A Manifesto for Data Sharing in Social Media Research

Katrin Weller
GESIS Leibniz Institute for the Social Sciences
Computational Social Science
Phone: 0049 (0) 221 47694 472
katrin.weller@gesis.org

Katharina E. Kinder-Kurlanda
GESIS Leibniz Institute for the Social Sciences
Data Archive for the Social Sciences
Phone: 0049 (0) 221 47694 449
katharina.kinder-kurlanda@gesis.org

ABSTRACT
More and more researchers want to share research data collected from social media to allow for reproducibility and comparability of results. With this paper we want to encourage them to pursue this aim – despite initial obstacles that they may face. Sharing can occur in various, more or less formal ways. We provide background information that allows researchers to make a decision about whether, how and where to share depending on their specific situation (data, platform, targeted user group, research topic etc.). Ethical, legal and methodological considerations are important for making this decision. Based on these three dimensions we develop a framework for social media sharing that can act as a first set of guidelines to help social media researchers make practical decisions for their own projects. In the long run, different stakeholders should join forces to enable better practices for data sharing for social media researchers. This paper is intended as our call to action for the broader research community to advance current practices of data sharing in the future.

CCS Concepts
• Human centered computing→Collaborative and social computing→Social Media.

Keywords
Reproducibility; methodology; social media; archiving; data sharing; data archives; privacy; data protection; legal issues

1. PERSPECTIVES ON SOCIAL MEDIA DATA SHARING
Social media platforms are an interesting source of research data for researchers across various disciplines. This includes, but is not limited to, computer science, media and communication studies, library and information science, social sciences such as political science or sociology, psychology, linguistics, cultural studies, physics, education, economics, and medicine. Interest in Facebook, Twitter, blogs or Wikipedia as data sources was sparked somewhat independently in different communities that then started to apply idiosyncratic methods and tools to study social media users and/or social media platforms. Consequently various social media researchers with the most diverse research topics make use of (big) datasets directly collected from social media platforms. A growing body of publications that originate in different communities are published in different journals and conferences, apply different methods for data collection, data analysis and evaluation, and may even define social media in slightly different ways.

This particular constellation leads to a highly interesting scenario, in which there is freedom to explore new research questions and to approach them from various angles. Some of the researchers who study social media explicitly value the freedom and creativity in this area with relatively little constraints and standards [18]. On the other hand, there are more and more efforts to take social media research to the next level, e.g. by taking critical perspectives in big data analyses [2; 3; 10; 27; 28] or by uncovering details about the nature and quality of data obtained through the APIs of social media platforms [4; 22; 37].

In light of this situation, there are three main reasons why data sharing has become a crucial next step for the advancement of social media research:

1. To support validity by advancing reproducibility and comparability: reproducibility of research results is an important requirement for achieving validity in many scientific disciplines. Even in those domains where reproducibility is not required or even desired, it needs to be made transparent how results were generated. Access to the original data used for a study is often necessary in order to test reproducibility. Also, comparative studies are depending on the availability of data from previously published works.

2. To avoid 'digital divides' in data accessibility: answering meaningful research questions depends on the availability of suitable research data. Currently, individual connections to social media platform providers often pave the way towards access to interesting datasets. The 'Big Data poor' with no connections or generous funding to buy datasets are left behind [2]. The more datasets are being shared, the more likely it is that a researcher will get access to the dataset that is most suitable to answering a specific research question.

3. To save time and money in data collection processes: no-one knows how much time and effort currently goes into the collection of datasets that other researchers already possess. For example, for the US presidential election in 2012 there are at least 17 research papers that use different datasets collected from Twitter [40]. If at least some of these individual efforts instead had been working towards the creation of a universal and shared dataset, the benefit for the research community could have been multiplied.

Making research data accessible for reuse is in line with initiatives calling for a more 'open science' [26], which also include programs for enabling 'open access' to publications. Such efforts are often supported by funding agencies who increasingly insist on publications and research data to be shared publicly.

