The practical and ethical challenges in acquiring and sharing digital trace data: Negotiating public-private partnerships

Breuer, Johannes; Bishop, Libby; Kinder-Kurlanda, Katharina

Empfohlene Zitierung / Suggested Citation:
Breuer, J., Bishop, L., & Kinder-Kurlanda, K. (2020). The practical and ethical challenges in acquiring and sharing digital trace data: Negotiating public-private partnerships. *New Media & Society*, *22*(11), 2058-2080. https://doi.org/10.1177/1461444820924622

Special Issue Article

new media & society
2020, Vol. 22(11) 2058–2080
© The Author(s) 2020
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/1461444820924622
journals.sagepub.com/home/nms
SAGE

The practical and ethical challenges in acquiring and sharing digital trace data: Negotiating public-private partnerships

Johannes Breuer, Libby Bishop and Katharina Kinder-Kurlanda
GESIS—Leibniz Institute for the Social Sciences, Germany

Abstract

The ubiquity of digital devices and the increasing intensity of users' interactions with them create vast amounts of digital trace data. Companies use these data to optimize their services or products, but these data are also of interest to researchers studying human behavior. As most of these data are owned by private companies and their collection requires adherence to their terms of service, research with digital trace data often entails some form of public-private partnership. Private companies and academic researchers each have their own interests, some of which are shared, while others may conflict. In this article, we explore different types of private-public partnerships for research with digital trace data. Based on general considerations and particular experiences from a research project with linked digital trace data, we propose strategies for identifying and productively negotiating both shared and conflicting interests in these relationships.

Keywords

Data economy, data sharing, digital trace data, ethics, public-private partnerships, social media, web tracking

Introduction

Social scientists have used a variety of methods to study online behavior, including surveys, participant observation, online ethnography, and interviews. In recent years,

Corresponding author:
Johannes Breuer, Data Archive for the Social Sciences, GESIS—Leibniz Institute for the Social Sciences, Unter Sachsenhausen 6-8, 50667 Cologne, Germany.
Email: johannes.breuer@gesis.org

however, there has been increasing use of digital trace or tracking data. Two types of digital trace data that are often used for research are web tracking and comprehensive social media datasets. Data from web tracking and social media have been used to study diverse topics in social scientific research, including news consumption (Flaxman et al., 2016; Guess, 2015), political discussion (Vaccari et al., 2016), the effectiveness of advertising (Matz et al., 2017), and methodological questions, for example, about the accuracy of self-reported Internet use (Araujo et al., 2017; Scharkow, 2016).

Essentially, there are three ways that researchers can obtain web tracking or social media data: They can (1) collect the data themselves, typically either through web scraping, the use of application programming interfaces (APIs) provided by the platforms/services or the creation of their own dedicated tools, (2) cooperate with companies that produce or hold these data (e.g. social media platforms) to gain privileged access, or (3) purchase the data from market research companies or data resellers. Each of these options has specific benefits and limitations. Importantly, these benefits and limitations extend beyond data collection and affect all phases of the research data lifecycle, including analysis, publication, and archiving or sharing the data.

The three options constitute different forms of public-private partnerships between academic researchers and private companies. In this article, we explore types of private-public partnerships and their implications for research that uses tracking data from the web in general and social media platforms in particular. We specifically focus on the practical and ethical challenges in the following two phases of the research data lifecycle: (1) data collection and acquisition and (2) data publication and sharing. From the perspective of researchers, these two phases are crucial as data access is one of the main challenges in research with digital trace data (see for example, boyd and Crawford, 2012; Thomson and Kilbride, 2015; Weller and Kinder-Kurlanda, 2015). Overall, this article has the following three aims: (1) develop a typology of public-private partnerships for research with tracking data from the web and social media, (2) discuss potential conflicts in these relationships as well as some resolutions, and (3) outline and compare new models for public-private partnerships in this domain. To illustrate some of the arguments we present in this article, we use experiences we have made in a recently completed research project with digital trace data.

Types of public-private partnerships for research with digital trace data

There is a significant amount of literature on the subject of partnerships between universities (or similar academic institutions) and private enterprises. However, it mostly focuses on innovation and knowledge transfer, rather than data access (see for example, Perkmann and Walsh, 2007; Poyago-Theotoky et al., 2002). While some themes are relevant in both contexts (e.g., who controls resources), this article explores the specific challenges that arise over data access, sharing and archiving. We look in detail at different types of data access for research with digital trace data and the types of public-private partnerships they entail. After presenting each of the access options, we compare their major strengths and weaknesses and provide some guidance on how researchers can find the best option for their research.

APIs, web scraping, and self-made tools

Existing literature indicates that collecting data through APIs or web scraping have been the most widespread options for researchers working with digital trace data. Although collecting the data themselves—through APIs, web scraping or bespoke self-made tools—gives researchers the most control over the process, this approach has several downsides that make it unattractive or unsuitable for many researchers and scenarios. Apart from the required programming expertise, a full do-it-yourself solution also means that the software and—depending on the size and scope of the project—the hardware infrastructure behind it (e.g., a proxy server) have to be maintained by the researchers. In the case of web tracking, this approach also usually requires that researchers directly recruit, manage, and incentivize participants. In sum, full research independence (in the sense of not needing to rely on the services provided by commercial companies for collecting digital trace data) comes at a high cost. While it may be feasible for smaller (or pilot) studies, for larger projects this solution requires substantial resources and sustained effort.

When researchers collect the data through APIs, web scraping or self-made tools, the interactions between business and academic partners are typically minimal, and one might even argue that these do not constitute partnerships. However, whether researchers collect data through an API of or scrape data from websites, they have to agree and adhere to the terms of service (ToS) of the platform (and/or its API), and respect company interests with regard to copyright or intellectual property in general. As Halavais (2019) notes, "researchers are bound to follow these terms of service in three ways: state enforcement, enforcement by institutional ethics panels, and embodiment of these rules in the technical infrastructure of the platforms themselves" (p. 4). While the need for ToS-compliance places constraints on what researchers can do with the data, they are largely independent and free to pursue their chosen research questions. Importantly, however, this way of acquiring data is limited to public data, which are often restricted in both volume and variety. Another disadvantage of using public APIs is that they are subject to change and can be unreliable.[1] For many requests, APIs draw samples and—in most cases—it is not known to researchers how sampling is designed or modified (for a more comprehensive critical discussion of using APIs for research see Lomborg and Bechmann, 2014). As the reactions of Facebook in the wake of the Cambridge Analytica scandal[2] show, there is a sizable risk for research that "companies can restrict or eliminate API access at any time, for any reason, and without any recourse" (Freelon, 2018: 665). This risk clearly illustrates that relying on APIs for collecting digital trace data means that the possibility of answering specific research questions depends on decisions made by commercial companies (who generally have different interests than the researchers).

