Archiving information from geotagged tweets to promote reproducibility and comparability in social media research

Kinder-Kurlanda, Katharina; Weller, Katrin; Zenk-Möltgen, Wolfgang; Pfeffer, Jürgen; Morstatter, Fred

Veröffentlichungsversion / Published Version
Zeitschriftenartikel / journal article

Original Research Article

BIG DATA & SOCIETY

Big Data & Society
July–December 2017: 1–14

DOI: 10.1177/2053951717736336
journals.sagepub.com/home/bds
SAGE

Archiving information from geotagged tweets to promote reproducibility and comparability in social media research

Katharina Kinder-Kurlanda[1], Katrin Weller[1], Wolfgang Zenk-Möltgen[1], Jürgen Pfeffer[2] and Fred Morstatter[3]

Abstract

Sharing social media research datasets allows for reproducibility and peer-review, but it is very often difficult or even impossible to achieve due to legal restrictions and can also be ethically questionable. What is more, research data repositories and other research infrastructure and research support institutions are only starting to target social media researchers. In this paper, we present a practical solution to sharing social media data with the help of a social science data archive. Our aim is to contribute to the effort of enhancing comparability and reproducibility in social media research by taking some first steps towards setting standards for sustainable data archiving. We present a showcase for sharing social media data with the example of a big dataset containing geotagged tweets (several months of continued geotagged tweets from the United States from 2014 and 2015; nearly half a billion tweets in total) through a research data archive. We provide a general background to the process of long-term archiving of research data. After some consideration of the current obstacles for sharing and archiving social media data, we present our solution of archiving the specific dataset of geotagged tweets at the GESIS Data Archive for the Social Sciences, a publicly funded German data archive for secure and long-term archiving of social science data. We archived and documented tweet IDs and additional information to improve reproducibility of the initial research while also attending to ethical and legal considerations, and taking into account Twitter's terms of service in particular.

Keywords

Data archiving, data sharing, ethics, social media data, Twitter, geo-data

Introduction

In May 2016, a group of researchers shared a dataset online which they had compiled from the dating site OKCupid (Kirkegaard and Bjerrekær, 2014). The dataset included information such as usernames, gender, location and sexual preferences of nearly 70,000 users and its publication led to a critical discussion among researchers about several ethical questions such as the lack of OKCupid users' informed consent and the difficulties of anonymization.[1] In particular, it was criticized that the researchers claimed that it was ethically and legally defensible to publicly share the dataset as the data was already published. However, even if a user of an online platform knowingly shares a piece of information by posting it on the platform, Big Data analysis can publicize and amplify it in a way the user never intended or agreed to (Zimmer, 2016). Many of the basic requirements of research ethics – protecting user privacy, maintaining data confidentiality and minimizing harm – are not sufficiently addressed in this scenario. Zimmer (2010) argues that it is our responsibility

[1]GESIS – Leibniz Institute for the Social Sciences, Köln, Germany
[2]Bavarian School of Public Policy, Technical University of Munich, Munich, Germany
[3]Information Sciences Institute, University of Southern California, Marina Del Rey, California, USA

Corresponding author:
Katharina Kinder-Kurlanda, GESIS – Leibniz Institute for the Social Sciences, Unter Sachsenhausen 6-8, D-50667 Köln, Germany.
Email: katharina.kinder-kurlanda@gesis.org

as scholars to ensure that research methods and processes remain rooted in long-standing ethical practices: 'Concerns over consent, privacy and anonymity do not disappear simply because subjects participate in online social networks; rather, they become even more important' (p. 324).

Sharing social media content repurposed for research cannot be divorced from the wider debates about the role of ethical research practices within data science. There is currently a concern that the research practices of data science do not make use of established tools of research ethics regulation or that some practitioners are even rejecting ethics regulations outright (Metcalf and Crawford, 2016). However, as the number of case studies from various disciplines is growing, social media research is also increasingly aiming at improving practices, which includes an – often ethically motivated – strive for better standards for validity (e.g., Fricke, 2014), comparability and reproducibility of research results. Amongst others, issues of representativeness (Boyd and Crawford, 2012; Ruths and Pfeffer, 2014) are being addressed.

With this paper, we contribute to the aim of improving social media research practices by combining two previously distinct efforts: approaches to measure quality and representativeness of big datasets collected from Twitter on the one hand (contributed by Morstatter and Pfeffer), and initiatives to establish a framework for sharing datasets used in social media research through archiving in a sustainable and documented way on the other hand (contributed by Kinder-Kurlanda, Weller and Zenk-Möltgen). In this paper, we are also looking at the issue of data sharing as an area within a research project's lifetime that is of particular concern for many of the debated and as of yet unsolved ethical issues surrounding the use for research of data 'found' on the internet. Data sharing thus may help to address the ethical concerns surrounding reproducibility and 'best practice' research, but may also sometimes be hard to bring in line with ethical issues arising out of the origin of the data as user-generated content.

In an interdisciplinary effort all authors of this paper came together to archive[2] a large-scale dataset collected from Twitter. The dataset was collected specifically to allow for archiving and future reuse and to serve as a reference dataset for geotagged tweets. We explored the challenges when archiving several months of continued geotagged tweets from the United States from 2014 and 2015 (about half a billion tweets altogether). While the dataset was large, there was no guarantee that the data we collected was representative of the population that future researchers might wish to study, for example, all Twitter users or even all users posting geotagged tweets (Ruths and Pfeffer, 2014). Moreover, a collection of geotagged tweets provided by the Twitter API (which employs opaque sampling techniques) was not guaranteed to be complete or to provide a representative sample of all geotagged messages on Twitter (Morstatter et al., 2013).

Nevertheless, geotagged Twitter data is particularly useful for research. For example, it has been used to help first responders gain situational awareness in disaster scenarios (Verma et al., 2011) or to uncover global patterns of tourism travel (Hawelka et al., 2014). There is also potential with this kind of data to learn more about various other spatio-temporal patterns, for example, of biodiversity conservation activities (Di Minin et al., 2015). Knowing the location from which a user is tweeting is also useful to gauge the value of the data she is producing (Morstatter et al., 2014). In addition, the location information allows comparison of Twitter data with other, 'offline' data, for example, socio-demographic variables from censuses or surveys, health data, geographic information, environmental data, etc. Our dataset can also serve as a reference dataset for comparative work, for example, for researchers studying similar characteristics of geotagged tweets with other datasets (other periods of time and other geographical regions) who want to compare their results to existing work. Our dataset can also be used for quality control, for example, to test reproducibility of the existing dataset or to study the impact of platform changes on social media data, as Twitter changed the way geotagged information was created halfway through our data collection. To summarize, there are various possibilities for new research projects to be performed on this data which makes archiving even more desirable. At the same time, the fact that the data is geotagged makes it more sensitive in terms of user privacy, which requires special consideration during a formal archiving process. At present, there are first approaches to archiving Twitter-based datasets that comply with Twitter's Terms of Service, but they are usually not rooted in practical archiving experience and therefore lack measures for long-term availability or documentation. They also do not yet focus on the specific setup of geotagged data. With this paper, we want to fill these gaps by proposing a way to handle and archive geotagged tweets and to make a reference dataset available for reuse. To this end, we archived the dataset in the German Data Archive for the Social Sciences at GESIS, founded in 1960 as one of the first archives for social science data. It specializes in survey data of interest to social and political scientists and has accumulated expertise, tools and networks in this area. The archive recently started archiving social media data, which poses some unique challenges but also the opportunity to make new use of a well-established data sharing infrastructure. A pilot project archived a

