless revealed by studying researchers' everyday practices. Such studies reveal a considerable impact of external constraints on researchers' attempts to achieve validity and better re-search quality.

1 Introduction

The use of datasets based on user-generated content and user networks from social media platforms is becoming increasingly popular in a variety of scholarly disciplines. We summarize the research efforts that make use of social media data as social media research, although it needs to be pointed out that this is not a coherent research field with an established set of methodological approaches, evaluation and documentation standards, or common best practices for research ethics. Social media research can come in a variety of shades – e.g. using different types of datasets, different conceptual frameworks, various theoretical assumptions – and is often influenced by the different disciplinary backgrounds of the respective researchers, which may hail from the social sciences and humanities (e.g. sociology, political science, linguistics, communication science) as well as STEM disciplines (e.g. computer science, physics, health care). The studies that make use of "Big Data" and new approaches in data mining and data analysis (in contrast to, for example, qualitative work based on close reading and analysis of selected content at a much smaller scale) are facing particular challenges. Recently, and particularly in reaction to claims that Big Data may make many traditional approaches to studying human behavior obsolete (Anderson 2008), critical position papers have emerged that question different aspects of current Big Data analyses in social media research. Challenges are usually seen in three areas: methodology/epistemology (particularly concerning issues of data quality), ethics and, connected to both of these, the role of platform providers and streaming APIs. The following overview briefly illustrates the dimension of the ongoing debate, without claiming completeness.

boyd and Crawford (2012) criticize the focus of many Big Data researchers on quantification to the detriment of other, equally valuable approaches and remind the community to consider research ethics, data access and digital divides, appropriateness of data, and (constructed) objectivity. Bruns (2013) builds on their arguments and describes practices of data collection and documentation, and the lack of replicability in many studies. He also addresses how big social media research challenges current publication formats and practices. Formats may need to be adapted allow for room for detailed explanations of complex methodologies and to acknowledge the speed at which internet platforms, users and content change. Karpf (2012) also addresses this issue of 'internet time' and advocates for more experimental methods and transparency, which again requires detailed explanations of approaches. Ruths and Pfeffer (2014) call for higher methodological standards when working with Big Data from online platforms and discuss a variety of sources for bias. One main source for bias is a lack of knowledge about how the population of an online platform relates to the (larger) population to be studied – an issue which is often exacerbated by the fact that different platforms attract different types of users and that automated/spam accounts may influence the datasets. Bias can also occur and be difficult to identify or correct due to proprietary and nontransparent algorithms for data access, missing data that have not been stored by the platform provider, and incomparability of data analysis methods. Tinati et al. (2014) also consider the methodological challenges those face who engage with Big Data and offer tools that address both the ephemeral, changing nature of platforms such as Twitter and the many temporary or permanent networks that form within one platform. Other authors discuss the pitfalls of divorcing theory from Big Data analysis (e.g. Lazer 2014; Frické 2014).

The nature of the data available on different social media platforms shapes and sometimes limits the type of analyses that can be conducted based on these data (Giglietto, Rossi and Bennato 2012). Platform providers' policies and streaming APIs' affordances shape the nature of research projects. For example, Twitter's policies affect data access and research opportunities (Puschmann and Burgess 2013). Due to a lack of information about data collection procedures it is mostly unclear what data provided by the Twitter API is representative of (Bruns and Stieglitz 2014; Morstatter et al. 2014).

Ethical issues are another important challenge in social media research and usually concern the difficulties of anonymization and a lack of informed consent in social media users (e.g. Zimmer 2010). While the recent controversy around the Facebook contagion experiment (Kramer et al., 2014) has created critical debates on the ethical and legal boundaries of social media research, many issues in this debate are not inherently new to social media research but rather are traditional ethical challenges reaching a new dimension in this context (Schroeder 2014a). Beninger et al. (2014), however, are among the first to investigate social media users' expectations of how their data are being treated by researchers.

With this paper we want to contribute to this ongoing discussion on social media research methods and their quality. In particular, we add considerations on the constraints in social media research which are posed by challenges in the research environment and the infrastructures for data access. By addressing these constraints, which are often outside of researchers' control, we aim to lay a foundation for considering them in the discussion of social media and Big Data research methodology. We therefore focus on aspects of data access, data sharing and opportunities for sharing and publishing research results, rather than on the applied research methodologies, and researchers' attempts to achieve validity and reliability of research results within the described constraints.