Direct cooperation with data-generating companies

The second option, direct collaboration with private companies, has the advantage that it can potentially provide the richest data in terms of both variety (many different attributes and actions are tracked) and number of cases (as the companies have data from all of their users). These collaborations come in many forms. Two of the most common are contractual agreements between a research institution and a company regarding the provision of or access to data, and the model of the embedded researcher. In the latter, a researcher becomes a consultant, intern or part-time employee of the company and, thus, gets access to internal data. Of course, the details of these arrangements can differ widely, so being an embedded researcher can mean very different things in terms of data access and beyond. A prominent example of direct collaborations between companies and academic researchers is the work by Burke (who is a research scientist at Facebook) and Kraut (e.g. Burke and Kraut, 2016).

A common criticism of this model from researchers, however, is that contractual agreements typically place specific restrictions on researchers' access to and use of data. In addition, the companies might want to have a say in what is and can be done with the data which can, essentially, limit the independence of the researchers. Of course, how independent research in these collaborations can be strongly depends on the individual agreements between the partners. As these agreements are often confidential the independence of research resulting from such partnerships can be difficult to assess. With regard to the specific model of the embedded researcher, Ruths and Pfeffer (2014) also note that "the rise of 'embedded researchers' (researchers who have special relationships with providers that give them elevated access to platform-specific data, algorithms and resources) is creating a divided social media research community" (p. 1063) by proliferating disparities in data access.

Purchasing data

The option of purchasing data from resellers or market research companies incurs substantial monetary costs. When purchasing data, researchers do not need to set up a data collection infrastructure or to recruit participants. However, they typically have limited control over the data collection process, and complete information about this process may not be available. Hence, some researchers are reluctant to work with data that they have not collected themselves as this is the only way they can control its quality (see Weller and Kinder-Kurlanda, 2015). Nevertheless, purchasing data from a market research company or reseller can be a suitable option for many research purposes, especially if data were not otherwise accessible. Examples of studies that bought data from market research companies include the work with web tracking data by Scharkow (2016) and Araujo et al. (2017) or the analysis of large-scale Twitter data by Pasek et al. (2019).

Example case

In a recently completed research project, we opted to pay a market research institute to access data from their web tracking panel for a fixed time period. The reasons for choosing this option were mainly practical. It was the cheapest and quickest way to get individual-level web tracking data for a small project with limited funds and time. Moreover, we assumed that starting from a sample of people who had already agreed to their browsing behavior being tracked would also lead to higher consent rates for linking additional social media data. This hope was, in fact, confirmed as we found much higher consent rates for the linking of social media data than other studies (such as, for example, Al Baghal et al., 2019). The research interest of our project is both methodological and substantive. The methodological focus is on informed consent, data privacy, and data sharing, specifically, the circumstances under which people are willing to share parts of their social media data and to have it linked with web tracking and survey data. We want to develop best practices and recommendations for handling these data, including solutions for sharing it with other researchers, with the goal of finding a good balance between (re)usability and privacy. The substantive focus of the project lies in the exploration of the predictors and effects of different types of online news consumption.

The structure of the web tracking data that we collected in this project, as well as the ways in which we can use these data, are the result of a negotiation process between the researchers and the company.[3] A very positive outcome of some of the early negotiations is that we as researchers have access to the individual-level data. While the company usually delivers these data in an aggregated format (through an interactive online dashboard) to its industry clients, they understood our need for individual-level data and granted us access to it. However, an important restriction of these data is that we get information about only the domains that panel participants visited and not the full URLs. As the substantive focus of the project is the use of online news media (especially for political news), having access to the full URL would have provided us with valuable information, such as which article people read on a particular site. In addition, the full URL would have allowed us to perform a web scraping that allows for a more detailed content analysis of the news articles (or at least of a sample of those). With data available only about the domains that people visited, we can study online news diets or create typologies of users based on the news sources they seek out, but our ability to assess interest in particular topics is limited. The market research company that we work with collects the full URLs but does not make them available as part of their standard data access package.[4] We could have bought them on top of our access to the domain-level URLs. However, as the price for full URLs was quite high and would have exceeded our project budget, we refrained from this option. As mentioned earlier, we also conducted online surveys in the project. Since the market research company took care of inviting the tracking participants to these surveys and incentivizing them for their participation, they also had the opportunity to review each survey. In one of these surveys, we conducted an experiment that was meant to influence the browsing behavior of a subset of participants for several weeks. When we presented this idea to the market research company, they were somewhat skeptical as they sell the data to various industry customers and influencing the browsing behavior of tracking participants might also affect the insights that these other customers could glean from the data. However, after some discussion (both, with us and internally), the company agreed to the experiment. The experiment only involved a small subset of participants, but the main reason the company gave for their decision is that they are also interested in the findings and want to support academic research as they build on results as well as methods generated by basic research in academia to develop and improve their products and services. This was not only fortunate for us but also provides a good example of the ways in which the interests of academic researchers and private data-owning companies be aligned.

In addition to using the data for scholarly publications, we want to share the data from the project through a public repository to enable replications of our analyses as well as other new uses of the data. Since it is difficult and costly to obtain such data, making data available to the research community may also contribute to the reduction of inequalities in data access. Moreover, finding and documenting a solution for sharing this type of data can aid other researchers in sharing their digital trace data. However, we will not be able to share the full raw data. There are the following two reasons for this: (1) Data protection (especially protection of the participants' privacy as, for example, full URLs can contain personal identifiers, such as e-mail addresses or user names/IDs) and (2) business interests (as the market research company sells the data and other companies could use the data shared by us for commercial purposes without paying them). While the business interest is exclusive to the market research company, data privacy is a concern that we as researchers also share.

In a combined dataset including web browsing and mobile app use, sociodemographic information (and other answers) from an online survey, and social media activities, it is potentially possible to identify individual participants even after some common anonymization measures for survey data have been taken. In addition, these data also may include information on "special categories" as defined by the GDPR, as the survey includes questions about political opinions and religion. In general, online content is very difficult to anonymize as it is generated in "increasingly public, archivable, searchable, and traceable spaces" (Markham, 2012: 334). For these reasons, these data will have to be processed before they can be shared. Nevertheless, even after such processing, access to these data will have to be controlled to some degree. Re-use of these data will have to be limited to non-commercial academic purposes, which means that access to these data has to be restricted (an option that many data repositories/archives offer). While the specific solution for reducing the identification risk and controlling access to the data has to be worked out after a thorough exploration of the final linked dataset, both we as researchers and the market company agree that these data would need modifications before they can be shared as both parties value and respect the participants' privacy.

Choosing the appropriate model

All of the types of public-private partnerships for research with digital trace data described earlier differ with regard to the dimensions of control over data collection and quality, the associated monetary costs and effort as well as the required skills, the comprehensiveness and depth of the data, the possibility of making the data available (to other researchers or the general public), and overall independence of the research. Table 1 summarizes benefits and limitations of the three types of partnerships for each of these dimensions from the perspective of academic researchers. The evaluations of the different models presented in Table 1 (as well as the new models listed in Table 2 later on) are derived from the literature cited in this article, one author's experience as an "embedded researcher" doing corporate ethnographies, and the recently completed project with linked digital trace data described above. In addition, we consulted with colleagues from our own institution who are involved in one of the projects that have been granted access to Facebook data through the Social Science One model (see the section on "New models for public-private partnerships for research with digital trace data" for some details on this model).