Twitter dataset related to the Federal Elections in Germany 2013; as a result, a set of IDs from all tweets sent by election candidates and information about candidates' Facebook profiles was archived and is available for reuse (Kaczmirek and Mayr, 2015). The dataset used in the present case study goes beyond that, as it is bigger and more complex due to the inclusion of geo-information.

As shown in the section titled 'Challenges in archiving Twitter data', there are legal, ethical, practical and technical challenges when archiving tweets (and geotagged tweets in particular) which require novel approaches that deviate from the standard practices for survey data. Our solution for archiving the geotagged tweets balances three requirements: sharing legally and ethically, sharing to allow for reproducibility (e.g., precise documentation) and sharing to allow for novel questions and reuse (i.e. researcher friendly data provision).

First, we will explain some general principles for archiving research data with particular focus on experiences from the social sciences.

Sharing for reproducibility in specialized archives

Research data archives are built on the premise that science is not an individual endeavour; rather, researchers aim to achieve provisional results for others to criticize and build upon. In order for this continued critique to happen, researchers are required to make transparent and understandable the way in which they have come to their conclusions. In the quantitative social sciences, for example, research striving for objectivity is required to be reproducible (Popper, 1959). In addition to a sound methodology section in publications, research data sharing plays a major role in achieving reproducibility of research. It is for this reason that there is a long tradition of specialized archives that facilitate research data sharing in the social sciences and in empirical social research in particular. Conducting large surveys is also very costly and using the data to answer multiple research questions is therefore highly advantageous. However, there are also well-known obstacles to sharing data for reuse or reproducibility such as researchers' fear of opening their research to attack, legal limitations due to intellectual property and data protection issues, and in particular the effort required to prepare the data for reproducibility and reuse. Offering reproducibility always requires much work, effort and knowledge in documentation, preservation and curation of digital data (Borgman, 2012). Specialized data archives thus play an important role in sharing research data. The GESIS archive is publicly funded and offers mainly long-term preservation for digital survey data, which is reviewed, processed and documented to provide easily re-usable datasets to the scientific community following well-established archiving standards for social science survey data (e.g., Freese, 2007; King, 2011). The collection can be accessed via an online catalogue.[3] Over time, archives have developed sophisticated standards and procedures for archiving digital data. For example, there are archiving and data provisioning workflows which require (a) following *documented procedures* in archiving, (b) ensuring that preserved *information is understandable and findable* for researchers and (c) guaranteeing *long-term preservation* of the data (CCSDS, 2012).

The *documented procedures* applied in most archives concern the registration of research data with persistent identifiers, its archiving, distribution and long-term preservation. Archiving follows an established workflow (Schumann and Recker, 2013): First, acquisition takes place; second, processing is performed (e.g., quality control, documentation); third, storage occurs (i.e. preservation actions are taken); and fourth, dissemination and access are provided (e.g., include in catalogue).[4]

To ensure that the documented *information is understandable and findable*, documentation of research data at the GESIS archive occurs in accordance with the Data Documentation Initiative (DDI) metadata standard (http://www.ddialliance.org/). DDI is the most important standard for the description of quantitative social science research data and covers the whole data life cycle (Vardigan et al., 2016). The datorium repository used in this example contains DDI-compliant documentation for the datasets which allows to cite, find and better understand them. We will show below that this documentation is, however, currently still insufficient to document all the details required for reproducibility of a social media dataset.

Long-term preservation refers to the fact that digital information is by no means stable. Its accessibility and understandability depend on storage media, hard- and software environments and formats. In order to preserve digital objects such as the desired documentation of a data collection process for the long term, they need to be constantly altered and, for example, file formats require constant updating (Recker and Müller, 2015). Archives ensure long-term preservation in a continuous process of data curation.

In the past, the GESIS archive focused on providing survey data to social scientists. However, disciplinary boundaries are becoming less important when it comes to the sharing of research data, particularly for researchers who are trying to understand social phenomena on the web or who are working with new and computational approaches (Kinder-Kurlanda and Weller, 2014). A growing number of researchers are

using multiple or mixed methods and want to utilize data of different types and from various primary and secondary sources (Borgman, 2012). Social media data and particularly Twitter data are in especially high demand.

Research with Twitter data

There is a growing body of scholarly publications that utilize Twitter data to study various phenomena. In fact, Twitter has become one of the most frequently targeted platforms in social media research (Tufekci, 2014; Weller, 2015). This phenomenon cannot only be explained by the platform's popularity: While Twitter's active user community is relatively small, a search of the publications database Scopus retrieves more publications about Twitter than about the 'bigger' platforms YouTube and Wikipedia (Weller, 2015). Twitter's APIs make it relatively easy to access user data, Twitter has the advantage of relatively simple privacy settings (Zimmer and Proferes, 2014b), a clear structure and short texts; all of which may have also increased its popularity in research. Some researchers consequently are referring to Twitter as the 'model organism' of social media research – and Tufekci (2014) has outlined the problematic consequences this has for the field. Williams et al. (2013a, 2013b) as well as Zimmer and Proferes (2014a) have conducted meta-analyses of Twitter studies that illustrate their diversity, for example, when it comes to approaches used for data collection, sample sizes or research ethics. There are, however, only a small number of publications that use preexisting datasets (Zimmer and Proferes, 2014a). Most significantly, there is currently no benchmark dataset (in the sense of a 'standard' dataset, for example, for a specific timeframe or topic that fulfills agreed-upon quality requirements) and no possibility to easily relate data and results from one publication to those from another. Frequently, information given about data used for a study does not enable others to even reproduce the same dataset. All this puts the validity and expressiveness of Twitter research into question.