We are using the term 'data sharing' to summarize all activities of making datasets, which have been obtained from social media sources for research purposes, available to others. This broad definition follows argument that research data sharing can happen in many ways, from richly structured and curated to highly informal (Borgman 2013). While the process of acquiring datasets from social media platforms (e.g. through the API or third party tools) is usually referred to as 'data collection' in the context of social media research, data sharing assumes that the data has already been collected and stored by someone and is then handed over or made available to third parties. Data sharing and data collection can be considered as two different approaches to gain 'data access'; and this paper will contribute to the understanding of the relation between these two types of data access. The dimension of data sharing is currently rarely considered in discussions of social media research methodology. In this paper we shed some light on the practical implications of issues around data sharing as perceived by the research community. Such an exploration of data sharing experiences can also lay a foundation for introducing insights and solutions from other disciplines in which datasets are already frequently published and shared for reuse. Useful examples may originate in current practices in the social sciences and linguistics. Linguists frequently work with already available linguistic corpora (i.e. written texts or transcribed oral language, sometimes in form of annotated corpora, e.g. enriched with word class labels or other grammatical, semantic or historical information). Frameworks such as the Linguistic Data Consortium[1] collect and create reusable corpora for linguistic research. In the social sciences, the use of secondary data, i.e. data that have been collected in previous studies and made available for reuse, is common. Available datasets are usually survey data and well-developed standards exist for data documentation, for example provided by the Data Documentation Initiative (DDI)[2]. Some datasets are so well recognized that they act as so called benchmark datasets. At this point we do not assume that the sharing of social media data should follow any of these organized or even standardized efforts or conversely that it should remain highly informal.

While data access and challenges posed by infrastructure are the focus of this paper, they are only a single dimension of a broader ongoing research project that also looks at various other challenges in social media research, including epistemology, research ethics (Weller and Kinder-Kurlanda 2014) and interdisciplinarity (Kinder-Kurlanda and Weller 2014).

In order to address the various dimensions of social media research, we conducted a qualitative study of social media

[1] https://www.ldc.upenn.edu/about [13 April 2015]

[2] http://www.ddialliance.org/ [13 April 2015]

researchers from various disciplines. In interviews and observations we explored their reflections on everyday research practices when engaged in data collection, analysis, sharing and publication. In this way we captured the expertise and the opinions of social media researchers in order to explore a) their motivations for working with social media data, b) their methodological approaches and best practices, and c) the perceived challenges, pitfalls and drawbacks. Our qualitative ethnographic research design allowed us to explore various practical aspects of social media research and enabled us to capture the current state of this emerging research field. The research design is comparable to the approach by Schroeder (2014b) and Taylor et al. (2014) who interviewed social scientists and economists about their perspectives on Big Data research. In our work, we apply a broader scope in including different scholarly disciplines, but focus on a more specific type of Big Data, namely data from social media platforms.

This paper presents work in progress which will still be continued and enriched in the future, e.g. based on new additional interviews with social media researchers from diverse backgrounds and on in-depth analysis of the interrelations of different sub-topics discussed by the interviewees and developments over the course of time.

2 Method

The insights in this paper are based on face-to-face interviews, observations and the results from a qualitative questionnaire. We are an information scientist (with a focus on social media) and a cultural anthropologist (with a focus on science and technology studies and a strong background in ethnography) and we conducted all fieldwork together. Following an iterative approach between theory and fieldwork we entered the field with some assumptions based on the discussion of social media ad big data in the literature. Findings from our study informed further literature reading which in turn caused adaptations of our interview guide. We conducted 42 interviews with social media researchers at four international conferences in 2013 and 2014. As conference participants ourselves we also observed researchers' discussions surrounding data usage and their practices of knowledge exchange. In particular we paid attention to discussions of presentations but also to those happening in breaks and at receptions. The anthropologist took notes at the conferences. The notes were discussed and amended by both researchers after the conferences. Interviewees also filled in a qualitative online questionnaire in the days or weeks following the interview. A link to the questionnaire was provided by email and participants were also sent a reminder to fill in the questionnaire. 35 participants filled in the questionnaire (33 completed forms, 2 incomplete forms) which mostly contained more detailed questions about tools used, platforms studied and methods applied. The aim of the questionnaire (which contained both multiple choice questions with an additional text field and entirely open questions) was to allow participants to add details at their leisure after the interview and to collect more detailed information on data management processes from collection to publication.

While all conferences chosen for the interviews had an interdisciplinary outlook they still attracted different communities of scholars studying the Web, the Internet or online communication. We restricted our set of interviewees to researchers who engaged with content or data created by users on Web platforms, and excluded those who, for example, conducted surveys about social media consumption patterns. Potential candidates who matched our criteria were identified from the online conference programs of the four selected conferences and sent an email invitation. In three cases we arranged for interviews on-site at our institution as interviews could not be conducted at the conference itself.

Interviewed researchers had different disciplinary backgrounds, including, for example, computer science, media and communication studies, social science, physics, linguistics, and information science. We interviewed researchers working in Europe, the United States, Australia, South America and Western Asia (with only one researcher from South America and Western Asia each). Interviewees ranged from Master students to full professors in terms of professional levels (Master students, PhD students, postdoctoral researchers, senior researchers or professors). Most interviewees had experiences with research on social media data from several platforms. Researchers had specifically based research on data gathered from Twitter, blogs, Facebook, and many other platforms such as Four-square, Tumblr, 4chan or reddit. Almost all interviewees were working at universities or other non-profit research institutes with only two working in industry. Perspectives from researchers based within social media companies are missing, as only a single participant belonged to this group and only few participants worked in direct collaborations with such companies (e.g. at institutions that explicitly had part-nered up with social media companies for data access, as the MIT and Twitter have done recently).