In practice, it can be difficult for researchers to carefully and systematically weigh the particular strengths and weaknesses of the different types of partnerships. The suitability

Table 1. Benefits and limitations of different types of public-private partnerships for research with digital trace data from the perspective of academic researchers.

	API/web scraping/own tools	Direct cooperation (e.g. embedded researcher)	Purchase data from market research or data reseller
Monetary costs	potential costs for recruitment/incentives or hardware	normally no additional costs	high costs
Required effort and skills	substantial amount of time and technical skills required	depends on the agreement, but typically less than in a full DIY approach	recruitment and/or data collection are taken care of
Control over data collection	depends on available options or documentation	depends on the agreement, but possible conflicts of interest	researchers have to buy data "as is" but many data resellers, for example, offer options for creating bespoke data collections
Comprehensiveness and depth of the data	depends on sample and/or API limitations	potentially richest source of data	depends on how data are collected
Data sharing	subject to ToS	subject to contractual agreements, but typically restricted	subject to contractual agreements, but typically restricted
Independence	only limited by options of the API or tool	companies might want to have a say in what is/can be studied	researchers can choose what data to purchase based on their research interests

API: application programming interface; DIY: do it yourself; ToS: terms of service.

of the options heavily depends on the specific research interest. However, answering a set of relevant questions can aid researchers in finding the best solution for their purposes. The most important question is, "What kind of data do I need for my research?." If, for example, researchers need historical Twitter data, using the public REST API is not a good option because it only allows the collection of a maximum of the 3200 most recent Tweets from an account. In such a scenario, the better option for researchers would be to buy data from a reseller. Generally, it is always a useful first step to assess which options can provide the data you need.

The second question is, "What resources do I have for my research?" Resources include time, money, and skills. The answers to these questions have to be benchmarked against the requirements or costs of the available options. For the example of Twitter data, researchers would have to assess if they (or somebody from their team or group of collaborators) have the necessary skills to collect the data through the API or if they have sufficient funds for buying data from a reseller. If the answer to both of these questions is "no," researchers could investigate whether they could use data from publicly available

Table 2. Key benefits and limitations of the three new models for public-private partnerships for research with digital trace data.

Model	Benefits	Limitations
King and Persily (2018)	• Comprehensive and direct access to data through the company • Companies themselves provide solutions for restricted/safe data access • Data access associated with funding for the research	• Research topics defined by calls issued by the consortium (this also determines what data are accessible) • Composition of commission can influence selection of topics and applicants • Data access only for selected researchers • Companies might want to have a say in the selection of topics, researchers who are granted access or selection committee members
Bruns (2018)	• Research not limited by scope of specific calls (≠ Social Science One model) or collection policies/foci (if archives collect/acquire data) • Data can be accessed by all researchers (given they have the necessary skills to work with the APIs)	• APIs might be restricted or shutdown altogether • Researchers are responsible for data protection as well as archiving and sharing the date responsibly
Data archives as trusted 3rd parties/ intermediaries	• Archives have expertise and experience in the area of data protection • Solutions for restricted/safe data access already in place at most archives • Archives as representatives of specific academic disciplines means more leverage in negotiations with companies than individual researchers • Data in archives potentially available to all academic researchers	• Unlikely that archives can collect or acquire all of the data from a platform (esp. if the platform has a huge number of users and collects large and complex data); focus (e.g., on platforms or topics) could be decided on based on consultations with researchers

archived Twitter data from collections like *TweetsKB* (Fafalios et al., 2018), the Tweet ID Datasets from *Documenting the Now* (see https://www.docnow.io/catalog/) or large datasets stored in data archives (see, for example, Kinder-Kurlanda et al., 2017).

A third question to consider is "Do I want to, or have to, share my data and with whom?" Unless researchers reuse data that is already publicly available, it may be worthwhile or even necessary that they make the data they collect available for other researchers. For example, there may be requirements from funders or the researchers' institutions regarding data sharing, or from journal publishers who are increasingly requiring data to be made available. However, researchers also have to abide by the contractual agreements with companies or the ToS of an API. So one important task is to figure out if or how sharing requirements and agreements with companies are or can be compatible. Finally, even if these obligations can be reconciled, the researchers have to evaluate the reuse value of the data to be able to make an informed decision about the right way to share the data (the topic of choosing an appropriate solution for sharing research data is beyond the scope of this article, but interested readers can consult Klein et al., 2018 for some general guidance).

Shared and potentially conflicting interests in public-private partnerships for research with digital trace data

Public-private partnerships of the kind described in the previous sections are common in the area of research with digital trace data, and without such partnerships, much research would be impossible or at least substantially more difficult to conduct. It is important to note, however, that researchers and companies each have their own interests. Some of them may be aligned, while others may diverge. In the following, we will discuss three areas in which the interests of researchers and companies in public-private partnerships for research with digital trace data may overlap or conflict: (1) data protection and privacy, (2) data quality, and (3) data sharing.

Data protection and privacy

From the point of view of research ethics, there are various challenges to using data that originate out of user interactions with interconnected social media platforms, devices, and services. Ethical decision-making in Internet research is challenged by systems' complexity and changeability, making it hard to define general rules or standards. Rather, the variety of scenarios requires researchers to find solutions inductively and contextually, taking into consideration the specific level of vulnerability of a community or user, and to reevaluate decisions at various steps of a research project (Markham and Buchanan, 2012).

Several common challenges to working with digital trace data may be exacerbated depending on the mode of access chosen. Foremost, researchers need to contend with the fact that while users may have formally agreed to their data being used as stated in the platform's ToS, they may not actually be aware of being observed by researchers, and have not given informed consent to participate in the specific research project at hand (e.g. Hutton and Henderson, 2015; Williams et al., 2017). Absent explicit consent, researchers are obliged to assess how public or private is the original source, the extent of interaction with participants, topic sensitivity, and subject vulnerability (McKee and Porter, 2009). In response to the rapid growth in use of digital trace data, new guidance for how to process and share such data ethically is being published (Van den Eynden et al., 2019).

An approach that entails involving users in making decisions about which data to include in a research project (also see the section on "Data donation by users"), has its own challenges as Lomborg's (2013) research shows. For example, users' understanding of what constitutes sensitive content may be highly idiosyncratic and the researcher, who has an overview of the project, may be better equipped for making sound ethical and analytical judgments about privacy protection. Protecting users' privacy per se is also not easy as traditional anonymization measures, such as deleting profile names and pictures, still allow identifying authors through content using search engines or through linking the datasets with data from other sources.