WebSci '16, May 22-25, 2016, Hannover, Germany
ACM 978-1-4503-4208-7/16/05.
DOI: http://dx.doi.org/10.1145/2908131.2908172

In some disciplines there is already a long tradition of sharing research data through specialized archives. For example, in the social sciences (and in empirical social research in particular) there are several international data archives devoted to the digital long-term preservation and dissemination of research data, often survey data. Conducting large surveys is very costly, especially if it is done in several waves or over long periods of time. As a consequence, it is desirable to enable more than one researcher or team to make use of the data. Using the same dataset to answer more than one research question is desired – and researchers in the social sciences are used to the practice of accessing data archives for receiving high quality data that they can then use for their specific questions, often in novel combinations with other datasets [26]. Linguists are also used to having access to reusable datasets, so-called linguistic corpora (e.g. [5]), which may also serve as standards or benchmarks for new research.

The lack of social media data sharing initiatives has been noticed before. For example, boyd and Crawford have already prominently criticized the unequal chances of getting access to social media data: "those with money – or those inside the company – can produce a different type of research than those outside. Those without access can neither reproduce nor evaluate the methodological claims of those who have privileged access" [2]. Social media platform providers play a crucial role in this situation, as their policies affect data access and may shape research opportunities – as for example illustrated by Puschmann and Burgess for Twitter [25]. Giglietto et al. also demonstrate the limitations of data available from different social media platforms (Facebook, Twitter, Youtube) [11].

An environment that is conducive to the sharing of research data depends both on the availability of infrastructures to support archiving and data sharing, and the willingness of researchers to make their data available through this infrastructure. There have been several attempts to study researcher behavior when it comes to the obstacles for data sharing [1; 8]. Obstacles range from lack of incentives to the fear of opening one's own research to attack. One of the most crucial findings is that researchers do not share their data because offering it in a usable way usually requires additional work, personal effort and maybe even acquiring new knowledge in the areas of data documentation, preservation and curation [1; 9]. As we could show in earlier work there is often a high willingness amongst social media researchers to share research data [18] – however the main obstacle in this context appears to be lack of clarity about which type of data may be shared and how this can be done. In order to reduce the effort of sharing data for social media researchers we aim to make two initial contributions with this paper: We summarize current practices of sharing social media data as guiding examples (section 2) and we provide a framework for strategies on data sharing based on ethical, legal and methodological considerations (section 3).

2. CURRENT STRATEGIES FOR SHARING SOCIAL MEDIA DATA

A prerequisite for social media data sharing is that the data is archived and managed first. Research data management is a process that may or may not follow standard procedures for documenting and preserving primary data (such as data management plans or structured metadata systems). Data management involves various activities throughout a research project. It can happen more or less publicly: data management can be useful on the individual level to enable a researcher to keep track of their own data and results, activities can then be extended to groups of collaborating researchers, and finally could even be broadened to occur in a public data sharing setting.

To many social media researchers, data management principles may not yet be very well-known. With no clear guidance on how to archive and share social media research data, different approaches have been improvised by the community – with varying success and different degrees of formality. The approaches mainly differ in terms of (a) size and range of available datasets, e.g. single datasets vs. entire collections of (similar) datasets, or samples of data from one social media platform vs. full data for a given platform, (b) the different stakeholders behind the approach, e.g. individual researchers, publishers, or archiving institutions, and c) formality of the data sharing approach e.g. in terms of data documentation standards, access rights, or guaranteed long-term availability. Examples for current approaches include:

- Social media datasets are being *shared informally* on what can be described as a grey market for social media data. For example, researchers with a lab setup that allows for easy, often opportunistic collection of social media data may share datasets which they are not using themselves [41]. Such approaches often lack sufficient documentation of how the data was collected and processed. Consequently, some researchers state that they would rather not use datasets collected by someone else, as they would not be able to judge the quality. Others happily receive such datasets, especially if they do not have the skills to collect data themselves [41].
- Some researchers have made *single datasets publicly accessible via their own websites*. One example for such an approach is the paper by Hadgu and Jäschke [12]. Another example is the data used by Cha et al. [6]. This dataset used to be available on the authors' website – but now the website only features a statement that "Based on Twitter's explicit request, we are only sharing the anonymized topology of the Twitter social network. Please understand that we are not allowed to share any tweet information" [23]. This demonstrates how such approaches may lack in persistence over time. While in this particular case the intervention of a social media company caused the failure of a particular sharing approach – discontinuity may also simply be caused, for example, if a researcher no longer maintains his or her website.
- In addition to publishing datasets on personal websites, there are also cases of *researchers who make datasets available through the Internet Archive*. One example is a set of tweets related to the shooting of Michael Brown in Ferguson, MO, in August 2014 [32]. In principle, this should enhance long-term availability.
- In rare cases, researchers have already attempted to make available entire *collections of datasets*. In the case of CrisisLex datasets that share the same thematic context, namely crisis communication, were made available [7]. In the case of KONECT datasets that share a specific format, namely network structures [19] (which in this case includes but is not limited to social media data), were made available.
- Conference organizers and other publishers of scholarly work have started providing *spaces to share datasets in connection with publications*. For example, the International Conference on Web and Social Media (ICWSM) allows sharing datasets on its website [14]. Authors of accepted papers are asked to provide datasets along with their papers. Other researchers can later request access to these datasets by emailing a usage agreement [15]. The conference also features an award for the best dataset. This provides incentives for researchers to share the data. Also, this approach reduces much of the effort of data sharing – as it is integrated with the conference publication procedures.