Validity and reproducibility in Twitter research

Validity of Twitter research is also to some degree questioned by the lack of clarity about its representativeness (Boyd and Crawford, 2012). This problem has already been addressed in different ways. Ruths and Pfeffer (2014) urge to consider whether a given dataset is representative of the actual population one wishes to study. Bruns and Stieglitz (2014) describe different levels on which representativeness of Twitter data needs to be considered, the most important distinction being between whether a Twitter-based dataset is representative of all Twitter data or of specific user groups and whether Twitter data can be representative of (parts of) society at all. They highlight how the specific characteristics used to retrieve a dataset (e.g., hashtags) may filter specific types of users from the entire user community (e.g., those speaking a specific language). Busch (2014) also points out how decisions made in sampling and filtering may highlight certain aspects of targeted data and obscure others. Other research has investigated the bias in the APIs provided to researchers by Twitter (e.g., Morstatter et al., 2013). In our case, the completeness of the geotagged tweets data returned from the Twitter Streaming API varied: Over 90% of all geotagged tweets for the United States were returned through the service for the first half of the data (collected in 2014). The second half of the data (collected in 2015), however, indicates a considerable sampling bias which still warrants further exploration and was connected to Twitter fundamentally changing the way it reported its data during this time: It moved from an automated geotag-only feature to including the 'place' feature which allowed users to set their own location. When Twitter switched to the place feature, we received substantially less data with the geotag query.

Understanding (and documenting) underlying biases of collection methods is the first step towards methodological standards in social media research and towards enabling validity and reproducibility. The next important step is to share the data and code used in a research project.

Challenges in archiving Twitter data

Twitter is but one, albeit a remarkable, case for investigating possibilities for sharing and archiving social media data. On the one hand, it is the only major social media platform so far that has made significant attempts to publicly preserve their entire history of data (Stone, 2010), although their agreement with the Library of Congress (2013) has not yet led to any tangible outcome (Zimmer, 2015). On the other hand, Twitter itself currently[5] prohibits sharing of collected Twitter datasets to a considerable degree; the Twitter Terms of Service are the first challenge when archiving Twitter data. The Twitter Developer Agreement (Twitter, 2014) states: 'if you provide Content to third parties, including downloadable datasets of Content or an API that returns Content, you will only distribute or allow download of Tweet IDs and/or User IDs'. Some exceptions are being made for lists of tweets in PDF and excel files,[6] but it is not allowed to provide bigger datasets in their native JSON format for third party usage. Twitter has even demanded that individual researchers

stop sharing Twitter datasets. For example, the dataset used by Cha et al. (2010) was originally provided on the authors' institutional website for download but has since been removed upon Twitter's request (MPI-SWS, 2010). Legal challenges based on the Twitter Terms of Service thus are the first set of obstacles for archiving and sharing Twitter data.

Research ethics. A second set of issues arises out of ethical concerns for users' privacy and their consent to being studied. Particularly, when dealing with large volumes of user generated content, obtaining informed consent to research from the users of a social media platform is difficult or even impossible (Zook et al., 2017). While users may have formally agreed to their data being used as stated in the platform's Terms of Service by clicking 'OK', they may not even be aware of being observed by researchers (Hutton and Henderson, 2015). It thus becomes even more important to take steps to protect users' privacy. However, there is a high risk that individuals may be re-identified from publications, published datasets or additional material (Zimmer, 2010). Researchers working with social media data need to make decisions about how to protect users in ways appropriate for the specific context, research topic and user-group (Weller and Kinder-Kurlanda, 2015). For example, data from vulnerable groups may require different handling than politician tweets.

Technical challenges. Technical challenges to archiving Twitter data are connected to the way the data needs to be shared: Providing a list of tweet IDs as suggested by Twitter (instead of a file containing the complete tweet texts and metadata such as author and timestamp) means that to recreate the dataset, every individual tweet needs to be requested again via its ID from the API – a process referred to as rehydrating. This process requires some additional time and tools during the research process and we offer assistance in our archiving solution (see below). The Twitter Developer Agreement (Twitter, 2014) requires researchers to 'Delete Content that Twitter reports as deleted or expired; Change treatment of Content that Twitter reports is subject to changed sharing options (e.g., become protected); and Modify Content that Twitter reports has been modified'. Sharing only tweet IDs ensures that a user's decision to delete a tweet will also result in that tweet being removed from future and (ideally also) existing data collections. However, sharing only tweet IDs also means that someone who wants to re-build the dataset from the IDs may find that many of the tweets are no longer available. It may not be possible to create the exact Twitter dataset twice and archiving tweet IDs is thus no guarantee for reproducibility of Twitter-based research. In fact, every rehydration may result in a different dataset as more and more tweets are deleted over time. What is more, content may no longer be available once platform functionalities change, for example, once Twitter modifies the way in which geotagging works (see below).

Documentation standards. Archiving Twitter data is also challenged by a lack of established standards for documentation of the different stages of data collection, processing and analysis for social media data. The examples for social media data sharing listed in the section 'Current approaches to sharing datasets collected from Twitter' mostly include little explicit data documentation or definition of what the data's essential properties to be preserved are. One solution to this issue is to extend the well-established standards for metadata and documentation of social science survey data to also be applicable to social media data.

Current approaches to sharing datasets collected from Twitter

Despite the challenges, several approaches for sharing Twitter data with other researchers or with the broader public already exist. The current approaches to sharing datasets collected from Twitter come from the following groups (also see Thomson, 2016):

1. Individual researchers or projects provide the datasets they have used for publications on their own website, on university websites or through third party platforms.
2. Conference organizers and other publishers of scholarly work provide the datasets used for publications. For example, the ICWSM conference has been including datasets for accepted conference papers since 2012.[7] Users request access by emailing a usage agreement.[8]
3. Third parties, commercial companies or individual laymen make available datasets via their websites.
4. Libraries and (web) archives provide sets of collected tweets. The GESIS archive's social media election data (Kaczmirek and Mayr, 2015) and the geotagged tweets in this paper are some of the first examples of this.

While most of these examples already allow for secondary use of Twitter datasets, data is not yet presented in a way that advances social media research more generally. Datasets are not presented in a sustainable format that can be referenced consistently (e.g., via a digital object identifier (DOI)) and that is guaranteed to be available over time. Most fundamentally, they also usually lack detailed information about how data was collected and processed. Documentation standards or