Various literature has described the way in which Internet companies seek to make data profitable by extending data collection to previously inaccessible domains within private and public life and increasing the scope and detail of user profiles (e.g. Alaimo and Kallinikos, 2017; Couldry and Yu, 2018; Van Dijck, 2014; Zuboff, 2015). However, generally, data protection and privacy are shared interests of academic researchers and the private companies they work with, for both legal and ethical reasons. For example, both academic associations (see, for example, Markham and Buchanan, 2012) and market research companies (ICC/ESOMAR, 2016) have guidelines for ethical research practices and are subject to the same legal regulations (at least if they operate in the same jurisdiction). However, many countries have special laws for scientific (non-commercial) research. Hence, legal compliance tends to be more important for private companies, whereas ethical duties and obligations are more intensely discussed and laid out in formal review procedures in academic research (e.g. through Institutional Review Boards or peer review for scholarly publications).

As stated before, apart from ethical issues, there is also the need for legal compliance with data protection law. For researchers based in Europe, the recent introduction of the General Data Protection Regulation (GDPR) affects how digital trace data can be used. Importantly, the GDPR is also of relevance for researchers outside Europe as many countries have started to introduce similar new data protection laws and companies (including social media platforms) have updated their privacy policies for all users to be GDPR-compliant. What researchers should keep in mind is that GDPR applies only to personal data, that is, data that identify living individuals, and not to anonymized data. If personal data are processed, the following six principles must be respected: (1) lawfulness, fairness, and transparency; (2) purpose limitation; (3) data minimization; (4) accuracy; (5) storage limitation; (6) integrity and confidentiality (for an introduction to GDPR, see https://gdpr.eu/). Moreover, there must be a legal basis for processing. For academic research, this is usually consent of the data subject or public interest of the research.[5] Commercial entities often use "legitimate interest" as their basis for processing. As most types of digital trace data used for research constitute personal data, it is important for researchers to at least have some basic knowledge of the GDPR regulations to be able to assess if or how they apply to the data they collect/use.

Data quality

Another interest that research institutions and data-holding private companies generally share is the quality of the data, as they both want to maximize the information richness of the data. Both want data that are high in resolution, cover a broad range of behaviors, and are representative of the user population. On an abstract level, both academic researchers and private companies are interested in gaining or producing insights from digital trace data. The important difference, however, is that the production of new knowledge is an interest in and of itself for academic researchers, whereas for private companies it is a means to an end to increase profit.[6] In general, companies are more interested in prediction and less in finding and explaining underlying causal or explanatory mechanisms. While the rise of machine learning approaches in the social sciences and psychology have also led to an increased interest in prediction, these disciplines are still predominantly interested in explaining human behavior and its causes (Yarkoni and Westfall, 2017). Also, on a more concrete level, data quality can mean different things for the parties in a public-private partnership. For example, they might have different views on what constitutes the population from which they should/want to sample or if high compliance and fast response rates are a good thing or a sign of potential (selection) bias. Although this is beyond the scope of this article, it might help for the study of public-private partnerships for research to conduct some qualitative/ethnographic research. Such research might help in uncovering potential differences in the understanding and prioritization of different aspects of data quality.

Data sharing

An area where the interests of academic researchers and private companies may conflict is the publication and sharing of research data. While the sharing of research data varies across academic disciplines (including the social sciences, see Zenk-Möltgen et al., 2018), standards such as reproducibility, reusability, and transparency are becoming increasingly important with the spread of the open science movement. To make research findings reproducible, replicable, and robust (LeBel et al., 2018), the underlying data should be made available for other researchers, ideally in the least-processed form possible (Ellis and Leek, 2018). Apart from reproducibility, in the case of (large) digital trace data, there are several other reasons why the data should be shared. The data may, for example, be used as training data for statistical models or algorithms, to build linguistic corpora (in the case of textual data) or for future historical research. Of course, even if they want to be open and share their data, researchers have to balance openness with data/privacy protection (Lundberg et al., 2018).

In the end, there is always a tradeoff between the availability of and access to data on the one side and privacy and data protection on the other side. Concerns regarding privacy and data protection are, in fact, among the most common reasons why researchers are hesitant to share their data (Borgman, 2012; Dehnhard et al., 2013; Fecher et al., 2015; Houtkoop et al., 2018; Vanpaemel et al., 2015). In addition, there are several challenges for data sharing that are unique or at least more pronounced for digital trace data, such as their personal and often sensitive nature or the size of typical datasets (Bishop and Gray, 2017; Kinder-Kurlanda et al., 2017; Van Atteveldt and Peng, 2018; Weller and Kinder-Kurlanda, 2015, 2017).[7] In the case of public-private partnerships, the opportunities for data sharing might be further reduced if the company does not allow the data to be shared or permits sharing only in a specific way (e.g. highly aggregated and/or only on request). Issues of data protection and privacy certainly play a role in these restrictions. However, the companies also want to protect their business assets, and making data publicly and freely available may be a risk as competitors could make use of it or potential customers could use it without paying the company that collected it.

Limitations of current models of public-private partnerships for research with digital trace data

All types of public-private partnerships strike a different balance between independence, required resources, and access to information for the researcher. What is distinctive for these partnerships in the area of digital trace data is that there is an imbalance of power as it is private companies who hold and control the data. This is what causes an ironic situation in which "social scientists have access to more data than they ever had before to study human society, but a far smaller proportion than at any time in history" (King and Persily, 2018: 3). While the volume of potentially available data is continuously increasing, the key question is not only whether or how researchers can access it, but also how useful it is for answering societally relevant questions. In general, data quality is more important for academic research than data quantity. Nevertheless, the current situation, as described in the quote by King and Persily, increases the importance of public-private partnerships for academic research as large parts of the digital trace data that is of interest to researchers is otherwise unavailable.

Up to now, most direct collaborations for research with digital trace data have been based on individual, customized agreements between companies and researchers or institutions. This model, however, has limitations. Companies are more likely to enter partnerships with larger and more prominent institutions that typically have more funds and are more likely to generate publicity.[8] This creates the risk of a division of researchers and institutions into "the Big Data rich and the Big Data poor" (boyd and Crawford, 2012: 674). Importantly, even for the same company, different researchers/institutions might reach different agreements with them, resulting in inequalities in data access or rights of use.