- In some cases, conference organizers also provide *datasets for specific research challenges*. The probably most well-known examples are the Twitter datasets that have been provided by the Text Retrieval Conference (TREC) organizers in collaboration with Twitter Inc. for studying microblog retrieval [36]. In this way, different groups of researchers can compare their retrieval algorithms in terms of performance on the same dataset. While this is not an approach that helps individual researchers to share their data, it still constitutes an interesting effort towards comparability of research through data sharing.
- *Dedicated digital data archives* have started archiving social media data. For example, a pilot project archived a Twitter dataset related to the Federal Elections in Germany 2013 in the German Data Archive for the Social Sciences; as a result a set of IDs from all tweets sent by election candidates was archived and is available for reuse [16]. Another social media dataset is in the process of being published [24]. Institutionalized archives, libraries and other continuously funded organizations allow for an archiving that follows recognized standards of digital long-term preservation and documentation. These institutions are involved in developing and harmonizing standards for archiving, access control, documentation, and data management in general.
- *Web Observatories* aim to bring together diverse communities engaging with data science and web-based resources [39]. The aim is to create "a network of separate web observatories, collections of datasets and tools for analysing data about the Web and its use, each with their own use community" [35]. Web observatories are built out of communities with a detailed knowledge of social media data and corresponding methods. They are also currently starting to be more engaged in wider research data communities and organizations such as the Research Data Alliance. The combination of archiving both Web data and tools to collect and analyse those data in the same framework is a particularly comprehensive idea.
- In some cases, *social media companies themselves are providing large datasets* to the public, which may then also be used for research purposes. Wikipedia, for example, regularly publishes so-called Wikipedia dumps, i.e. copies of a Wikipedia version at a specific point in time [42]. However, this example can be considered as an alternative access points to "raw" social media data (in addition to, e.g. public APIs), rather than as an archived research dataset. We would locate the Twitter Archive at the Library of Congress on the same level – if it was already available to researchers: the collaboration between Twitter and the Library of Congress has been announced in 2010 [29], but has not yet led to any usable outcome [20].
- A very unique case was presented in 2015, when a reddit user announced that he had prepared a *complete corpus* of the entire reddit data from 2007-2015 and shared this data via reddit [30] and as a torrent on Archive.org [31].

While some of these approaches may work well for a specific purpose, there are still a lot of unsolved issues. Social media data sharing approaches often do not follow any broader or long-term agenda and they are usually not grounded in the general principles that underlie institutionalized data archiving and sharing: data are thus not presented in a way that advances the social media research field or reproducibility more generally; data are mostly not presented in a sustainable format that can be referenced consistently (e.g. via a DOI) and that is guaranteed to be available over time; and datasets lack (standardized) metadata, i.e. detailed information about how it was collected and processed that would allow for new projects using the same dataset.

The presented list does not claim to be complete – there may be additional examples which we have not listed. Some of them are being summarized by [34]. Furthermore, we have to point out that most of the approaches are focusing on Twitter data, while other social media platforms are still under-represented.

We believe that this list is a useful illustration of first strategies towards social media data sharing. But in a next step, we will have to consider different perspectives on sharing that should be considered before deciding on a specific sharing approach.