metadata guidelines (comparable to the DDI standard mentioned above) are not yet available for social media research. Possibly as a consequence of these gaps, available datasets are not frequently re-used in the research community. During a series of expert interviews, it was shown that at least more advanced researchers were reluctant to re-use datasets collected by others if they had no way to judge quality due to a lack of information about how the data had been produced (Weller and Kinder-Kurlanda, 2015). Documentation and dissemination in accordance with archival best practice may improve datasets' attractiveness. Archives decide which 'significant properties' of a digital object are indispensable and therefore need to be preserved for the long term 'to ensure its continued access, use, and meaning, and its capacity to be accepted as evidence of what it purports to record' (Grace et al., 2009). One of several strategies for determining significant properties (Faniel and Yakel, 2001) is a people-centric approach that considers the producers and re-users of data and their requirements. At the GESIS archive, these considerations of past and future data users are paired with a process-centric approach that also takes into account the different steps of the collaborative social science research process (Schumann and Recker, 2013). Required information about the social science research process includes how, when and why data was created and details of how it was processed and analyzed. In the case of survey data intended to be preserved for re-use by quantitative social scientists who want to test hypotheses, it is required that contextual information such as the composition of the target population and the selection of respondents is preserved as well as the data itself (Recker and Müller, 2015). When archiving a geotagged Twitter dataset, we are addressing a different target audience, namely computational social scientists as well as other social media researchers; a different type of data, namely geotagged Twitter data; and potentially more diverse methodologies that we need to make allowances for. Not everyone may want to test hypotheses and 'kludginess' of methods is both common and required in internet-based research (Karpf, 2012). We have to consider the needs of the producers and users of the archived dataset, as well as the overall characteristics of this new type of dataset. In our example, contextual pieces of information that we need to preserve in addition to the tweets themselves are shapefiles[9] and the codes used for collection, cleaning and analysis. The codes together with the documentation also provide information about collection details such as time frames, geographic selection of tweets and assignment of states by geographic coordinates. In other cases, such necessary contextual information may include, for example, explanations about the hashtags used for collection. We will now take a closer look at our specific dataset and its properties before explaining our approach to archiving it.

The dataset

The dataset used in this work consists of geotagged Twitter posts obtained by setting up our own data collection pipeline (see Figure 2). In this section, we outline the methodology and tools used to collect these tweets.

Targeting a specific dataset

To assemble the dataset used in this work, we collected geotagged tweets from within the United States from 1 June 2014 until the end of November 2014 and again for the same time period in 2015. Tweets with geolocation were selected within a geographic bounding box (−128.6, 24.5), (−59, 50) and then allocated to US states using a script. The tweet location was provided by the users who tagged their tweets using Twitter's 'geolocation' feature: a feature that allows the user to leverage the latitude/longitude information from the device's GPS sensor to add location information to the tweet. The geolocation could also be derived from the wifi or IP addresses in the case of a tweet being authored on a desktop or laptop computer. While wifi provides a very accurate location with an average error of 40 centimeters,[10] an IP address is less accurate[11]; but IP address is rarely used in these cases. The geolocation information was thus accurate to the sub–city-block level and was provided in real time. Another benefit of this information source was that it came from the user's current position and not from the user's profile location field. This gave us the benefit of timelines as we did not need to rely on the user to change her location manually in real time. It also gave us the benefit of accuracy. One drawback of this data outlet was that only about 1% of all users chose to geolocate their tweets, making data from this outlet relatively sparse.

A shift in location information

In April 2015, Twitter substantially changed the way users shared their location. Users were still able to post their geolocation through the geotagging feature, but they were now also prompted to tag their location with the 'place' where their tweet was produced.[12] Before the change, the only information that was revealed was the latitude/longitude pair indicating the location on Earth where the tweet was created. After the change, the user chose the name of their location from a dropdown list thus allowing for a potentially richer dataset that could also include the place name (e.g., a coffee shop in New York City), an ID and in the

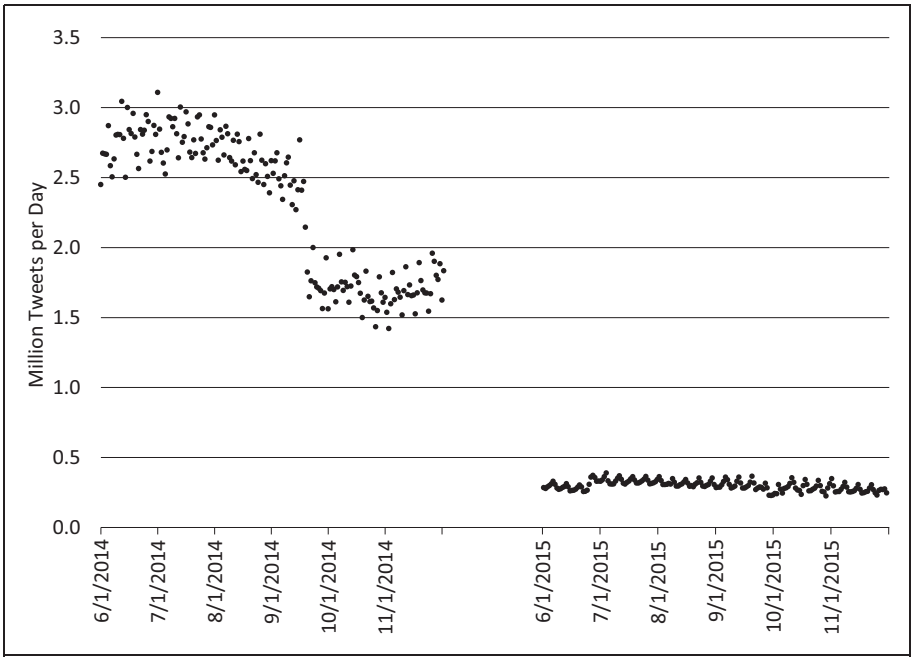

Figure 1. Number of tweets per day in our dataset.

case of larger places (e.g., cities), a geographic bounding box. The downside of the 'place' is that the information is prone to human error, may reflect whimsical decisions and that it is much easier to add 'wrong'[13] locations. Our data collection script only collected the coordinates field (i.e. the geolocation provided by GPS, wifi or IP address) as this was all that was available at the start of the crawl. When Twitter unveiled the 'place' feature, we did not modify the crawler to collect it. However, as one can see in Figure 1, the change in the way Twitter handled locations had a major impact on the number of actual geotagged tweets in our sample. The overall number of tweets with geographic information went up due to the new geotagging feature, and thus the data collection with the API got sampled more dramatically. Previously, collecting tweets with a geographic bounding box ensured a high sampling rate (Morstatter et al., 2013). This was no longer the case. While the change between the first and the second half of the data was obviously due to this platform effect, the smoother change in the number of geotagged tweets happening between mid-August and mid-September 2014 cannot be explained at this point.

Collection methodology

Twitter provides two API endpoints called the 'Streaming APIs'.[14] We collected the data using the Filter API, which provides tweets in real time. The Filter API allows to supply parameters that direct the crawl. The API further allows for three types of parameters: keywords, usernames and geographical regions that we wish to crawl. Geographic regions are provided as geographic bounding boxes in the form of a southwest and northeast latitude/longitude pair.

