Integrating Survey Data and Digital Trace Data: Key Issues in Developing an Emerging Field

Stier, Sebastian; Breuer, Johannes; Siegers, Pascal; Thorson, Kjerstin

Veröffentlichungsversion / Published Version
Zeitschriftenartikel / journal article

Social Science Computer Review
2020, Vol. 38(5) 503-516

DOI: 10.1177/0894439319843669
journals.sagepub.com/home/ssc

Integrating Survey Data and Digital Trace Data: Key Issues in Developing an Emerging Field

Sebastian Stier[1], Johannes Breuer[1], Pascal Siegers[1], and Kjerstin Thorson[2]

Abstract

While survey research has been at the heart of social science for decades and social scientific research with digital trace data has been growing rapidly in the last few years, until now, there are relatively few studies that combine these two data types. This may be surprising given the potential of linking surveys and digital trace data, but at the same time, it is important to note that the collection and analysis of such linked data are challenging in several regards. The three key issues are: (1) data linking including informed consent for individual-level studies, (2) methodological and ethical issues impeding the scientific (re)analysis of linked survey and digital trace data sets, and (3) developing conceptual and theoretical frameworks tailored toward the multidimensionality of such data. This special issue addresses these challenges by presenting cutting-edge methodological work on how to best collect and analyze linked data as well as studies that have successfully combined survey data and digital trace data to find innovative answers to relevant social scientific questions.

Keywords

data linking, surveys, digital trace data, social media, sensors, informed consent, linkage bias, data sharing

This article is part of the SSCR special issue on "Integrating Survey Data and Digital Trace Data", guest edited by Sebastian Stier, Johannes Breuer, Pascal Siegers (GESIS—Leibniz Institute for the Social Sciences) & Kjerstin Thorson (Michigan State University).

[1] GESIS—Leibniz Institute for the Social Sciences, Köln, Germany
[2] Michigan State University, East Lansing, MI, USA

Corresponding Author:
Sebastian Stier, GESIS—Leibniz Institute for the Social Sciences, Unter Sachsenhausen 6-8, D-50667 Köln, Germany.
Email: sebastian.stier@gesis.org

Background

Traditionally, quantitative social scientists have mostly used surveys to study topics like social networks, media use, political and civic participation, health-related behaviors, and others. However, especially against the backdrop of the ever-expanding use of digital technologies, studying contemporary human behavior with survey methods has several downsides. The most important one is the limited reliability of self-reported behavioral measures (typically assessed retrospectively), which is amplified by the diffuse and fast-paced information environments humans encounter in their daily lives via their smartphones, tablets, computers, or other digital devices. Moreover, surveys suffer from declining response rates, especially among strata of the population most likely to rely on digital technologies.

By contrast, studies from the relatively young field of computational social science (CSS) collect digital traces of human behavior in a nonintrusive way, with high precision and granularity. These digital trace data can be roughly defined as "records of activity (trace data) undertaken through an online information system (thus, digital)" (Howison, Wiggins, & Crowston, 2011) and can be collected from a multitude of technical systems, such as websites, social media platforms, smartphone apps, or sensors. At the same time, CSS approaches also have limitations. Importantly, most studies relying exclusively on digital trace data lack relevant information on individuals' activities across several venues online and offline, their attributes (e.g., sociodemographic characteristics or personality traits) or key outcome variables (e.g., voting, social, or political attitudes). Hence, these data alone cannot answer questions about individual-level determinants of human behavior.

Given the drawbacks of each of these two types of data, combining data and methods from survey research and CSS is a promising way to account for their respective weaknesses. However, in order to be able to meaningfully combine these data, three challenges need to be addressed. First, while there is great pluralism in the different ways to link surveys and digital trace data, we still lack a conceptual framework guiding researchers through the benefits and pitfalls of different approaches to *data linking*. Second, there are no shared standards regarding *methodological and ethical issues* such as the recruitment of participants, informed consent, the transformation of digital trace data into meaningful measurements, and data sharing. As of now, pioneering researchers have been active in different fields with only limited exchange between disciplines. Third, despite informative individual findings, we still lack an empirically informed *meta-perspective* on what the added benefits of such integrated research designs are (notable exceptions are Resnick, Adar, & Lampe, 2015; Wells & Thorson, 2015).