3. A ROAD MAP TO SOCIAL MEDIA DATA SHARING

In the following we describe three perspectives that are critical for successful data sharing in social media research: methodological, legal and ethical perspectives. For each of the three perspectives our aim is a) to outline current frameworks that need to be acknowledged by individual researchers to better understand the parameters of their own sharing decision and b) to highlight open challenges or questions that need to be addressed by the research community and additional stakeholders (e.g. archival institutions, publishers) in the future. The latter is especially important, as in most areas crucial questions still remain unresolved (and cannot be resolved without broader discussions in the community).

The three perspectives are based upon our experiences in working in a digital data archive combined with results from a research project that investigated social media researchers' methodological issues and data management practices [18; 41].

3.1 Methodological perspective: Documentation and metadata for reproducibility

In addition to the information about data and methods provided in a publication other information is required in order to be able to comprehend how exactly a piece of research was performed. In particular, new and exploratory research designs may require access to various additional information in unusual formats in order to retrace individual steps taken in the analysis. In the case of social media data user-friendly documentation may include e.g. explanations about the hashtags used for collection or lists of accounts [17], which also need to be archived in order for the research to become reproducible. Detailed documentation of the dataset and the provision of code and syntax allow everyone to check how collection, cleaning and analysis were performed. Additional information on the collection setup and, for example, potential sources for data loss (such as server outages or rate limits) would be useful as well.

Ideally, researchers would need to agree on documentation standards for social media datasets: which basic information needs to be documented in order to understand how a dataset was collected or in order to reproduce another dataset with the same parameters? Documentation standards could be different for each possible platform that data can be collected from and could also act as guidance for methods sections in publications.

3.2 Legal perspective

3.2.1 Data protection legislation

Data sharing must occur in accordance with legal requirements. These may pertain to various factors and be different depending on where the repository is located. For example, data protection legislation across Europe, while becoming more and more harmonized is still different in different countries. Even more legal frameworks may need to be considered on a global scale.

Archiving solutions offered for research data, for example by the UK Data Service or the GESIS Data Archive in Germany, currently make available data in different access categories, including particularly secure ones to ensure that the privacy of study participants in protected, as is demanded by data protection legislation. Consequently, data may be openly accessible to everyone without registration, accessible only after registration, restricted to be accessed only after the data depositor has accepted a request, accessible only after an embargo period, or even only accessible in safe rooms and after signing complex usage agreements. Access restrictions have the potential to control who has access to what data, while still striving for user-friendliness. In social media research, researchers would currently have to figure out which setting may be most appropriate for their dataset – also in light of platforms' terms of services and user privacy, as we will see below. Discussions about best practices are necessary.

3.2.2 Usage agreements and terms of service of platform providers

Social media research is per se closely tied to the commercial companies that run social media platforms. Consequently, the companies' instructions on how data may be accessed, collected and used play a major role in all phases of social media research. While usage agreements or terms of services can be of different specificity, they mostly aim to restrict access to the data and some also specify restrictions on how data can be shared. For example, Twitter data may not be shared in the form of full text tweets with their metadata (e.g. in JSON) – but only in the form of lists of tweet IDs [38] (which is largely respected in the approaches listed in section 2, but again has implications on the usability and persistence of the data, see [33]). In some countries additional challenges may emerge if data sharing is prohibited but if there are other legislations that require research data to be made accessible.

An ideal next step would be to establish a direct dialogue between companies and researchers on the exact interpretation of terms of services and the implications this has for data collection and data sharing – with the aim to establish feasible interpretations that allow researchers to at least share data for the sake of quality control and reproducibility. While social media companies are interested in retaining control over the data as the data itself is part of their business model, they may also benefit from higher quality social media research. It could count as an advertisement for their platform on one side and lead to insights about their users on the other side.

3.2.3 Copyright

When sharing data a researcher also needs to consider the copyright of the dataset. Data sharing systems such as datorium, a sharing platform for social science research data which enables researchers to deposit, document and publish their data [43] allow choosing a license for the dataset, preferably one of the creative commons licenses. Thus, researchers can ensure e.g. that there is an attribution of their work. In the case of social media data they may not hold the copyright themselves or copyright may not apply. If sharing occurs at an archive or other institution staff may be able to advise on this issue. This means, of course, that specialized archival institutions which host the respective expertise are needed and should be strengthened in the future.