Once these parameters are specified, the Filter API returns the tweets immediately after they are produced. One caveat is that the Filter API returns a sample of the tweets matching the parameters, at most 1% of all of the data available on Twitter. We leveraged the API to collect geotagged tweets originating within the United States[15] using the Twython[16] library, a popular Python Twitter API interface. All of the collected data was stored in a MySQL database. Figure 2 shows the steps involved in collecting and archiving the data. First, the query with the geographic bounding box is sent to the Filter API. The API returns the sample of the matching tweets to the MySQL database used to store the data. For archiving, we exported all Tweet IDs into separate text files for months and counties. Finally, these were compressed into ZIP-files to archive them in datorium.

Archiving the project dataset

So far we have explored (a) general practices in archiving research data, (b) general challenges of handling and archiving Twitter data and (c) the particular characteristics of our case study dataset composed of geotagged tweets. In the following, we are combining these three dimensions and present our solution for archiving

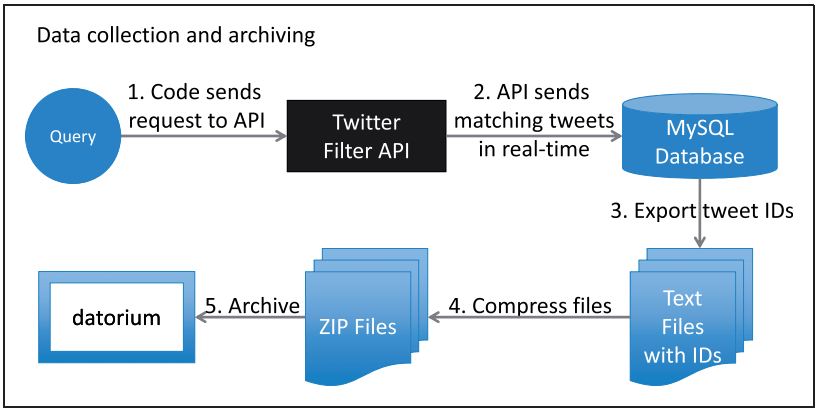

Figure 2. Schematic of the collection pipeline.

the dataset considering requirements from data archiving practice and Twitter specifics. For future researchers to use any archived dataset in a way that allows them to assess its explanatory power in the context of their research question, the dataset and complementary material need to be documented in such a way as to make the exact circumstances of its origination transparent without overwhelming researchers with the level of detail of documentation. These two requirements need to be balanced. A third requirement for archiving is that it occurs in accordance with legal and ethical frameworks, which also means to balance the obligation for transparency and reproducibility with privacy considerations. Out of these three sets of requirements, we can distill the following questions and answers that need to be dealt with when archiving our dataset:

1. What is required for archiving and sharing to occur legally and ethically? We need permission to share all data that the dataset is based on, for example, we need to check usage agreements not only of the geotagged Twitter data but also of any additional data that it is enriched with. In our case, this concerns the shapefiles, but it could be other information, for example, census data. We also need to decide how to contend with the issues around privacy and missing informed consent.
2. What is required for supporting reproducibility? We need to document in detail how the dataset was constructed. For example, this includes information about analysis scripts that show how we controlled for longitude/latitude issues in the dataset. For reproducibility, we also need to archive the tweets exactly in the way in which they were used in a specific study – which may not be the way in which the data should be presented to make it easy to reuse it for other new projects.
3. What is required for allowing new questions to be asked of the data? We need to provide the data in a user-friendly way so that it is easy to download and allows researchers to quickly assess its value for answering different research questions. For example, we may want to provide all tweets during a certain period or from a certain geographical area.

Selecting an archival setup

We decided to archive the dataset via the datorium repository at GESIS[17] datorium is a light-weight sharing platform for social science research data which enables researchers to deposit, document and publish their data (Wira-Alam et al., 2015). It thus allows researchers across the world to upload and share their datasets. The fact that it is hosted by a publicly funded research service institute guarantees that data is stored securely and persistently. For each dataset, datorium guarantees that it will be available for at least 10 years. Each uploaded dataset also receives a DOI which is persistent and unique and makes the dataset citable and accessible. All metadata of the dataset are made available via the DOI registration services da-ra[18] and DataCite[19] by assigning the DOI to the datorium dataset. Our dataset is citable as Pfeffer and Morstatter (2016). datorium uses different access categories for researchers to share the uploaded data. Consequently, data may be openly accessible to everyone, accessible only after registration, restricted to be accessed only after the data depositor has accepted a request or accessible only after an embargo period. We chose to make our dataset accessible upon request after an embargo period.[20] To request the data, researchers need to state, amongst other information, their institutional affiliation and the topic of the planned research. The datorium repository also allows choosing a license for the dataset, preferably one of the creative commons licenses. Thus, researchers can ensure, for example, that there is an attribution of their work or that

usage is for scientific purposes only. Archive staff advise on and eventually approve the level of access and licensing to ensure that sharing complies with legal and ethical requirements. The datorium repository thus allows to control who has access to what data. Archive staff also apply (partially automated) checks of the quality and level of disclosiveness of the data to avoid accidental sharing of sensitive information.

Data aggregation and date/time information

We aggregated information extracted from tweets to the level of US counties, of which there are 3108 in the continental part of the United States (without Alaska, Hawaii and territories). The county-level information can later be aggregated to the state level. The grouping of tweets to counties was based on the tweets' geolocation. Tweets are generated in different time zones. If we want to aggregate and analyze them on a daily basis, we need to make a decision on whether to split the days at midnight in each time zone or for one specific time zone. We decided on the latter as we were just covering tweets from four time zones and events most likely are discussed in real-time within these time zones. As a consequence, we stored all tweets in UTC time zone in the database when we collected the tweets and subtracted 6 hours (= US Central Time Zone) to decide on the day of a tweet. We ignored daylight saving differences in time zones (e.g., Arizona).

Archived data, code and shapefiles

The dataset itself was archived in the form of text files. These files did not include any raw tweets, that is, not the tweet text, user names, timestamps or other metadata usually provided through the API – and also not the explicit geocodes from the tweets. Instead, we archived tweet IDs only in order to comply with the Twitter Terms of Service, which means that anyone who wants to reuse the data will have to rehydrate the tweets based on their IDs. To improve user friendliness, we also archived some basic analysis results, namely aggregated counts of hashtags. Overall, we archived 53 files (48 files with tweet IDs, two scripts and three help files) with about 21 GB of data in zipped files, organized as follows (for [date], we used the month of data collection, six each in 2014 and 2015):

- state_id_[date].zip: Text files per state for one day including IDs of geotagged tweets of the states.
- county_id_[date].zip: Text files per county for one day including IDs of geotagged tweets of the counties.
- state_hash_[date].zip: Text files per state for one day including hashtag counts of geotagged tweets of the states.
- county_hash_[date].zip: Text files per county for one day including hashtag counts of geotagged tweets of the counties.
- state_codes.txt: US states mapped to two-digit codes.