For the interviews we used an interview guide which covered the main topics to be addressed rather than specific questions. In addition to allowing explanation, correction, thinking-aloud and discussion, the face-to-face interview situations also proved to be beneficial in providing an atmosphere of trust and non-judgment when addressing sensitive topics such as ethical concerns or the lack of best practice standards in data management. The interview guide comprised topics that would allow studying various dimensions of research practices in dealing with social media data including also, for example, data management, data analysis, methods and epistemology, as well as collaboration and interdisciplinarity.

Interviews usually lasted around 30 minutes each. All interviews were transcribed into text. The written transcripts were then interpreted using a 'lean coding' approach in which themes were built through reducing and combining categories found in the data (Creswell 2013). Codes were discussed and iteratively defined by both authors.

3 Results and Discussion

When studying the interdisciplinary nature of social media research based on our first set of interviews, we observed that researchers who had no background in computer science often relied on collaborations with more technically oriented colleagues in order to set up data collection processes. This division of tasks would lead to practical challenges in the everyday work, such as establishing feasible workflows and a common language for researchers from different disciplines (Kinder-Kurlanda and Weller 2014). Interdisciplinary collaboration in social media research also faces a challenge beyond individual workflows: The lack of technical expertise can lead to an increased inequality in terms of data access as described by boyd and Craw-ford (2012). Data sharing would be one of the most obvious solutions to bridge the divide between 'data haves' and 'data have nots' by adding an additional mode of data access besides data collection – but we will show here that the sharing of social media data is problematic and that researchers who aim to share often face an ethical dilemma.

The challenges of inequality and the ethical implications of data sharing result out of the context of various (technical) challenges around data collection approaches in social media studies, which we will first explore.

3.1 Data Collection Challenges

In the interviews and questionnaires, many researchers elaborated on the "data collection problem". The most frequently mentioned issues pertained to:

- *Quality issues in the data provided by APIs*: Many researchers faced problems with data quality, such as a lack of clarity with regard to the bias in the sample provided by an API, insufficient documentation of the data, and opaque collection processes.
- *Rate/volume limits*: Researchers were often restricted in data collection by the various ways in which APIs and platform providers limited the type and amount of data they could retrieve.
- *Ephemeralism of platforms and data, especially updates to an API or changes to the platform*: Changes to the structure of social media websites during the data collection period often had to be accommodated and could even cause certain data that an analysis was based on to "disappear".
- *Technical difficulties on the researchers' side*: Many researchers faced technical difficulties in their collection setup such as instabilities in the collection infrastructure, code crashing, or running out of disk space.

About two thirds of the interviewees reported that they had encountered such typical problems during data collection, most of them had encountered several or even all of the issues listed here.

3.2 The Need for Data Sharing to Enable Data Access

When asked in which areas they saw a need for assistance, two topics were mentioned most often by researchers: data collection and data sharing (i.e. the two different types of data access). Many of our interviewees had at some point experienced challenges in accessing social media data – although there were significant differences in how severe these challenges would prove.

Sharing Can Alleviate Inequalities in Data Access

We found that there was a strong inequality with regard to the amount and quality of social media data that the interviewed researchers had access to. The majority of interviewed researchers had never paid for access to social media data (which would allow for almost wholesale access during data collection), but seven researchers had experiences with buying data, spending up to 'several thousand dollars' in large projects. Most researchers described the social media research field as a highly uneven landscape where some – especially those researchers working in or in collaboration with industry – had access to large amounts of high-quality data whereas others were restricted by legal and financial concerns to relying on data retrieved through the APIs or even collected manually[3]. Inequality of access, as also recently observed by Ruths and Pfeffer (2014, p. 1063) was perceived as a challenge for the broader research community: "The rise of 'embedded researchers' (researchers who have special relationships with providers that give them elevated access to platform-specific data, algorithms, and resources) is creating a divided social media research community." Many of the interviewed researchers wished for social media companies to be more obliging and cooperative with regard to data collection. We also witnessed that researchers who had entered the field early on had initially been able to establish close collaborations with or receive data from individual companies, while over the course of time data collection had become more and more difficult. Social media platform providers were seen as closed-off and difficult to

[3] In fact, as mentioned above, the majority of interviewees were researchers not working in industry or at institutes with direct collaboration pro-jects with social media companies.

approach, with several researchers even reporting that it had become impossible to receive an answer to email requests.

Obstacles to data collection could, in addition to money, also be found in a lack of language skills or in cultural constraints (e.g. differences in popularity of specific social media platforms across cultures and countries). Some countries or disciplines were seen to be 'behind' others with regard to knowledge in how to collect data and in access opportunities[4]. Certain communities had therefore not been able to develop the field of social media research. One interviewee described observing a new community entering the field of social media research in Italy: "And it's the first time that I saw so many studies on Twitter and Facebook [in this particular community]. I was amazed. But they are really starting from scratch."