3.3 Ethical perspective

The requirements for sharing derived out of an ethical perspective are concerned with how data should be shared in order to comply both with general ethical decision making and with internet research ethics in particular. Although legal and ethical questions may be closely related in many cases, ethical requirements also go beyond the legal requirements as not everything that researchers are allowed to do is something that they should do. While other factors may come into play, we here focus on ethical decision making in the face of two, often incommensurable aims: ensuring users' privacy in the face of often lacking or questionable informed consent, and allowing reproducibility and peer-review to ensure 'good research'.

3.3.1 Privacy expectations of the content generators

Obtaining informed consent to use content generated by users of a social media platform is often difficult or even impossible. While users may have formally agreed to their data being used as stated in the platform's terms of service they may not actually be aware of being observed by researchers [13]. It thus becomes even more important to take appropriate measures to ensure that users' privacy is protected. Anonymization of social media data would be desirable in many cases, but is hard to achieve or even impossible. Even in large-scale social media studies where individual users are not of interest there is a high risk that individuals may be re-identified from publications, published datasets or additional material [44]. In addition, ethical decision making about whether and how to share data may depend on the specific context, research topic and user-group [41]. For example, data from vulnerable groups may require different handling than data from celebrities' or politicians' tweets.

There are still no standards for how to handle the lack of informed consent from social media users who may also have certain privacy expectations. Archiving social media data in the form of mere references to the actual content (such as tweet IDs) addresses these issues to a certain degree: If a user removes content from the social media platform this content will also not be retrievable from the shared references. Users therefore retain at least some form of control over their content. However, information is still shared from users who may be unaware of being the target of research. This is why it can make sense to additionally control access to datasets to ensure that only researchers from acknowledged research institutions who can be expected to adhere to common principles of research ethics access the data.

Discussions of research ethics need to continue on a broader level. More research is needed on users' privacy expectations on social media platforms. Ethical considerations should also be addressed in publications based on social media data, for example, reviewers should request information on underlying ethical considerations if applicable.

3.3.2 Peer-review as an ethical requirement

Researchers may feel that they need to share the data, not only to alleviate access inequalities but also to ensure 'good research' is taking place and that the field is being advanced. The requirements of reproducibility may clash with the requirements of ethical sharing – but there is a long tradition of dealing with this issue in data sharing and many ways have been found to ensure transparency and reproducibility without compromising on participants' privacy. Detailed documentation or complex anonymization techniques, synthetic data and restricted access solutions are but some of them. Even under the current conditions a researcher having performed e.g. a content analysis of a dataset obtained from the Twitter API could mostly fulfil the reproducibility requirements by sharing the tweet IDs at a conference website or in a repository, making available code for accessing, cleaning, processing and analyzing the data in a repository, for example via GitHub, and explain her/his methodology in as much detail as possible in publications.

3.4 So: How much should I share?

From the perspective of reproducibility and ethical obligations connected to it, sharing as much information about a research project as possible would be advantageous – but this may clash with the requirements for sharing legally and ethically (see fig. 1). Detailed data documentation can help to alleviate this dilemma, but some issues remain. For example, sharing tweet IDs only – as is required by the terms of service of the Twitter API and by ethical considerations with regard to user privacy and lack of consent – will always limit the degree of possible reproducibility. However, first insights have already been offered that at least allow to judge how much a Twitter dataset 'deteriorates' over time [33] or how data quality is impacted by different access points to the data [21]. Making a decision on how much and where to share can be taken by defining requirements for the sharing the three perspectives outlined above.

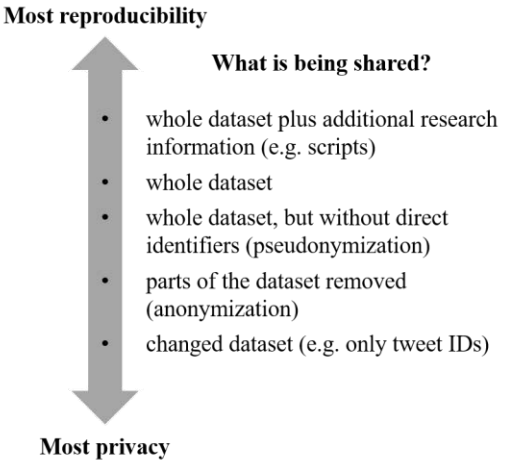

Figure 1. Reproducibility and privacy in data sharing.