Our aim with this special issue is to characterize these issues in more depth and provide suggestions how they can be addressed. By bringing together studies and scholars from different disciplines including communications, political science, sociology, and survey methodology, we want to foster the exchange about methodological approaches and provide the readers with an overview of how the collection and analysis of combined survey and digital trace data have been tackled in different fields.[1] The methodological contributions in this issue address key questions regarding data collection, recruitment, and informed consent. The contributions pursuing substantive research questions pay great attention to describing and discussing the methods they use and explicitly describe the added value of integrating surveys and digital trace data in their specific case. Finally, this editorial itself is meant to provide a meta-perspective on the specific benefits and the unique challenges of combining survey data with digital traces.

Two Paradigms: Survey Research and CSS

This special issue focuses on two methodological paradigms for studying how humans interact with digital technologies: survey research and CSS. The most obvious difference between these two

approaches lies in the types of data they use: survey responses (self-reports) and digital trace data. But there are also differences in terms of the predominant methods of analysis. For example, survey researchers frequently employ (multilevel) regression or structural equation modeling, whereas in CSS machine learning, text mining or social network analysis are more prevalent.

Both of these paradigms have specific limitations that might be overcome by combining them. For example, surveys of routine behavior are either highly abstract (e.g., "on average, how often do you check Facebook per day") or very demanding on respondents for more granular aspects of online behavior (e.g., "how often did you retweet posts by a politician or a party in the last month"). Given that respondents already have difficulties in assessing their offline behaviors (e.g., traditional media use, see Prior, 2009), the validity and reliability of self-reports further suffers in the contemporary high-choice digitalized media environments. As Araujo, Wonneberger, Neijens, and de Vreese (2017) note, "the increasingly fragmented and ubiquitous usage of [the] internet complicates the accuracy of self-reported measures" (p. 173). As a result, most established general population surveys refrain from using items on specific types of behavior in their standard questionnaires, especially regarding online behavior (with some exceptions, such as the social media studies from Pew Research Center).

Data collection methods from CSS are capable of gathering detailed and reliable objective data on human behavior without affecting user behavior itself ("nonintrusive measurement," see Lazer et al., 2009). Compared to surveys, these data can reveal more fine-grained behavioral patterns, such as the number and contents of social media posts by a person over time, the sharing and tagging of photos, or geographic movement patterns of smartphone users. Researchers can, for example, collect data through the Application Programming Interfaces (APIs) of platforms like Facebook or Twitter or web crawlers. In addition, researchers can opt for more proactive types of data collection by incentivizing users to install plug-ins on their desktop computers and smartphones or hand out sensors such as GPS/movement trackers or wearable badges to study participants. These data collection schemes produce data which can take the form of clickstream data providing detailed information on web browsing sessions, networks of face-to-face conversations, time stamps and contents of human communication or geo-located traces of movements.

Digital trace data alone, however, are of limited use for social scientific research as it usually provides incomplete, imprecise, or no information about relevant attributes and attitudes of the individuals whose data are collected. Moreover, the data are most often based on biased samples, making it hard to link online behavior to microlevel theories from the social sciences (Jungherr, 2018). Many CSS studies, therefore, remain largely descriptive as the nature of their data offers very limited opportunities for theory-driven (causal) analyses. Accordingly, authors with a background in survey research are quite critical in evaluating the representativeness, validity, and reliability of these data (Diaz, Gamon, Hofman, Kiciman, & Rothschild, 2016; Japec et al., 2015; Schober, Pasek, Guggenheim, Lampe, & Conrad, 2016).

Linking Surveys and Digital Trace Data: Ways of Linking and Benefits for Research

There are various ways to link different data types in survey research and the social sciences in general. We chose a specific focus on surveys and digital trace data for this special issue. This excludes the combination of survey data with, for example, paradata (e.g., Roßmann & Gummer, 2015), geo data (e.g., Schweers, Kinder-Kurlanda, Müller, & Siegers, 2016), administrative records (e.g., Schnell, 2014), eye tracking (e.g., Vraga, Bode, & Troller-Renfree, 2016), or traditional forms of media content such as television or newspaper coverage (for a review, see de Vreese et al., 2017).