We also observed inequalities between university institutions within the same country. Large and well-funded research institutions were seen to be mitigating the gap between academia and researchers working in industry or at platform providers' own research departments by either being able to negotiate special agreements for data collection or by being able to buy data. Researchers located in Europe (or other places outside of the US) or those working for smaller institutions were more dependent on own efforts in data collection. Discipline-specific funding programs also played a role in whether researchers were in a position to pay for data collection: Researchers with a background in the natural sciences, for example, were used to spending large amounts of money for expensive data collection tools, and funding calls in certain areas were traditionally making allowances for such costs. Overall, opportunities for access to data often seemed to follow the established economies of attention, funding and publications found in international academia.

Data sharing was thus perceived as necessary to include different communities and to close the inequalities gap.

Sharing can Prevent Labor-Intensive and Time-Consuming Data Collection

Given the current challenges in collecting social media data as outlined above, there is a second argument that calls for feasible ways to share research data in social media research: data collection takes up a large amount of time and effort for the individual researchers.

Researchers stated, for example, that "data cleaning is painful" or that "too much energy goes into data collection". The central argument here was that sharing of (high-quality) datasets would prevent spending high proportions of time, effort – and eventually research funding – on tasks that had already been performed by others. As one interviewee phrased it, currently "lots of people […] are all doing sort of the same thing with different levels of efficiency or success. And it's just horrible." This holds for developing tools for data collection as well as for selecting the collection parameters, such as developing and maintaining a list of keywords to track Twitter activities.

Sharing is necessary to enable better quality of research

A third reason for why data sharing may improve social media research was also to be found in issues surrounding access and concerned the impact of data availability on research quality. Interviewees viewed the barriers to access as crucial constraints that had a considerable impact on research success. Data availability could influence research efforts in many ways, for example making it impossible to answer some research questions or forcing researchers to adapt research questions during the course of a project. Most interviewees agreed that the barriers for data access were too high and had a negative impact on the quality of social media research. However, the lack of data availability not only prevented interesting research, it also caused a different difficulty. Many researchers voiced concerns that issues around availability often led to opportunistic approaches being applied in data analysis (see also Bruns, 2013) with researchers making use of the data that was available rather than data that was most suitable for answering the specific research question. The preference for easy-to-collect data also led to some platforms being over- or underrepresented in social media research and to other biases. This is a point which we will explore further in future work on epistemology in social media research.

Sharing datasets was also seen to enable comparative studies, reproducibility of results and peer-review of other researchers' work. Some interviewees had experience with using other types of secondary data (e.g. surveys, linguistic corpora) and wished for comparable infrastructures for accessing reusable datasets in social media research.

3.3 Data Sharing as a Legal and Ethical Dilemma

Given the situation outlined so far, it is understandable that we found many social media researchers to be very open towards the idea of sharing 'their' datasets – something that should be highlighted because it is not necessarily the case in other research fields with more established standards in working with secondary data (Fecher et al. 2015; Borgman 2013). The desire to share data was pointed out by participants themselves, who described how they had experienced social media researchers not to be 'guarding' their data in the way that researchers from other disciplines usually would.

Researchers Feel an Obligation to Share Data

In many cases we even found expressions of re-searchers who felt an ethical obligation to share their datasets, either with other researchers or with the public. Data sharing within research groups was a common practice. One re-

[4] Skills such as programming also played a role and we have explored this issue in another paper (Kinder-Kurlanda and Weller 2014).

searcher claimed that “we share datasets with everybody, actually. We don’t feel we own that.” The core characteristic of social media data as being created by the social media users for other purposes seemed to play a critical role for this:

> “It’s all public, it doesn’t belong to us, we don’t create the data, we don’t evoke it, I mean it’s natural. I don’t think you have the right to really keep other people from it, no.”

Some researchers expressed the desire to share their datasets with the public and with participants. We also learned about single cases in which funding bodies required data sharing or in which a publisher (journal or conference proceedings) asked for the data to be submitted with the manuscript (see section 3.5).

Obstacles to Data Sharing

Although many researchers expressed a desire to share, few social media datasets are in fact publicly available for reuse as secondary data. Mainly this seemed to be due to an insurmountable lack of clarity with regard to the legality of sharing social media data. Most researchers were unsure whether they were allowed to share the data collected or what repercussions they would face in the case that data was accidentally or deliberately shared. For example, one researcher said: “probably, possibly we’d be actually operating outside the spirit if not the letter and the law of the latest Twitter API.” While some platforms specifically prohibited sharing in the terms and conditions of APIs, the situation was very unclear if, e.g. no API was used or clear terms and conditions were not easily available. We found a general sense of unease and uncertainty surrounding the topic of the legality of sharing.