4. CONCLUSION

In order to advance social media research as an evolving new field, control over data quality is crucial. Making research data accessible within the research community is a first necessary step to support reproducibility and comparability of social media research. This paper is intended as our call to action for the broader research community to advance current practices of data sharing in the future. As immediate steps we propose the following measures to be approached by different stakeholders involved in social media research:

Individual researchers or research groups:
Researchers should be aware of the positive implications that data sharing can have for the field of social media research and should constantly consider contributing to this, e.g. via the following initial steps:

- Documentation and data management: Researcher should prepare documentation of major sets of data they have collected for their own use. This will be useful for their own work routine – but could also enable to publicly share the data at a later point in time, e.g. once certain legal uncertainties have been solved.
- Sharing internally: For some cases it may be sufficient to share research data within one's own research group or with colleagues. This could inspire discussion and collaboration. It could be done via internal repositories or intranets (sharing via cloud storage may already violate some terms of service).
- Sharing data to enable reproducibility and for advancing the field: In the case of sharing and documenting data for a specific piece of work the easiest way would be to deposit the data in a repository provided by the conference or the journal where the work will itself be published.
  In cases where data should be shared on a broader scale – e.g. to allow for new research questions to be asked of the same data and to enable access to researchers beyond one's core community it may be more suitable to archive data in an archive institute or web observatory.

Based on our interviews with social media researchers [18; 41] and on the various 'grassroots' initiatives to publish social media data as outlined in section 2, we are optimistic that many researchers are willing to take these steps and to share their data.

Archiving institutions
Archiving institutions such as research data archives should continue to address the challenges of long term preservation and unique identifiers for research datasets (e.g. DOIs). Furthermore, they are the ones who should fuel the discussions on:

- suitable documentation practices and metadata standards,
- different models for data access (e.g. embargoes, access to sensitive data),
- practices for anonymization of social media datasets.

Publishers and conference organizers
Publishers and conference organizers can support researchers in sharing datasets by publishing data in addition to written papers. If this is not feasible, they may consider the following steps:

- In the current formation phase of research methodologies and standards in social media research, it is important to also publish non-standard works, such as papers that describe a dataset or a data collection process – or works that use previously published datasets in new contexts or for comparing data quality. Publishers should create spaces for such works.
- Reviewers should be encouraged to check for conformity with legal requirements or ask for the description of ethical considerations, if applicable.

Research associations
Finally, major research associations could act as a collective voice that may start a dialogue with social media companies. Many challenges can only be addressed if such a dialogue takes place. Research associations could also establish guidelines on data sharing with the research community.

We hope that this call to action will inspire fruitful discussions and novel approaches to data sharing in the future.

5. REFERENCES

[1] Borgman, C. L. 2012. The conundrum of sharing research data. Journal of the American Society for Information Science and Technology 63(6):1059–1078.

[2] boyd, d., Crawford, K. 2012. Critical questions for Big Data: Provocations for a cultural, technological, and scholarly phenomenon. Information, Communication & Society, 15(5), 662–679. DOI: http://doi.org/10.1080/1369118X.2012.678878

[3] Bruns, A. 2013. Faster than the speed of print: Reconciling 'Big Data' social media analysis and academic scholarship. First Monday 18(10). DOI: 10.5210/fm.v18i10.4879 .

[4] Bruns, A., Stieglitz, S. 2014. Twitter data: What do they represent? it Information Technology 59(5):240-245. DOI: 10.1515/itit-2014-1049

[5] COCA. no date. The Corpus of Contemporary American English (COCA) Retrieved from http://corpus.byu.edu/coca/ (accessed Feb 12, 2016, archived by WebCite® at http://www.webcitation.org/6fFjBR3mZ)

[6] Cha, M., Haddadi, H., Benevenuto, B., Gummadi, K.P. 2010. Measuring User Influence in Twitter: The Million Follower Fallacy. In Proceedings of the International AAAI Conference on Weblogs and Social Media (ICWSM), May 2010.