We hence published the tweet ID data in such a way that it easily can be correlated with other data that is available on the county level, for example, census data, health data, environmental data or other measurement data. In order to understand collection and processing of the data, scripts and shapefiles were shared. Specifically, we shared the Python script used for data collection with the Twitter API and the Python script used for sorting geotags. In the documentation of our dataset, we also link to a Github page containing the Python scripts used for collection and analysis[21] to allow for more interaction and feedback between researchers. The county shapefile (with geo-polygone information of US counties) originated from the Unites States Census Bureau website,[22] the other shapefile (with geo-polygone information of US states created from the county file) was created with the Dissolve geo-processing tool in QGIS 2.6.1 Brighton (QGIS Development Team, 2009).

Almost all potential uses of the dataset rely on the 'rehydration' of the list of tweet IDs by the future user. To lower the barriers to using the dataset, we offer assistance for rehydration in two ways: We archived in datorium a 'Python Script to rehydrate Tweets from Tweet IDs'[23] to retrieve tweets from the API. We also link to the 'Hydrator' tool in the documentation.[24] This tool also allows rehydration and its documentation links to more information about the process of rehydration.

A first step towards setting standards for future archiving

We were trying to improve best practice by setting an example, but were limited in what we could accomplish in terms of setting actual standards for archiving social media data at the current point in time. Setting such standards will require involving others, such as, for example, the DDI community, other archives and researchers working with different kinds of social media data. Nevertheless, our sharing solution in the archive satisfied the demands mapped out at the beginning of this section.

Sharing legally and ethically. Sharing of the geotagged tweets occurred with legal requirements and ethical

considerations in mind. As described above, only tweet IDs and aggregated information (i.e. the counts of hashtags) are stored to comply with Twitter's requirements. The datorium system further allows us to retain control over who has access to the data and for what purposes it will be used. There are still no standards for handling the lack of informed consent from social media users who may also have certain privacy expectations. Archiving only tweet IDs addresses this issue to a certain degree: Removing a tweet from Twitter also means that the tweet will be removed from future versions of the dataset as deleted tweets cannot be rehydrated. Users therefore have at least some form of control. However, we are still sharing information from users who may be unaware of being the target of research. This is why we decided to additionally control access to ensure that only researchers who we can expect to adhere to common principles of research ethics access the data. Each access request is decided on individually based on the information provided in the application (e.g., research topic, methods, etc.) Additionally, the geolocation information may pose a particular threat to privacy. Thus, while this feature needs to be explicitly activated by users, we do not provide detailed geoinformation in datorium – it can however still be accessed on Twitter itself once the tweets have been rehydrated. After some deliberation, we decided that in the interest of replicability such a solution would be acceptable, if not ideal. The shapefiles posed no issue as they can be used and shared in publications, as long as the US Census Bureau is acknowledged as their source (which we do). Our sharing solution thus aims to balance privacy requirements and the (ethical) obligation to make research reproducible and comparable. We expect that it will require deliberation and careful consideration for every individual social media dataset shared in the future in order to find similar (and never ideal) compromises that balance conflicting demands.

Sharing to ensure reproducibility and comparability. While not all data is publicly available to everyone in our solution (it is to researchers after request, though), the detailed documentation of the dataset in datorium using the DDI standard and the provision of code allows everyone to check how collection, cleaning and analysis were performed. This applies to geographic and time coverage, technical infrastructures used for data collection and assignment of geographic coordinates to US states. However, more detailed information could only be described in a non-standardized way or not at all due to the lack of metadata standards covering the specific tweet data type and related data collection issues. Examples of such additional information required for reproducibility of geotagged Twitter data are: API biases (i.e. the API would only return a sample of respective tweets upon a specific query); changes in data availability and formats (e.g., Twitter made changes to its geotagging feature in the middle of data collection); or explanations about code and (ready-made) scripts used in collection, cleaning and analysis (in addition to the code itself).

Sharing to allow for new research projects. The data was aggregated into easily downloadable, user-friendly units containing tweet IDs of single months/days. Additional aggregated information was added to allow other researchers to base new investigations upon the data. Standardized metadata (compatible with DDI) for the whole dataset was created within datorium to inform secondary users about the research methods and context. The documentation includes information such as the title of the dataset, the primary researchers, the dataset's availability (restricted), the subject area, a topic classification, the collection period, notes, publications already made using this dataset and other information. A citation suggestion is also included. The usage of a DOI for citation of the exact dataset in datorium enables unambiguous identification of the data and should be considered a major benefit. We also shared a script to assist researchers in the process of rehydration.

Conclusion

We successfully archived a dataset of geotagged tweets that can now be used in various ways: It can be used for secondary research, that is, if researchers want to address new research questions based on the existing dataset. This could, for example, be sentiment analyses of tweets per US state or content analyses of how certain topics have been discussed across the United States. The dataset can also be used to enrich other existing datasets, such as survey data or statistical data (e.g., surveys on internet usage or census data) or data from other social media platforms. It can also be used for comparisons with other, similar datasets, for example, tweets from other periods of time and for testing reproducibility. The dataset is particularly interesting for studying the impact of platform effects on social media data as Twitter changed the way geotagged tweets were handled between the two time periods of our data collection and as we witnessed different sampling effects of the API.

Our dataset will be useful for researchers across several disciplines, both for those who work with social media data to answer specific research questions and for those who want to assess the quality of social media data in general. In addition, this particular dataset also is a use case of archiving that serves the more

general purpose of establishing a routine for documenting and sharing social media data. We show that Twitter data can be archived in the form of meaningful collections of tweet IDs, that is, stored as filtered by region and time and enriched with documentation of collection methods (via shared code, additional data such as shapefiles and metadata about the collection process). Assigning a persistent DOI to the dataset allows us to establish referencing standards. This and the guaranteed availability for at least 10 years underscore the official nature of this dataset. Restricting the access to the dataset as well as storing only tweet IDs helps to secure compliance with Twitter's Terms of Service and users' privacy protection. Although this constitutes significant achievements, there is more work ahead and some open challenges remain.