When thinking about ways to categorize different linking types combining digital trace data and surveys, one first has to focus on the unit of analysis. While linking at the individual level always

Table 1. Linking types with examples from the literature.

Ex Ante Linking	Ex Post Linking
(A) Aggregate level • Analysis of audience overlaps (e.g., Mukerjee et al., 2018; Nelson & Webster, 2017) • Analysis of aggregate audience statistics (e.g., political ideology, Nelson & Webster, 2017)	(B) Aggregate level Linking survey responses to digital trace data . . . • Temporally: both are generated during the same time period (e.g., Mellon, 2014; O'Connor et al., 2010; Stier et al., 2018) • Topically: both focus on the same topic (e.g., Pasek et al., 2019) • Geographically: both can be located within same geographic area (e.g., Beauchamp, 2017)
	(C) Public actors Link publicly available digital trace data of public actors (e.g., politicians or organizations) to their survey responses (e.g., Karlsen & Enjolras, 2016; Quinlan et al., 2017)
(D) Individual level Ask individuals in surveys for informed consent to record in real time: • Website visits (e.g., Guess, 2015; Jürgens et al., 2019; Möller et al., 2019; Vraga & Tully, 2018) • Smartphone data (e.g., Boase & Ling, 2013; Jürgens et al., 2019; Kreuter et al., 2019) • Sensor data (e.g., Génois, Zens, Lechner, Rammstedt, & Strohmaier, 2019)	(E) Individual level Ask individuals in surveys for informed consent to collect their historical digital trace data . . . • From social media APIs (e.g., Al Baghal et al., 2019; Haenschen, 2019; Hofstra, Corten, van Tubergen, & Ellison, 2017; Hopp, Vargo, Dixon, & Thain, 2018; Vaccari et al., 2015; Wells & Thorson, 2015) • via data donation, for example, personal Google or Facebook histories (e.g., Thorson et al., 2018)

means that data from different sources on the same participant are combined, it is also possible to link data at the aggregate level. This distinction is similar to the typology by de Vreese and Neijens (2016) who identified "user-centric" and "site-centric" approaches. Yet, in the case of digital trace data, a crucial distinction is whether the data are linked ex post or ex ante. *Ex ante* means that the linking is part of the research design and that the data are created for linking purposes. Collecting such digital traces involves active participation of respondents, for example, by installing a tracking tool on a device like a desktop computer or smartphone. In contrast, ex post linking uses data that are available through the APIs of social media platforms or collected by web scraping technologies. Combining the relevant dimensions results in the 5-fold table of linking types presented in Table 1.[2]

The studies listed in Table 1, Panel A, build on ex ante linking designs to investigate *aggregate audience behavior* in the consumption of online contents. Analogous to television meters used for decades in communications and market research, commercial companies incentivize users to install tools tracking their website visits in real time. In these cases, the data are aggregated because the companies transform data from their panels of users into statistics at the level of website domains. Analyses of audience overlaps, despite ongoing methodological debates, reveal a clear concentration of audience attention on major brands, which refutes the popular assumption that online audiences self-segregate into echo chambers (Mukerjee, Majó-Vázquez, & González-Bailón, 2018; Nelson & Webster, 2017). For the U.S. study by Nelson and Webster (2017), panelists also reported their political ideology, which allowed the authors to show that news websites across the political spectrum generally have an ideologically diverse audience, even partisan websites like Breitbart.

Such results are impossible to obtain, let alone validate with self-reported survey data for a broad set of websites in the long tail of visits. However, as researchers cannot directly survey people who participate in these commercial tracking panels, this type of data linking limits the analysis to macrolevel characterizations of audience behavior.