[7] CrisisLex. No date. CrisisLex. Retrieved from http://crisislex.org/ (accessed Feb 12, 2016, archived by WebCite® at http://www.webcitation.org/6fFm4G3Jx))

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

[9] Fecher, B., Puschmann, C. 2015. On the limits of openness in science: between aspiration and reality when sharing research data. Information – Wissenschaft und Praxis 66(2-3):146–150.

[10] Frické, M. 2014. Big Data and Its Epistemology. Journal of the Association for Information Science and Technology 66(4): 651-661. DOI: 10.1002/asi.23212

[11] Giglietto, F., Rossi, L., 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. DOI: 10.1080/15228835.2012.743797

[12] Hadgu, A. T., Jäschke, R. 2014. Identifying and analyzing researchers on twitter. In Proceedings of the 2014 ACM conference on Web science. New York: ACM Press, 23-32. DOI:10.1145/2615569.2615676

[13] Hutton, L., and Henderson, T. 2015. "I didn't sign up for this!": Informed consent in social network research. In Proceedings of the Ninth International AAAI Conference on Weblogs and Social Media (ICWSM), 178–187.

[14] ICWSM. 2012. ICWSM Dataset Sharing Service. Retrieved from: http://icwsm.cs.mcgill.ca (accessed Feb 6, 2016, archived by WebCite® at http://www.webcitation.org/6fC7JfFyR)

[15] ICWSM. 2015. Usage Agreement for ICWSM Contributed Datasets. Retrieved from http://www.icwsm.org/2015/datasets/datasets/icwsm_user_agreement_v1.pdf (accessed Feb 12, 2016, archived by WebCite® at http://www.webcitation.org/6fFl9SHLu).

[16] Kaczmirek, L., Mayr, P. 2015. German Bundestag Elections 2013: Twitter usage by electoral candidates. ZA5973 Data file Version 1.0.0. DOI: dx.doi.org/10.4232/1.12319

[17] Kaczmirek, L., Mayr, P., Vatrapu, R. et al. 2014. Social Media Monitoring of the Campaigns for the 2013 German Bundestag Elections on Facebook and Twitter. DOI: http://arxiv.org/abs/1312.4476

[18] Kinder-Kurlanda, K.E., 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 conference on Web Science, 91-98. New York: ACM.

[19] KONECT. No date. The Koblenz Network Collection. Retrieved from http://konect.uni-koblenz.de/ (accessed Feb 12, 2016, archived by WebCite® at http://www.webcitation.org/6fFmJQs4w).

[20] McLemee, S. (2015). The archive is closed. Inside Higher Ed. Retrieved from https://www.insidehighered.com/views/2015/06/03/article-difficulties-social-media-research (accessed Feb 6, 2016, archived by WebCite® at http://www.webcitation.org/6fFldRaKg).

[21] Morstatter, F.; Pfeffer, J.; Liu, H.; Carley, K. M. 2013. Is the sample good enough? Comparing data from Twitter's streaming api with twitter's firehose. In Seventh International AAAI Conference on Weblogs and Social Media.

[22] Morstatter, F., Pfeffer, J., Liu, H. 2014. When is it biased? Assessing the representativeness of twitter's streaming API. In Proceedings of Web ScienceTrack at the 23rd Conference on the WWW, 555–556. New York: ACM. DOI: 10.1145/2567948.2576952

[23] MPI-SWS. no date. The Twitter Project Page at MPI-SWS. Retrieved from http://twitter.mpi-sws.org/ (accessed January 26, 2015, archived by WebCite® at http://www.webcitation.org/6VsuuxQlU)

[24] Pfeffer, J., Morstatter, F. 2016. Geotagged Twitter posts from the United States: A tweet collection to investigate representativeness. DOI:10.7802/1166

[25] Puschmann, C., Burgess, J. 2013. The politics of Twitter data. HIIG Discussion Paper Series No. 2013-01. DOI: http://dx.doi.org/10.2139/ssrn.2206225

[26] Recker, A., Müller, S., Trixa, J., Schumann, N. (2015). Paving the Way For Data-Centric, Open Science: An Example From the Social Sciences. Journal of Librarianship and Scholarly Communication, 3(2), eP1227. DOI: http://dx.doi.org/10.7710/2162-3309.1227

[27] Ruths, D., Pfeffer, J. (2014). Social media for large studies of behavior. Science 346(621):1063-1064. DOI: 10.1126/science.346.6213.1063

[28] Schroeder, R. 2014. Big Data and the brave new world of social media research. Big Data & Society 1(2):1-11. DOI: 10.1177/2053951714563194.