The uncertainty about the legality of data sharing had resulted in many researchers never publicly sharing data and instead having found various individual ways and strategies of sharing. Some would only share data with researchers involved in the same project while others would be happy to share data with researchers from other institutions. Some even claimed the emergence of a ‘grey market’ in which “everybody has all kinds of stuff on their machines (...). If people ask us for a data set because they are working on something similar (...), then we might share it, but we can never publicly do that.” These sharing practices seemed to mainly rely on personal contacts, thus exacerbating the inequality in access mentioned above. Social media research data was being ‘hidden’ in attempts to protect oneself and to deal with uncertainty.

Other factors prohibiting sharing included concerns common in sharing research data in general, such as ethical concerns surrounding users’ privacy and security concerns when sharing sensitive data. Overall data sharing in social media research currently differs fundamentally from well-established practices of using secondary datasets in other fields in that it is far less connected to the publication process. Datasets are rarely openly published but often shared privately in hidden environments.

We also encountered instances of datasets being passed on to other researchers which had not been used for analysis by the researcher who collected them. In this case we may not speak of secondary data in the traditional sense, as there is no primary analysis of the collected data. Out of fear that they would not be able to access certain data retrospectively, some researchers were collecting a variety of Twitter datasets in real-time, just in case that they would need them one day – and thus had more data than they would ever be able to use themselves. In single cases this raised questions about authorship if one person had only collected and provided a dataset but did not work with it him-self/herself.

3.4 Documentation and Technical Challenges in Data Sharing

While social media datasets were being shared, it was not always easy to reuse them. Despite being open to the possibility of sharing one’s own data, many re-searchers were reluctant to use data collected by others. We found that a significant number of researchers had experiences in working with data collected by others. Thirteen researchers reported in our questionnaire that they had worked with data obtained from a colleague. From the more detailed answers in the interviews we learned, however, that datasets obtained by others were used with caution. Some researchers would only reuse data obtained from close colleagues or directly coming from a social media company. One phrased it as follows:

> “I actually only use [other researcher’s datasets] where I’m very sure about where it comes from and how it was processed and analyzed. There is too much uncertainty in it.”

No interview partner explicitly referred to publicly available collections of datasets such as the ones available for the ICWSM conferences - but we did not always explicitly ask for this constellation, so there may still have been instances of using such collections. In future interviews we may include this in the interview guide or try to incorporate other ways to study the acceptance of publicly available datasets (e.g. via identifying citations of such datasets).

Technical Challenges and Lack of Standards

We found a general skepticism towards datasets collected by others. Often researchers were seen to argue that when using others’ data they would have less control over how data were collected. Different data collection strategies as well as different tools for data collection were seen to make datasets incompatible with each other:

“I think probably a couple times we’ve asked around if anyone else happened to have a particular dataset. […] but not so much, because they probably have tracked in a different data format, and then merging the two together actually becomes quite difficult as well."

Reuse was especially problematic for those researchers who possessed the technical skills and the expertise to understand the challenges inherent in data collection (as outlined in section 3.1). We also encountered cases where researchers with little experience or with fewer technical skills would happily receive any data they could get without questioning the technical details.

Many researchers were working on topics that were covered by other researchers as well and described how each research team would use their own, idiosyncratic datasets collected from the same platform, but in different ways, making comparisons difficult or impossible:

> "We would need replicability working. If there are three studies on […] Arab spring we were all working with the same dataset that would actually be scientifically decent instead of five studies with five different datasets."

Researchers also critically discussed the possibility of one or more centralized research institutions who could serve as providers of social media research data. While some researchers wished for "one basis of data" (in the sense of a benchmark dataset) or data being made available by one central, independent provider to alleviate the costs of data collection, others saw the advantages of collecting data from scratch in the fact that it ensured better data quality and allowed more insights into the data's provenance. They argued that a central social media data provider for researchers would therefore need mechanisms to ensure reliability, comparison and interchangeability of the data. Moreover, a danger of establishing a central institute for data sharing was seen in the fact that such an infrastructure may be driven by the desire to promote its own services rather than by establishing comparability with other sources. This argument leads to another big challenge in social media research, the quest for traceability and missing documentation standards.

Challenges for Documenting Workflows

While we witnessed scripts for data collection or data analysis being shared and reused within the community, not much information about processes was available. Some of this information may be considered "tacit knowledge" (Polanyi 1966), i.e. implicit knowledge that is difficult or even impossible to verbalize or that experts may not even be aware of. However, we found that researchers were well aware of many challenging details in data collection and were also able to provide detailed verbal descriptions of significant parts of their workflow-related knowledge. We therefore decided to speak of 'hidden knowledge' in data collection practices, rather than tacit knowledge, as we were mainly dealing with knowledge that could have been verbally expressed but for practical reasons was not being published. Tools for data collection were often not well-documented. Even researchers using the same third party tools for data collection (e.g. YourTwapperkeeper for Twitter data) found it hard to compare or reuse datasets from others, because they lacked knowledge of certain details, such as "How many other keywords the server was tracking at the time, for example, which influences how much data it is actually receiving." The lack of such details prevented an assessment of the quality of the data.