First of all, we cannot solve the challenge that tweets can only be archived and shared in the form of tweet IDs, which means additional effort to rehydrate the tweets and data loss due to deleted tweets. For a researcher accessing the data via the archive, it is impossible to recreate the original dataset exactly. Unless Twitter changes its policy here (e.g., by allowing to save and share the entire tweet texts for research purposes), full reproducibility of a given dataset may not become possible. In fact, sharing tweet IDs has just become more difficult: Twitter's new Terms of Service, in effect since June 2017, state that it is not allowed to distribute more than 1,500,000 Tweet IDs without the express written permission of Twitter.[25] Had these terms already been in effect in our case, it would not have been possible to share our dataset without applying for an individual permission from Twitter – which they may or may not have granted. It thus has become even more difficult to share data for the sake of research transparency. However, researchers are also in constant negotiation with Twitter about improving data sharing options. For example, Twitter reacted to researchers' concerns over the new Terms of Service by clarifying that it would still be possible for researchers from accredited institutions to share more than 1,500,000 tweets.[26]

Second, possibilities for documentation of existing repositories need to be extended and adjusted to fit both traditional and novel requirements for reproducibility. As data can only be shared in very limited ways, the sharing of other information and materials becomes even more important. For example, additional fields in datorium to allow documentation of API bias, usage of popular scripts and social media platform functionality changes and the points in time at which they occur would improve future social media archiving projects. As a next step, we will propose a draft of a metadata format for social media data archiving to be discussed by the research community and by DDI practitioners in order to set a standard format. In addition to petitioning platform providers to allow more data sharing for research purposes, we see the sharing of additional information and materials as the most promising way to make social media research more transparent and to improve its quality.

Third, while we aimed to protect social media users' privacy, we only achieved this to a certain extent and we could not solve the issue of the lack of informed consent. Our aim to protect privacy on the one hand and to make research reproducible and comparable on the other forced us to compromise on both.

We see our attempt at archiving the Twitter dataset as a step towards establishing archiving best practice for social media data. Due to the lack of standards in this area, we followed well-established archiving standards for survey data. While many of the steps performed in survey data archiving could be applied to a social media dataset (e.g., establishing significant properties to be preserved for the long term), some had to be adapted (e.g., documentation of data collection procedures) as best as possible within the existing infrastructures and conventions. Based on this experience, we suggest the following requirements for future standards for archiving social media data:

To ensure that others can understand or even reproduce the process of constructing the initial dataset, various additional information and materials need to be provided. This concerns programming code (scripts or codes used for collection and cleaning), information on the collection setup (APIs, software, services and hardware used), information about time and place of the collection, information about social media platform functionality changes and information on sampling caused by the way the data is provided to researchers (e.g., the API). In addition, it should be best practice to offer assistance to other researchers who want to reproduce the dataset, for example, by sharing a rehydration script.

To ensure that issues of representativeness can be addressed, any available information on the specific user groups as content providers needs to be documented.

To advance ethically reflective social media data sharing, it needs to be best practice to establish a carefully considered balance between protecting user interests and ensuring research transparency that is also in adherence with the data provider's terms of service. The sharing of information and materials in addition to the data needs to be facilitated to make it easier to find such a balance. Sharing legally and ethically also means to follow the changes and updates in terms of services and policies and to participate in negotiations about data sharing for the sake of reproducibility with platform providers.

Finally, it remains to be seen how much of the practices described here can be transferred to datasets collected from other social media platforms. In the meantime, researchers who study alternative social media platforms may profit from archived Twitter datasets as a source for enabling cross-platform studies.

Acknowledgements

We thank the datorium team and especially Thomas Ebel for their very professional and friendly assistance in archiving this dataset.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

Notes

1. http://openpsych.net/forum/showthread.php?tid=284
2. As explained in the section titled 'Archiving the Project Dataset', we archive the dataset in the *datorium* repository. The dataset was submitted and accepted by the repository in January 2016. Documentation and metadata are accessible under the DOI:10.7802/1166. Data itself was under an embargo until May 2016 when it was opened for access.
3. https://dbk.gesis.org/dbksearch/
4. For more details on the archive's workflows and procedures, see GESIS Data Archive (2014).
5. Twitter's Terms of Service are subject to change over time which may open or close possibilities for sharing research data. We discuss one such change (a limit to the number of shareable tweet IDs) in the text.
6. https://developer.twitter.com/en/developer-terms/policy
7. http://www.icwsm.org/2015/datasets/datasets/
8. http://www.icwsm.org/2015/datasets/datasets/icwsm_user_agreement_v1.pdf
9. Shapefiles include geospatial vector data and are used in geographic information system software to generate representations of geographic data.
10. See https://www.technologyreview.com/s/542561/wi-fi-trick-gives-devices-super-accurate-indoor-location-fixes/
11. See https://support.clickmeter.com/hc/en-us/articles/211035626-How-accurate-reliable-is-IP-GeoLocation-
12. https://support.twitter.com/articles/122236?lang=en
13. A user may add 'London' as the location because she is tweeting about events there, but she could be anywhere else in the world.
14. See Twitter's developer documentation: https://dev.twitter.com/streaming/public
15. The geographic bounding box we used for the United States was the Southwest and Northeast point (−128.6, 24.5), (−59, 50), respectively. Alaska, Hawaii and territories were thus not included.
16. https://twython.readthedocs.org/en/latest/
17. https://datorium.gesis.org/
18. http://www.da-ra.de
19. http://www.datacite.org
20. For our reasoning see below: 'Sharing legally and ethically'.
21. https://github.com/fredzilla/mysql-tweet-crawler-bigdata
22. ftp://ftp2.census.gov/geo/tiger/TIGER2015/COUNTY/
23. The script to rehydrate the tweets can be found at http://doi.org/10.7802/1504 and has the persistent identifier doi:10.7802/1504
24. https://github.com/docnow/hydrator#readme
25. See https://dev.twitter.com/overview/terms/agreement-and-policy
26. See https://twittercommunity.com/t/policy-update-clarification-research-use-cases/87566

References

Borgman CL (2012) The conundrum of sharing research data. *Journal of the American Society for Information Science and Technology* 63(6): 1059–1078.

Boyd D and Crawford K (2012) Critical questions for big data. *Information, Communication & Society* 15(5): 662–679.

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

Busch L (2014) A dozen ways to get lost in translation: Inherent challenges in large scale data sets. *International Journal of Communication* 8(2014): 1727–1744.

CCSDS (2012) Reference model for an open archival information system (OAIS). Recommended practice, No. CCSDS 650.0M2. Available at: http://public.ccsds.org/publications/archive/650x0m2.pdf (accessed 20 May 2016).

Cha M, Haddadi H, Benevenuto F, et al. (2010) Measuring user influence in Twitter: The million follower fallacy. In: *Proceedings of the fourth international AAAI conference on weblogs and social media (ICWSM)*, Washington, DC, 23–26 May 2010, pp. 10–17. Menlo Park, CA: AAAI Press..

Di Minin E, Tenkanen H and Toivonen T (2015) Prospects and challenges for social media data in conservation science. *Frontiers in Environmental Science* 3(2015): 63.

Faniel IM and Yakel E (2001) Significant properties as contextual metadata. *Journal of Library Metadata* 11(3–4): 155–165.

Freese J (2007) Replication standards for quantitative social science: Why not sociology? *Sociological Methods & Research* 36(2): 153–172.

Frické MH (2014) Big data and its epistemology. *Journal of the Association for Information Science and Technology* 66(4): 651–661.