[29] Stone, B. 2010. Tweet preservation. *Twitter Blog* (14 April 2010). Retrieved from https://blog.twitter.com/2010/tweet-preservation (accessed Feb 6, 2016).

[30] stuck_in_the_matrix. 2015a. I have every publicly available Reddit comment for research. ~ 1.7 billion comments @ 250 GB compressed. Any interest in this? Retrieved from https://www.reddit.com/r/datasets/comments/3bxlg7/i_have_every_publicly_available_reddit_comment (accessed Feb 12, 2016, archived by WebCite® at http://www.webcitation.org/6fFpMhWNk).

[31] stuck_in_the_matrix. 2015b. Complete Public Reddit Comments Corpus. Retrieved from

https://archive.org/details/2015_reddit_comments_corpus (accessed Feb 12, 2016).

[32] Summers, E. 2014. Ferguson-tweet-ids. Retrieved from https://archive.org/details/ferguson-tweet-ids (accessed Feb 6, 2016).

[33] Summers, E. 2015. Tweets and deletes: silences in the social media archive. Retrieved from https://medium.com/on-archivy/tweets-and-deletes-727ed74f84ed#.pay32r3eu (accessed Feb 6, 2016; archived by WebCite® at http://www.webcitation.org/6f6KxoikL)

[34] Thomson, S.D. 2016. Preserving Social Media. DPC Technology Watch Report. Retrieved from http://dpconline.org/publications/technology-watch-reports

[35] Tiropanis T., Hall, W., Hendler, J., de Larrinaga, C. 2014. The Web Observatory: A Middle Layer for Broad Data. Big Data. September 2014, 2(3): 129-133. DOI:10.1089/big.2014.0035.

[36] TREC. 2011. Tweets2011. Retrieved from http://trec.nist.gov/data/tweets/ (retrieved Feb 12, 2016, archived by WebCite® at http://www.webcitation.org/6W1ZVkk8o)

[37] Tufekci, Z. 2014. Big Questions for Social Media Big Data: Representativeness, Validity and Other Methodological Pitfalls. In ICWSM'14: Proceedings of the 8th International AAAI Conference on Weblogs and Social Media.

[38] Twitter, Inc. 2015. Developer agreement & policy. Retrieved from: https://dev.twitter.com/overview/terms/agreement-and-policy (accessed Feb 6, 2016).

[39] Web Science Trust. No date. Web Observatory. Retrieved from http://webscience.org/web-observatory/ (accessed Feb 12, 2016, archived by WebCite® at http://www.webcitation.org/6fFnJwWSa).

[40] Weller, K. 2014. Twitter und Wahlen: Zwischen 140 Zeichen und Milliarden von Tweets. In R. Reichert ed., Big Data: Analysen zum digitalen Wandel von Wissen, Macht und Ökonomie. Bielefeld: transcript, 239-257.

[41] Weller, K., Kinder-Kurlanda, K.E. 2015. Uncovering the Challenges in Collection, Sharing and Documentation: The Hidden Data of Social Media Research? In Standards and Practices in Large-Scale Social Media Research: Papers from the 2015 ICWSM Workshop. Proceedings Ninth International AAAI Conference on Web and Social Media Oxford University, May 26, 2015 – May 29, 2015, 28-37. Ann Arbor, MI: AAAI Press. Retrieved from http://www.aaai.org/ocs/index.php/ICWSM/ICWSM15/paper/view/10657 (accessed Feb 12, 2016).

[42] Wikipedia. No date. Wikipedia:Database_download. Retrieved from https://en.wikipedia.org/wiki/Wikipedia:Database_download (accessed Feb 12, 2016, archived by WebCite® at http://www.webcitation.org/6fFnfeGKS).

[43] Zenk-Möltgen, W. 2014. Datorium: Benefit from Data Sharing. Presentation at IASSIST 2014. Retrieved from http://www.iassistdata.org/conferences/2014/presentation/3834 (accessed Feb 12, 2016).

[44] Zimmer, M. 2010. But the data is already public: on the ethics of research in Facebook. Ethics and Information Technology 12(4):313–325.