Being able to retrace all steps of collecting, processing and cleaning the data was seen as crucial for assessing data quality and ensuring that the data really held what it promised. In order to share data a detailed and comprehensive documentation of the data collection environment is required in order to allow judging the factors that could have led to data loss or reduced data quality. We learned that even collection of the very same data (e.g. from Twitter) on two different servers in parallel could result in two different datasets. But information on server downtimes or reboots was rarely documented and shared with the community. This also entailed that there was not enough public discussion of the "nitty-gritty nuts and bolts" of data collection and processing, with many researchers reporting that they learned about the pitfalls of specific tools or APIs either through trial and error or through informal communication with colleagues – which again led to duplicate efforts across the research community and to inequality in access to this hidden knowledge. Duplicate efforts could also arise as more and more researchers decided to create own tools that matched their particular criteria for data collection, which also led to new challenges such as adapting tools to the changing features of social media platforms and their APIs.

Another challenge arose out of the interdisciplinary nature of social media research: different disciplinary backgrounds would require different strategies for meaningful data collection and to understand approaches from different disciplines precise documentation was needed. Methodological standards, however, were discussed critically with some researchers arguing that they may curb the current phase of exploration and experimentation which was highly desired. Yet flexibility in methods seemed to call for even more precise documentation of what had been done and how, in order to judge the quality and facilitate building on the results.

Documentation Practices

Current practices in documenting data collection and other research activities often did not live up to the desired quality. In some cases, researchers admitted that they already had experienced difficulties in keeping track of their own activities and in understanding what exactly they had done in order to collect or clean a specific dataset. For example, one researcher reported that he had had to go back to the raw data after not being able to retrace data cleaning steps.

More frequently we encountered various levels of (mostly successful) improvisation around documentation of workflows. This included, among others, the following approaches to documentation:

- Preparing handwritten notes or notes in a text file
- Considering the actual code/dataset as documentation of what had been done (e.g. "I'm using R. Everything is in the programming language. So all my decisions are explicitly written")
- Publishing details in blog posts
- Using a wiki for capturing decisions
- Archiving all email communication in a project

Still some researchers considered their approaches as not as precise as they could have been ("I'm not doing that systematically", "unfortunately, I don't document much"). And while one researcher remarked that "so far no one asked for the documentation behind the tool we created", many interviewees expressed a desire to learn more about these details in the work of other researchers.

Documentation in working groups was particularly challenging. In interdisciplinary teams, different members had different expectations of what needed to be documented. If student assistants or technical specialists were hired by social scientists for some parts of data processing, there was the danger that they might leave the project before the end and take their expertise with them. Documentation was described as a thankless task with much effort and little recognition in scholarly reputation practices. One suggestion included that this task should therefore be outsourced to specialists such as librarians and archivists. Thus, the question remains of how to document work-flows and experiences in such a way, as to make data and documentation accessible to the wider community, thus allowing sharing of background knowledge in order to enable comparative research and reproducibility of results.

3.5 Publication Formats

Current traditional publication formats were perceived as rather unprepared for the in-depth documentation of research workflows.

Social Media Research Challenges Review Processes in Some Disciplines

Depending on their disciplinary backgrounds, researchers would face different challenges when trying to publish social media research in their disciplines' main journals or outlets. For example, one researcher reported that the reviewers used to certain types of qualitative analyses based on close readings of textual material failed to understand her concerns about not wanting to include verbatim quotes from user-generated content ("You don't have any quotes! How do I know that you did this study?"). Computer scientists on the other hand found it difficult to achieve publication of detailed descriptions of data collection infrastructures, as this was often not considered original research ("[…] the scripts run for days or weeks and I have to restore it and edit it and this is months of work that is basically not at all represented in the actual paper"). Some researchers had moved away from their core disciplines but still found it difficult to get their work published at all: "It's difficult to publish, right. Because you don't belong anywhere."

Although we came across instances, where it was discussed whether data collection should be a criterion to claim authorship or how to acknowledge a colleague who 'merely' collected a dataset without being involved in the rest of the research, this aspect was only brought up in few interviews. Also, the question of how to properly cite datasets collected by other was barely discussed. We did not explicitly and universally ask for current practices in data citation and authorship, however.

Furthermore, best practice when dealing with sensitive user data and other considerations of the ethical issues in social media data collection (including documentations of Institutional Review Board (IRB) approval decisions) were rarely shared with the published results. Ideally, publishers of journals and proceedings would take a lead in requesting these details. But some cases where publishers or other bodies did in fact request data and documentation again proved difficult. For example, one interviewee reported that one conference would not accept his paper because a reviewer thought that terms of service might have been violated. As a consequence the author tried to check back with Facebook for clarification – but never received a response. Only few researchers reported that they ever had been required to hand in their raw data together with a manuscript.