Another limitation has arisen in the post-Cambridge Analytica era as a result of companies' fears about risks they incur when sharing data, even with presumably responsible researchers. Providing largely uncontrolled access to sensitive personal data without making a clear distinction between commercial and academic research was at the core of the Cambridge Analytica scandal that prompted what is essentially a lockdown of the Facebook API for academic research (Bruns, 2018). And recently, Twitter has also announced that it will regulate API access more strictly (see https://developer.twitter.com/en/developer-terms/more-on-restricted-use-cases.html). The consequences of these changes for academic research on these platforms are quite substantial, to such an extent that Freelon (2018) speaks of a "post-API age" for computational research. Such restrictions have an impact on the type of academic research that can be done on the use and effects of social media. The curtailment or closing of APIs further adds to the imbalance in data access, with access to large collections of digital trace data limited to those researchers and institutions that have the financial means to acquire data from resellers or the prestige (and necessary contacts) to enter into direct collaborations with the data-holding companies. However, as Halavais (2019) aptly puts it "a reduction in the range of perspectives, approaches, and backgrounds of those engaged in such work reduces the opportunity for innovative work and establishes the platforms themselves as more substantial gatekeepers of social research" (p. 2).[9] In light of these developments, Freelon (2018) argues for the increased importance of web scraping for computational research as it makes researchers independent from the decisions of commercial companies regarding the availability and scope of APIs. Notably, web scraping is not only more difficult (as typically requires at least some coding skills) but may also violate the ToS of a platform (or website). Although Freelon lays out the benefits of web scraping and suggests that it is more flexible and future-proof than access through APIs, he also notes that researchers should always try to use authorized (ToS-compliant) methods and be aware of the risks associated with violating ToS. At the same time, he points out the difference between compliance with ToS and the protection of participants' privacy. While web scraping is one specific method that can be employed as an alternative to access through APIs, there are several new models for public-private partnerships for research with digital trace data that have been developed or (further) promoted in response to the restrictions imposed on various platform APIs.

New models for public-private partnerships for research with digital trace data

Overall, the tensions inherent public-private partnerships for research with digital trace data have led to calls for new models that avoid, or at least better manage these tensions. In this section, we discuss and compare three new models for public-private partnerships and introduce a potential alternative.

King and Persily (2018)

One of the most detailed suggestions for a new model for public-private partnerships for academic research comes from King and Persily (2018). In their working paper, they first outline a general "new model for industry academic partnerships" and then present an implementation of this model for research using data from Facebook. They propose that a group of academics form a commission that acts as a trusted third party to mediate between the company and researchers. Together with the company, the commission defines topic areas or specific research questions to solicit proposals from academic researchers. A subcommittee of the commission (potentially together with additional external experts) reviews and selects the proposals. The researchers whose proposals are selected receive access to the company's data as well as funding from independent non-profit foundations. Other than in the choice of commission members and topic areas, the company is supposed to have no influence on the research process and its outcomes. The model has first been applied to Facebook (now put into practice as Social Science One: https://socialscience.one/) as its formulation is a direct consequence of the Cambridge Analytica scandal and the following shutdown of most functionalities of the Facebook Graph API. However, King and Persily (2018: 10) present it as a general model that "would still work, with few modifications and without any added difficulties" for "other industry-academic partnerships—such as with smaller companies, in less politicized environments, or in substantive areas where funding from nonprofit foundations is unavailable". The only change they suggest for such scenarios is that the company could directly fund the commission and researchers.

The model proposed by King and Persily (2018) is noteworthy as an alternative to individual agreements between companies and single institutions. While the solution they propose for academic access to Facebook data may be better than no access at all, some of the suggestions are debatable. Although the authors provide a good explanation why the commission should be composed of senior researchers (they cannot apply for data access and funding themselves) and explicitly state that "diversity in methodological approach, substantive area, geographic region, race, ethnicity, gender, ideology, and political party" (King and Persily, 2018: 10) is important for the selection of commission members and research proposals, a key question is how the commission members and the topics or specific research questions are chosen. Depending on the composition of the commission and the (direct or indirect) influence of the company and the nonprofit foundations, there may be biases in the choice of topics and research questions. With regard to the generalizability of the mode, the effort it requires may not be feasible for smaller companies.

Bruns (2018)

A comprehensive critique of Facebook's reaction of essentially closing its Graph API for academic research as well as the model suggested by King and Persily (2018) comes from Bruns (2018). He argues that the platform providers should not "be allowed to position themselves as the gatekeepers for the research that investigates how their platforms are used" (Bruns, 2018: online). The article makes four suggestions how to improve public-private partnerships for research with digital trace data: "1) Straightforward scholarly data access policies; 2) Custom APIs for research purposes; 3) Accept the use of research data repositories; 4) Open and transparent engagement with the research community" (Bruns, 2018: online). With regard to data access, Bruns (2018) argues that it should not be limited to select researchers as this can exacerbate the divide between what boyd and Crawford (2012) have called the Big Data rich and the Big Data poor. An open and transparent engagement with the academic community also means that research with data from social media sites and other online platforms should be open to a variety of disciplines and topics. The point about the use of research repositories is a response to the restrictions that many companies pose on the sharing of data, even when this is done exclusively for the purpose of academic research. Given what happened with the Cambridge Analytica data, the data sharing solution that King and Persily (2018) propose for Facebook is more restrictive. They suggest that researchers can "access minimally necessary data on company infrastructure with specialized, locked-down systems with continuous auditing" (King and Persily, 2018: 9). While this means that the raw data cannot be stored elsewhere or shared, they state that one requirement for funding will be that researchers produce “replication data files” (King and Persily, 2018: 9) and make them publicly available through the Dataverse repository.[10]

Data archives as trusted third parties

As an extension of the arguments presented by Bruns (2018), we suggest that research repositories are ideally suited for the role of a trusted third party in public-private partnerships for research with digital trace data. They have experience with and technical solutions for storing and controlling access to personal and sensitive data (typically called safe havens or secure data centers), and their goal is to maximize the use of research data while also protecting the privacy of individual participants. Specifically, many archives have gradations of access, from open (public and free) to highly restricted (e.g., a physical or virtual safe room where data are only accessed, but not downloaded, and research outputs are inspected for disclosure risk).

Over the last few years, several archives have begun to develop solutions for archiving digital trace data and making it available to researchers. Many archives targeted (mainly) at researchers from the social sciences which have previously focused on survey data are engaged in various efforts to extend their portfolio to include digital trace data. The ICPSR is, for example, developing a separate Social Media Archive (SOMAR; see Hemphill et al., 2018). Other curated archives for the social sciences, such as GESIS in Germany (see Kinder-Kurlanda et al., 2017), have started to integrate datasets containing digital traces into their regular catalogs. Importantly, archives not only work on solutions for storing and documenting digital trace data, but also for dealing with the specific legal and ethical concerns related to working with this type of data (with regard to both user/participant privacy as well as platform ToS). A good example of these efforts is the work package that deals with legal, ethical and quality issues of new forms of data in the SERISS project (see https://seriss.eu/about-seriss/work-packages/wp6-new-forms-of-data-legal-ethical-and-quality-issues/) in which the Consortium of European Social Science Data Archives (CESSDA) is involved.

In additional to the distinctive benefits described earlier, data archives may have more leverage in negotiations with companies than individual researchers (and potentially also more than individual universities, especially if the archives cooperate with scholarly associations to represent the interests of one or several disciplines). Their capabilities in managing, documenting, and storing data can also be attractive to smaller companies that do not have a technical infrastructure for storing/archiving data like Facebook, Twitter or Google. A central institution like a data archive might also be able to better shoulder the weight of creating purpose-built panels for the collection of digital trace data, whether this be through APIs (potentially/ideally with extended access), data donation by the users (see the following section on this) or some other mode of data collection. Similar to the types of public-private partnerships for research with digital trace data that we compared in Table 1, the new models described in this section also have specific strengths and limitations. Table 2 provides an overview of the key benefits and limitations of the three new models.