GESIS Data Archive (2014) Data seal of approval assessment report. Available at: https://assessment.datasealofapproval.org/assessment_116/seal/html/ (accessed 20 May 2016).

Grace S, Knight G and Montague L (2009) Investigating the significant properties of electronic content over time. In: *SPECT: Final Report*. Available at: http://www.significantproperties.org.uk/inspect-finalreport.pdf (accessed 20 May 2016).

Hawelka B, Sitko I, Beinat E, et al. (2014) Geo-located Twitter as proxy for global mobility patterns. *Cartography and Geographic Information Science* 41(3): 260–271.

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 web and social media (ICWSM)* (eds Hogan B and Quercia D), Oxford, UK, 26–29 May 2015, pp. 178–187. Menlo Park, CA: AAAI Press.

Kaczmirek L and Mayr P (2015) German Bundestag Elections 2013: Twitter usage by electoral candidates. *GESIS Data Archive,* ZA5973 Data file Version 1.0.0.

Karpf DA (2012) Social science research methods in internet time. *Information, Communication & Society* 15(5): 639–661.

King G (2011) Ensuring the data-rich future of the social sciences. *Science* 331(6018): 719–721.

Kirkegaard E and Bjerrekær JD (2014) The OKCupid dataset: A very large public dataset of dating site users. Open differential psychology [Originally posted at https://osf.io/p9ixw/]..

Library of Congress (2013) Update on the Twitter archive at the Library of Congress. Available at: http://www.loc.gov/today/pr/2013/files/twitter_report_2013jan.pdf (accessed 20 May 2016).

Metcalf J and Crawford K (2016) Where are human subjects in big data research? The emerging ethics divide. *Big Data and Society* 3(1): 1–14.

Morstatter F, Lubold N, Pon-Barry H, et al. (2014) Finding eyewitness tweets during crises. In: *Proceedings of the ACL 2014 workshop on language technologies and computational social science,* Baltimore, MD, 26 June 2014, pp. 23–27. Menlo Park, CA: AAAI Press.

Morstatter F, Pfeffer J, Liu H, et al. (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 (ICWSM)*, Cambridge, MA, 8–11 July 2013, pp. 400–408. Menlo Park, CA: AAAI Press.

MPI-SWS (2010) The Twitter project page at MPI-SWS. Available at: http://twitter.mpi-sws.org/ (accessed 20 May 2016).

Pfeffer J and Morstatter F (2016) Geotagged Twitter posts from the United States: A tweet collection to investigate representativeness. Version: 1. GESIS Data Archive. Dataset. http://doi.org/10.7802/1166.

Popper K (1959) *The Logic of Scientific Discovery*. London: Hutchinson.

QGIS Development Team (2009) QGIS geographic information system. *Open Source Geospatial Foundation*.

Recker A and Müller S (2015) Preserving the essence: Identifying the significant properties of social science research data. *New Review of Information Networking* 20(1–2): 231–237.

Ruths D and Pfeffer J (2014) Social media for large studies of behavior. *Science* 346(6213): 1063–1064.

Schumann N and Recker A (2013) De-mystifying OAIS compliance: Benefits and challenges of mapping the OAIS reference model to the GESIS data archive IASSIST Quarterly 36(2): 6–11.

Stone B (2010) Tweet preservation. In: The official Twitter blog. Available at: https://blog.twitter.com/2010/tweet-preservation (accessed 20 May 2016).

Thomson SD (2016) Preserving social media. DPC technology watch report. Available at: http://www.dpconline.org/docman/technology-watch-reports/1486-twr16-01/file (accessed 09 January 2017).

Tufekci Z (2014) Big questions for social media big data: Representativeness, validity and other methodological pitfalls. In: *Proceedings of the eighth international AAAI conference on weblogs and social media (ICWSM)*, Ann Arbor, MI, 1–4 June 2014. Palo Alto, CA: AAAI Press.

Twitter (2014) Developer agreement & policy: Twitter developer agreement. Available at: https://dev.twitter.com/overview/terms/agreement-and-policy (accessed 20 May 2016).

Vardigan M, Granda P and Hoelter L (2016) Documenting survey data across the life cycle. In: Fielding N, Lee RM and Blank G (eds) *The SAGE Handbook of Survey Methodology*. London: Sage, p. 443.

Verma S, Vieweg S, Corvey W, et al. (2011) Natural language processing to the rescue? Extracting "situational awareness" tweets during mass emergency. In: *Proceedings of the fifth international AAAI conference on weblogs and social media (ICWSM)*, Barcelona, Spain, 17–21 July 2011, pp. 385–392. Menlo Park, CA: AAAI Press.

Weller K (2015) Accepting the challenges of social media research. *Online Information Review* 39(3): 281–289.

Weller K and Kinder-Kurlanda KE (2015) Uncovering the challenges in collection, sharing and documentation: The hidden data of social media research? In: *Proceedings of the ninth international AAAI conference on web and social media (ICWSM), standards and practices workshop* (eds Hogan B and Quercia D), Oxford, UK, 26–29 May 2015, pp. 29–37. Menlo Park, CA: AAAI Press.

Williams SA, Terras M and Warwick C (2013a) How Twitter is studied in the medical professions: A classification of Twitter papers indexed in *PubMed*. *Medicine* 2.0 2(2): e2.

Williams SA, Terras M and Warwick C (2013b) What do people study when they study Twitter? Classifying Twitter related academic papers. *Journal of Documentation* 69(3): 384–410.

Wira-Alam A, Müller S and Schumann N (2015) Datorium: Sharing platform for social science data. In: *Proceedings of the 14th international symposium on information science* (eds Pehar F, Schlögl C and Wolff C), Zadar, Croatia, pp. 244–249. Glückstadt: Verlag Werner Hülsbusch.

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

Zimmer M (2015) The Twitter Archive at the Library of Congress: Challenges for information practice and information policy. *First Monday* 20(7).

Zimmer M (2016) OkCupid Study reveals the perils of big-data science. *WIRED Magazine*. Available at: https://www.wired.com/2016/05/okcupid-study-reveals-perils-big-data-science/ (accessed 20 October 2016).

Zimmer M and Proferes N (2014a) A topology of Twitter research: Disciplines, methods, and ethics. *Aslib Journal of Information Management* 66(3): 250–261.

Zimmer M and Proferes N (2014b) Privacy on Twitter, Twitter on privacy. In: Weller K, Bruns A, Burgess J, et al. (eds) *Twitter and Society*. New York: Peter Lang, pp. 169–167.

Zook M, Barocas S, Boyd D, et al. (2017) Ten simple rules for responsible big data research. *PLoS Computational Biology* 13(3): e1005399. https://doi.org/10.1371/journal.pcbi.1005399.