Not Enough (Room for) Discussions of Method

We found a frequent perception that current publications did not include enough discussion of methods and ethics in social media studies. Sometimes researchers also admitted that they themselves also did not describe their data collection and processing in sufficient detail (i.e. in a way that would actually allow others to replicate every step). Part of this was due to space limitations. For example, one researcher explained:

> "It's difficult because if you really wanted to document how you got the data, what you did with the data, how you analyzed, how you processed, visualized the data, if you did all that then you'd already have a paper writ-ten before you even get to documenting the outcomes and findings themselves. So in some ways we would almost always would have to write two papers. One is the methods that we used and one is the analysis that we've done."

Another interviewee described that “when I am over on word count in a submission the methods section is the first thing that’s going.” Others agreed with this practice and proposed a separate publication genre or supplementary material especially focused on describing the methods in more detail. Less pressure on publication output (i.e. fewer papers and less focus on rankings) was also seen as a possibility to give researchers the time to truly focus on one project and describe this in sufficient detail.

Of course other factors than publication formats may also play a role in not describing methods in more detail, such as a fear of one’s methods becoming vulnerable to critique.

Alternatives for Dealing with Shortcomings in Publication Channels

The lack of documentation and the inability to learn all necessary details from published papers made social media researchers highly dependent on informal communication and other forms of exchange. Researchers listed summer schools, in-group communication, specialized workshops, and tips and tricks from a senior researcher as important resources to learn about approaches in data collection and analysis. Others had explicitly reached out to authors of published papers via email in order to learn more about the exact methods used.

3.6 Discussion

This entire situation is not necessarily unique to social media research and in fact some researchers also drew comparisons to other fields where similar issues existed. However, given the early explorative stage of social media research and the variety of research methods originating from the high level of interdisciplinarity in the field (as well as little explicitly developed training and education programs for social media researchers), inaccessibility of both data and documented background knowledge may lead to more serious challenges in accessing the validity and quality of research results. What made the situation of research data sharing particularly unique within the social media field was the fact that sharing faced even more obstacles. Data sharing is an ‘intricate and difficult problem’ (Borgman 2012). The increased restrictions posed, for example, by API’s usage agreements and the public attention to social media privacy issues make data sharing even more difficult in the social media field. We found social media data sharing to be fraught with insecurity, uncertainty and aggravation. Yet we also witnessed an intrinsic, often strong motivation to share data for the sake of improving access to social media data and to be able to develop and improve on methods in the fast developing field.

We believe that more formalized approaches and standards are required and possibilities for curation work need to be created in order to facilitate sharing of social media research data to improve validity. In order to advance the field and to allow for the experimentation and transparency that, for example, Karpf (2012) advocates, less formal modes of sharing need to be furthered. Informal sharing can accommodate the variety of approaches and allow for the ephemeral nature of data and contexts. We believe these two approaches to sharing be commensurable; in fact the idea of metadata as static may need to be challenged. Edwards et al. (2011) propose an alternative view of metadata as part of a process of scientific communication rather than as an enduring outcome. If the true potential of Big Data can only be unlocked by sharing social media data across disciplines and by experimenting with methods sharing frameworks need to be flexible and adaptive.

4 Conclusion and Outlook

Challenges around accessing, sharing and publishing social media research (data) are already being addressed in some of the literature as important factors in discussions of methodology and epistemology of social media and Big Data research. However, currently both the actual research data and much of the technical knowledge in social media research remain hidden unless uncovered by studying researchers’ everyday practices. We have conducted such a study which revealed a considerable impact of external constraints on researchers’ attempts to achieve validity and better research quality. Social media researchers’ approaches and practices in general were highly influenced by certain constraints to their work. Constraints mainly concerned the access to data, the sharing of data, and the publication of information about data collection and processing. We thus contribute a better understanding of how researchers deal with access, sharing and publication challenges in their everyday practices to lie a foundation for the consideration of these challenges in further methodological and epistemological discussions of social media and Big Data. It may be concluded that in the current state, both data and knowledge are hidden in social media research. Data is hidden in the sense that research datasets are rather being shared privately than formally published for reuse. More than in other disciplines, use of secondary data is difficult and informal data sharing is connected to many uncertainties regarding legal and ethical questions. Knowledge is hidden in the sense that important background information about data collection technology and research methodology remains unpublished (and sometimes entirely unrecorded by the researchers who document too little even for personal usage). Currently, social media researchers are still to a considerable degree concerned with challenges of data access (based on the technical restrictions applied to proprietary social media data), so that although they are generally aware of the need of benchmarks and strategies of ensuring validity, they often do not yet find themselves in a position to establish the basic standards required. A first step towards this would be more transparency of current practices by fostering better documentation of the technical and methodological background knowledge which is often hidden. Only after this background knowledge is uncovered will it be possible to establish standards for data documentation and data formats. In a next step, criteria for benchmark datasets may be defined, e.g. also considering issues of anonymization. And finally, new challenges such as research data alliances for creating and maintaining such benchmark datasets, licensing models, or citation standards can be addressed. The dimension of better curating datasets through licenses, benchmarks and citation standards will have to be subject to future work and additional interviews.