Data donation by users

Another way of avoiding potential conflicts of interest in public-private partnerships for research with digital trace data in addition to new cooperation models is to find alternative sources of the data, for example, by asking users to donate data. Halavais (2019) describes this as one of the ways of "partnering with users to collect big data" (p. 8). As the phrasing chosen by Halavais (2019) indicates, the partners of the researchers in this model are not the companies that own the platforms and services but the people who use them. Many of the major online services, including Facebook, Twitter, and Google, allow their users to export their data (usually called an archive) with just a few clicks. In many cases, this feature was implemented in response to new legislation, such as the GDPR in the EU. The popularity of the technology and (online) services related to the concept of the quantified self (Ruckenstein and Pantzar, 2017; Sharon and Zandbergen, 2017) can be used to promote research projects that employ such a data donation model and recruit participants.[11] Provided they offer a safe solution for the upload and pseudonymization of the data, researchers can ask users to share their data with them, thus, donating it for scientific research.[12] This approach solves the ethical issue of data being collected without users being aware of it. Instead of being a research "subject," participants become an active part of the research process, which is in the spirit of the popular idea of "citizen science." Of course, this approach necessitates that researchers obtain informed consent from participants and inform them in detail about how their data will be used. This consent still needs to be handled with care as users may not be aware of the full implications of what they are consenting to. Data donation does not free researchers of the ethical obligation to protect users' privacy, especially as they are the ones who have a more comprehensive understanding of the projects and analysis tools used (Lomborg, 2013). Also, while data donation can circumvent data sharing restrictions imposed by companies, it is a novel approach, and methodological research is needed to assess how to best recruit and incentivize participants.

One drawback of this approach is that having people donate their own digital trace data for academic research introduces self-selection bias. As downloading one's personal data is often not that straightforward, relying on data donation from the users might introduce a further bias as tech-savvy participants will be less likely to face problems in accessing their own data.[13] A third source of potential bias is that users might manipulate their data, provided they have the necessary knowledge. However, data acquired through any of the three types of public-private partnerships outlined in the first parts of this article is typically also biased. For example, public APIs normally limit the amount of data that can be accessed and, in most cases, it is not fully transparent to the user how the data are sampled. And although the samples that market research companies provide tend to be larger and more diverse than those recruited by researchers themselves, they are equally affected by self-selection bias, and the recruitment strategy and panel management measures of the company might introduce additional biases. Finally, even if researchers have full access to the complete data of a platform, these data are still biased in the sense that all platforms have different user populations which makes generalizations to Internet users as a whole (or even to other platforms) problematic.

As stated before, downloading one's own data from a social media platform is not always self-explanatory and can take some time (as the files containing one's data typically have to be generated and are not directly available for download). Hence, the data donation model requires that participants are appropriately instructed and, ideally, also offered support. On the positive side, such measures increase the transparency of the research process as participants can see (and control) what data the researchers use (Halavais, 2019). Ideally, this also bolsters trust in the researchers. Nevertheless, given the personal nature of the data and the effort it can take to share them with researchers, it may be necessary to offer participants substantial incentives (that are higher than those typically paid for participation in a survey).[14] Another important point with regard to the data donation model is that it requires solutions for the anonymization of data (e.g., the removal of names from social media posts). Ideally, this should be done automatically, so that the researchers never see the personal data. Of course, this is not trivial and manual work might still be necessary, even if technical solutions are developed that automatically take care of parts of the anonymization process.

Conclusion

Research with digital trace data has become increasingly popular in the social sciences over the last few years. There are different ways researchers can access these data. All of these options have their own benefits and limitations that researchers need to take into account when deciding how to collect data. Notably, the choices made with regard to data collection will have downstream effects on later stages of the research process, such as data publication and sharing. Of course, another important aspect that researchers need to factor in when choosing a data access model is their own situation—most importantly, their particular research interest and their resources (time, money, and skills).

All of the usual access options for digital trace data require some form of public-private partnership between academic researchers and private companies. What characterizes these relationships is that both sides have their own interests, some of which are shared while others may conflict. Our own experiences in negotiating with a market research company show that the differences in interest can be more pronounced when data sharing is discussed, while there is a large overlap in interests regarding data collection (quality and privacy). In order for these partnerships to be productive, it is important that both parties are aware of their own interests and make them explicit to each other as early as possible. Developing solutions for avoiding, minimizing or dealing with potential conflicts of interest can not only increase the efficiency of these partnerships, but, ultimately, also help to improve the ethical standards and quality of research with digital trace data.

The limitations of types of public-private partnerships for research with digital trace data that have been most common so far as well as the restrictions imposed on public APIs by some major social media companies have led to suggestions for new models. While we see that the new models for public-private partnerships proposed by King and Persily (2018) and Bruns (2018) both have merit and present interesting options for ensuring access to digital trace data, we believe that the model of data archives as intermediaries and trusted third parties has unique advantages that make it particularly interesting from the perspective of academic researchers. Another option that might prove useful for future research is direct cooperation with the users in the form of a data donation instead of partnerships with private companies.

The models for and alternatives to public-private partnerships for research with digital trace data we discuss in this article are by no means completely exhaustive, and new ideas are likely to emerge in the coming years. Moreover, in order to validate our typology and test some of its assumptions, we believe it may be worthwhile for future research to answer some of the questions we raise here through empirical work on the methodological practices of researchers working with digital trace data. Given the increase of research interest as well as recent developments on the side of social media and other Internet companies, we believe that public-private partnerships will remain an important issue for research with digital trace data and hope that our typology and the framework for identifying and dealing with potential conflicts of interests can aid researchers in finding the best solutions for their work with digital trace data.

Funding

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

ORCID iD

Johannes Breuer https://orcid.org/0000-0001-5906-7873

Notes

1. What is also important to note in this context is that data change as users change (e.g., because they are deleted, amended, or shared further). This applies to all social media data regardless of how they are collected. In addition, if there are platform changes, this also affects the data that the company running it have.
2. An app called "thisisyourdigitallife," was developed by researcher Alexander Kogan. Users were paid to use the app and agreed to have the data used for academic research. However, the app also collected data about the users' friends, resulting in data on 87 million users. These data, according to the Facebook ToS, could not be sold. Details are disputed, but Kogan transferred the data to the company Cambridge Analytica, where it was used for targeted political advertising (Metcalf and Fiesler, 2018).
3. It should be noted that the experiences described here are not the result of a systematic (auto-) ethnographic study. One researcher took notes at all the meetings and the status and results of the collaboration were discussed at regular intervals among all authors of this article. Hence, the descriptions here are anecdotal and should be interpreted as such. Their main purpose is to provide some concrete examples for the opportunities and challenges that such collaborations can hold.
4. We also only get data about which mobile apps were used and not what participants did within those apps. However, unlike in the case of the URL tracking, the reasons for this are primarily technical as the apps do not allow the extraction of detailed tracking data.
5. One common misperception about GDPR is that consent is mandatory, but this is not the case. Another is that data archiving is prohibited. This is also false as Article 89 of GDPR permits processing of personal data for historical, statistical, and archiving purposes (UK Data Service, 2019).