Acknowledgements

We would like to thank the researchers who participated in our study – we appreciate the valuable insights they shared with us! We would also like to thank our student assistants who helped to transcribe the audio recordings. Thanks are also due to our home institution, the GESIS Leibniz Institute for the Social Sciences for funding "The Hidden Data of Social Media Research" as an internal project.

References

Anderson, C. 2008. *The end of theory: Will the data deluge makes the scientific method obsolete?* Edge [Online], available at http://edge.org/3rd_culture/anderson08/anderson08_index.html (accessed 13. April 2015).

Beninger, K., Fry, A., Jago, N., Lepps, H., Nass, L. and Silvester, H. 2014. *Research using social media; Users' views*. Available at: http://www.natcen.ac.uk/media/282288/p0639-research-using-social-media-report-final-190214.pdf (accessed 19 March 2015).

Borgman, C. L. 2012. *The Conundrum of Sharing Research Data*. J. Am. Soc. Inf. Sci., 63: 1059–1078.

boyd, d., and Crawford, K. 2012. Critical questions for Big Da-ta: Provocations for a cultural, technological, and scholarly phenomenon. *Information, Communication & Society* 15(5):662–679.

Bruns, A. 2013. Faster than the speed of print: Reconciling 'Big Data' social media analysis and academic scholarship. *First Monday* 18(10).

Bruns, A., and Stieglitz, S. 2014. Twitter data: What do they represent? *it Information Technology* 59(5):240-245.

Creswell, J. 2013. *Qualitative Inquiry & Research Design: Choosing Among Five Approaches*. Los Angeles, London, New Delhi et al.: Sage.

Edwards, P. N., Mayernik, M. S., Batcheller, A. L., Bowker, G. C., Borgman, C. L. 2011. Science friction: Data, metadata, and collaboration. *Social Studies of Science* 41(5):667–690.

Frické, M. 2014. Big Data and Its Epistemology. *Journal of the Association for Information Science and Technology* 66(4): 651-661.

Fecher, B., Friesike, S., Hebing, M., Linek, S., Sauermann, A. 2015. A Reputation Economy: Results from an Empirical Sur-vey on Academic Data Sharing. *DIW Berlin Discussion Paper*, No. 1454, available at: http://www.diw.de/documents/publikationen/73/diw_01.c.497416.de/dp1454.pdf (accessed 19 March 2015).

Giglietto, F., Rossi, L., and Bennato, D. 2012. The open laboratory: Limits and possibilities of using Facebook, Twitter, and YouTube as a research data source. *Journal of Technology in Human Services* 30(3-4): 145-159.

Karpf, D. (2012). Social science research methods in Internet time. *Information, Communication & Society* 15(5):639-661.

Kinder-Kurlanda, K. E., and Weller, K. 2014. "I always feel it must be great to be a Hacker!" The role of interdisciplinary work in social media research. In *Proceedings of the 2014 ACM Web Science Conference WebSci'14*, Bloomington, IN, USA, 91-98. New York: ACM.

Kramer, A., Guillory, J., Hancock, J. 2014. Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences* 111(24): 8788–8790.

Lazer, D., Kennedy, R., King, G., Vespignani, A. 2014. The parable of Google Flu: Traps in Big Data analysis. *Science* 343 (14 March): 1203-1205.

Morstatter, F., Pfeffer, J., Liu, H. 2014. When is it biased? Assessing the representativeness of twitter's streaming API. In *23rd Conference on the WWW*, 555–556. New York: ACM.

Polanyi, M. 1966. *The Tacit Dimension*. Garden City, New York: Doubleday.

Puschmann, C., and Burgess, J. 2013. The politics of Twitter data. *HIIG Discussion Paper Series* No. 2013-01.

Ruths, D. and Pfeffer, J. (2014). Social media for large studies of behavior. Science 346(621):1063-1064. Schroeder, R. 2014a. Big Data and the brave new world of social media research. *Big Data & Society* 1(2):1-11.

Schroeder, R. 2014b. Big Data: towards a more scientific social science and humanities? In Graham, M., Dutton, W.H. (eds) *Society and the Internet*, 164–176. Oxford: Oxford University Press.

Taylor, L., Schroeder, R., and Meyer, E.T. 2014. Emerging practices and perspectives on Big Data analysis in economics: Bigger and better, or more of the same? *Big Data and Society* 1(2): 1–10.

Tinati, R., Halford, S., Carr, L., and Pope, C. (2014) Big data: methodological challenges and approaches for sociological analysis. *Sociology* 48(4):663-681.

Weller, K., Kinder-Kurlanda, K. E. 2014. "I love thinking about ethics!" Perspectives on ethics in social media research. Presentation at Internet Research (IR15), Daegu, South Korea, 22.-24.10.2014.

Zimmer, M. 2010, "But the data is already public": On the ethics of research in Facebook. *Ethics and Information Technology* 12(4):313-325.