6. Of course, one may argue that the end goal for individual researchers is to produce publications in order to further their careers. However, what we refer to here are the structures and ideals of science and economy as systems.
7. Because of these unique ethical challenges, several recent publications provide some guidance for researchers who want to collect and/or share (specific types of) digital trace data (e.g. Bishop, 2017; Kinder-Kurlanda et al., 2017; Mannheimer and Hull, 2017; Thomson, 2016; Weller and Kinder-Kurlanda, 2016; Williams et al., 2017).
8. The expected quality of the research output may also play a role here, although—unlike quantity—the quality of research is notoriously difficult to define and measure.
9. Following what essentially constitutes a lockdown of the Facebook API for researchers, Anja Bechmann invited researchers to contribute to a list of publications that would not have been possible without API access (see https://docs.google.com/document/d/15YKeZFSUc1j03b4lW9YXxGmhYEnFx3TSy68qCrX9BEI/edit) to illustrate how important this mode of access is for academic research.
10. Although the paper itself does not specify what these replication files should look like exactly, they are typically anonymized and aggregated data that are sufficient to reproduce the analyses in associated publications (see King, 1995).
11. A platform that completely relies on the donation of users' data is Open Humans (https://www.openhumans.org/).
12. Another option that Halavais (2019) mentions is to recruit participants and ask them to install browser plugins (or other pieces of software) that collect some of their digital traces and make them available to the researchers (for an example, see Haim and Nienierza, 2019).
13. On the contrary, these people might also be less likely to share their data with researchers as they are more aware of what the data contain (and what can be done with them) and/or have higher privacy concerns.
14. One way of motivating people to participate could be to offer them direct (and possibly also interactive) feedback about their data. This can also serve an educational purpose by making people aware of the digital traces the leave and what can be done with them.

References

Al Baghal T, Sloan L, Jessop C, et al. (2019) Linking Twitter and survey data: the impact of survey mode and demographics on consent rates across three UK studies. *Social Science Computer Review* 38(5): 517–532.

Alaimo C and Kallinikos J (2017) Computing the everyday: social media as data platforms. *The Information Society* 33(4): 175–191.

Araujo T, Wonneberger A, Neijens P, et al. (2017) How much time do you spend online? Understanding and improving the accuracy of self-reported measures of internet use. *Communication Methods and Measures* 11(3): 173–190.

Bishop EL (2017) Big data and data sharing: ethical issues. UK Data Service, UK Data Archive. Available at: https://www.ukdataservice.ac.uk/media/604711/big-data-and-data-sharing_ethical-issues.pdf (accessed 30 October 2018).

Bishop EL and Gray D (2017) Ethical challenges of publishing and sharing social media research data. In: Woodfield K (ed.) *The Ethics of Online Research*. Bingley: Emerald Publishing, pp. 159–187.

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: provocations for a cultural, technological, and scholarly phenomenon. *Information, Communication & Society* 15(5): 662–679.

Bruns A (2018) Facebook shuts the gate after the horse has bolted, and hurts real research in the process. *Medium*. Available at: https://medium.com/@Snurb/facebook-research-data-18662cf2cacb (accessed 30 October 2018).

Burke M and Kraut RE (2016) The relationship between Facebook use and well-being depends on communication type and tie strength: Facebook and well-being. *Journal of Computer-Mediated Communication* 21(4): 265–281.

Couldry N and Yu J (2018) Deconstructing datafication's brave new world. *New Media & Society* 20(12): 4473–4491.

Dehnhard I, Weichselgartner E and Krampen G (2013) Researcher's willingness to submit data for data sharing: a case study on a data archive for psychology. *Data Science Journal* 12: 172–180.

Ellis SE and Leek JT (2018) How to share data for collaboration. *The American Statistician* 72: 53–57.

Fafalios P, Iosifidis V, Ntoutsi E, et al. (2018) TweetsKB: a public and large-scale RDF corpus of annotated tweets. In: Gangemi A, Navigli R, Vidal M-E, et al. (eds) *The Semantic Web*. Cham: Springer, 2018, pp. 177–190.

Fecher B, Friesike S and Hebing M (2015) What drives academic data sharing? (ed Phillips RS). *PLoS ONE* 10(2): e0118053.

Flaxman S, Goel S and Rao JM (2016) Filter bubbles, echo chambers, and online news consumption. *Public Opinion Quarterly* 80(S1): 298–320.

Freelon D (2018) Computational research in the post-API age. *Political Communication* 35(4): 665–668.

Guess AM (2015) Measure for measure: an experimental test of online political media exposure. *Political Analysis* 23(1): 59–75.

Haim M and Nienierza A (2019) Computational observation: challenges and opportunities of automated observation within algorithmically curated media environments using a browser plug-in. *Preprint SocArXiv*. DOI: 10.31235/osf.io/xd63n.

Halavais A (2019) Overcoming terms of service: a proposal for ethical distributed research. *Information, Communication & Society* 22(11): 1567–1581.

Hemphill L, Leonard SH and Hedstrom M (2018) Developing a social media archive at ICPSR. In: *Proceedings of web archiving and digital libraries (WADL'18)*, Fort Worth, TX, 18 June 2018, article 4. New York: ACM.

Houtkoop BL, Chambers C, Macleod M, et al. (2018) Data sharing in psychology: a survey on barriers and preconditions. *Advances in Methods and Practices in Psychological Science* 1(1): 70–85.

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)*, Oxford, 26–29 May, pp. 178–187. Palo Alto, CA: The AAAI Press.

ICC/ESOMAR (2016) ICC/ESOMAR international code on market, opinion and social research and data analytics. Available at: https://www.esomar.org/uploads/public/knowledge-and-standards/codes-and-guidelines/ICCESOMAR_Code_English_.pdf (accessed 30 October 2018).

Kinder-Kurlanda K, Weller K, Zenk-Möltgen W, et al. (2017) Archiving information from geotagged tweets to promote reproducibility and comparability in social media research. *Big Data & Society* 4(2). DOI: 10.1177/2053951717736336.

King G (1995) Replication, replication. *PS: Political Science and Politics* 28(3): 444.

King G and Persily N (2018) A new model for industry-academic partnerships. Working paper. Available at: https://gking.harvard.edu/partnerships (accessed 30 October 2018).

Klein O, Hardwicke TE, Aust F, et al. (2018) A practical guide for transparency in psychological science. *Collabra: Psychology* 4(1): 20.

LeBel EP, McCarthy RJ, Earp BD, et al. (2018) A unified framework to quantify the credibility of scientific findings. *Advances in Methods and Practices in Psychological Science* 1(3): 389–402.

Lomborg S (2013) Personal internet archives and ethics. *Research Ethics* 9(1): 20–31.

Lomborg S and Bechmann A (2014) Using APIs for data collection on social media. *The Information Society* 30(4): 256–265.

Lundberg I, Narayanan A, Levy K, et al. (2018) Privacy, ethics, and data access: a case study of the fragile families challenge. arXiv preprint. Available at: https://arxiv.org/abs/1809.00103 (accessed 30 October 2018).

McKee HA and Porter JE (2009) *The Ethics of Internet Research: A Rhetorical, Case-based Process*. New York: Peter Lang.

Mannheimer S and Hull EA (2017) Sharing selves: developing an ethical framework for curating social media data. *International Journal of Digital Curation* 12(9): 1–15.

Markham A (2012) Fabrication as ethical practice. *Information, Communication & Society* 15(3): 334–353.

Markham A and Buchanan E (2012) Ethical decision-making and internet research: recommendations from the AoIR ethics working committee (Version 2.0). Available at: https://aoir.org/reports/ethics2.pdf (accessed 30 October 2018).

Matz SC, Kosinski M, Nave G, et al. (2017) Psychological targeting as an effective approach to digital mass persuasion. *Proceedings of the National Academy of Sciences of the United States of America* 114(48): 12714–12719.

Metcalf J and Fiesler C (2018) How Facebook can stop the next Cambridge Analytica. *Slate*. Available at: https://slate.com/technology/2018/03/cambridge-analytica-demonstrates-that-facebook-needs-to-give-researchers-more-access.html (accessed 28 June 2019).

Pasek J, McClain CA, Newport F, et al. (2019) Who's tweeting about the president? What big survey data can tell us about digital traces? *Social Science Computer Review* 38(5): 633–650.

Perkmann M and Walsh K (2007) University–industry relationships and open innovation: towards a research agenda. *International Journal of Management Reviews* 9(4): 259–280.

Poyago-Theotoky J, Beath J and Siegel DS (2002) Universities and fundamental research: reflections on the growth of university–industry partnerships. *Oxford Review of Economic Policy* 18(1): 10–21.

Ruckenstein M and Pantzar M (2017) Beyond the quantified self: thematic exploration of a dataistic paradigm. *New Media & Society* 19(3): 401–418.

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

Scharkow M (2016) The accuracy of self—reported internet use —a validation study using client log data. *Communication Methods and Measures* 10(1): 13–27.

Sharon T and Zandbergen D (2017) From data fetishism to quantifying selves: self-tracking practices and the other values of data. *New Media & Society* 19(11): 1695–1709.

Thomson SD (2016) Preserving social media. DPC Tech Watch Report. Available at: www.dpconline.org/docs/technology-watch-reports/1486-twr16-01/file (accessed 30 October 2018).

Thomson SD and Kilbride W (2015) Preserving social media: the problem of access. *New Review of Information Networking* 20(1–2): 261–275.

UK Data Service (2019) The DPA and GDPR. Available at: https://ukdataservice.ac.uk/manage-data/legal-ethical/obligations/data-protection.aspx (accessed 28 June 2019).

Vaccari C, Valeriani A, Barberá P, et al. (2016) Of echo chambers and contrarian clubs: exposure to political disagreement among German and Italian users of Twitter. *Social Media + Society* 2(3): 1–24.

Van Atteveldt W and Peng TQ (2018) When communication meets computation: opportunities, challenges, and pitfalls in computational communication science. *Communication Methods and Measures* 12(2–3): 81–92.

Van den Eynden V, Corti L, Woollard M, et al. (2019) *Managing and Sharing Research Data: A Guide to Good Practice*. London: SAGE.

Van Dijck J (2014) Datafication, dataism and dataveillance: Big Data between scientific paradigm and ideology. *Surveillance & Society* 12(2): 197–208.

Vanpaemel W, Vermorgen M, Deriemaecker L, et al. (2015) Are we wasting a good crisis? The availability of psychological research data after the storm. *Collabra: Psychology* 1(1): Article 3.

Weller K and Kinder-Kurlanda K (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 ICWSM workshop 2015*. Available at: http://www.aaai.org/ocs/index.php/ICWSM/ICWSM15/paper/viewFile/10657/10552 (accessed 30 October 2018).

Weller K and Kinder-Kurlanda K (2017) To share or not to share? Ethical challenges in sharing social media-based research data. In: Zimmer M and Kinder-Kurlanda K (eds) *Internet Research Ethics for the Social Age: New Challenges, Cases, and Contexts*. New York: Peter Lang, pp. 115–129.

Weller K and Kinder-Kurlanda KE (2016) A manifesto for data sharing in social media research. In: *Proceedings of the 8th ACM conference on web science —WebSci'16*, Hannover, 22–25 May, pp. 166–172. New York: ACM Press.

Williams ML, Burnap P and Sloan L (2017) Towards an ethical framework for publishing Twitter data in social research: taking into account users' views, online context and algorithmic estimation. *Sociology* 51(6): 1149–1168.

Yarkoni T and Westfall J (2017) Choosing prediction over explanation in psychology: lessons from machine learning. *Perspectives on Psychological Science* 12(6): 1100–1122.

Zenk-Möltgen W, Akdeniz E, Katsanidou A, et al. (2018) Factors influencing the data sharing behavior of researchers in sociology and political science. *Journal of Documentation* 74(5): 1053–1073.

Zuboff S (2015) Big other: surveillance capitalism and the prospects of an information civilization. *Journal of Information Technology* 30(1): 75–89.

Author biographies

Johannes Breuer is a senior researcher at GESIS - Leibniz Institute for the Social Sciences in Cologne, Germany where he now mostly works on the topics of digital trace data and data linking. His other research interests include the use and effects of digital media, computational methods, data management, and open science. Among other things, he has recently co-edited a special issue on "Integrating Survey Data and Digital Trace Data" for the journal *Social Science Computer Review*.

Libby Bishop is coordinator for International Data Infrastructures at the GESIS - Leibniz Institute for the Social Sciences in Cologne, Germany. Recently, she has published on the ethical issues in publishing and sharing social media data and the value of moral theory in analyzing ethical challenges in reusing data. She is currently developing a working paper that argues the threats to privacy from big data, including social media, are better challenged using an ethical framework based on human dignity—exemplified in Continental law—rather than a framework based on liberty as defined in the Anglo-American legal tradition.

Katharina Kinder-Kurlanda is a senior researcher and team leader at the GESIS - Leibniz Institute for the Social Sciences in Cologne, Germany. Katharina's publications include work on the epistemology of big (social media) data; research ethics; social games and data protection & security. She is co-editor (with M. Zimmer) of the book *Internet Research Ethics for the Social Age: New Challenges, Cases, and Contexts*.