Until recently, the most common way of linking surveys and digital trace data has been to collect data independently and then constructing measurements which allow for an *aggregated comparison of survey responses and digital trace data* (Table 1, Panel B). In this scenario, researchers compare items from survey data, such as the "most important problem" question, politicians' approval ratings, or consumer confidence to aggregated frequencies of related content on social media, most often Twitter but also other sources like search data from Google Trends (Mellon, 2014). At the aggregate level, surveys and digital trace data can be linked ex post according to the temporal dimension, for example, by comparing topic salience in the general public to online audiences for a period of interest like an election campaign (Mellon, 2014; Stier, Bleier, Lietz, & Strohmaier, 2018), geographically (Beauchamp, 2017), or topically (O'Connor, Balasubramanyan, Smith, & Routledge, 2010; Pasek, McClain, Newport, & Marken, 2019). For topical linking, the challenge lies in the construction of equivalent measures for both data types. The study by Pasek, McClain, Newport, and Marken (2019) in this special issue falls into this category as it compares the sentiment of tweets and approval polls for former U.S. President Barack Obama and finds that tweet sentiment and reported approval are only loosely related. While the long-term trends in these two measures are generally comparable (especially for certain demographic groups), they seem to measure different things and reflect different processes. By disaggregating similar research questions for a specific geographical unit of analysis, other studies have combined topical, temporal, and geographical linking approaches (e.g., Beauchamp, 2017).

In these studies, the respective other data source serves as a comparative baseline for answering research questions, such as whether certain transformations of digital trace data can complement or even substitute opinion polls, whether topic salience within online systems overlaps with that of the general public, or whether digital trace data allow for the prediction of more general human behavior.

Since governments, politicians, organizations, or other nonprivate entities are *public actors* and large parts of their communication are public, their digital trace data in the form of posts on social media platforms are more easily available for linking with surveys (Table 1, Panel C).[3] For example, Quinlan, Gummer, Roßmann, and Wolf (2018) show that the size of the campaign budget reported in a candidate survey predicts social media adoption by German candidates running for federal office and that the personality trait openness is positively related to Twitter use in this population. In a similar study, Karlsen and Enjolras (2016) used a candidate survey to distinguish between party-centered and individualized campaign goals that were related to differences in how candidates used and to which extent they were successful on Twitter. These examples illustrate how the combination of surveys and digital trace data has expanded the field of election research, even though there is a lot of potential to go beyond readily available social media metrics, for example, by analyzing the content of posts. Moreover, these research designs can be extended to domains other than politics. For example, surveys of nongovernmental organizations and other organizations already exist and could be linked with their public communication on digital platforms. Accordingly, integrating survey data and digital trace data can also improve studies at the mesolevel of politics and societies.

Another recent line of research is devoted to collecting data at the *individual level* (Table 1, Panels D and E). This means collecting digital trace data from a set of participants who also respond to one or more surveys. In the case of an ex ante linking (Panel D), data from survey respondents are linked to digital traces, for example, through smartphones, a streaming of tweets via the Twitter Streaming API, or a browser plug-in in real time—meaning that the collection of the two types of data happens in parallel. On the other hand (Panel E), researchers can also collect

historical digital trace data via a Facebook application collecting posts and friends lists or an upload of participants' histories from web browsers (Menchen-Trevino, 2016; Thorson & Wells, 2016). These historical digital traces can then be linked to surveys ex post. The introduction of the European General Data Protection Regulation (EUGDPR) in May 2018 opened up another opportunity for scientific data collection. The EUGDPR requires digital platforms to provide users with an option to access and export all of their personal data. Hence, study participants can download and donate these data for scientific purposes.

Importantly, linking digital trace and survey data at the individual level requires the informed consent of participants (see discussion on challenges below). Within the scope of this review, linking data at the individual level (ex ante and ex post) currently is the fastest expanding area as it allows researchers to study individual-level (online) behavior in an ecologically valid way while giving them control over all steps of the research process.

In a first line of research we identified, scholars from communications and political science have focused on the question whether self-reported media use corresponds to passively recorded behavior of study participants. Overall, studies reveal a low accuracy of self-reported frequency and duration of Internet use with observed behavior (Jürgens et al., 2019). Participants tend to overreport Internet use in general (Araujo et al., 2017; Scharkow, 2016) as well as smartphone use (Boase & Ling, 2013). Further, survey respondents cannot accurately recall visits to specific websites visits as well as their frequency (Revilla, Ochoa, & Loewe, 2017) and are particularly prone to overreporting visits to online news websites (Guess, 2015) and political articles (Vraga & Tully, 2018). A few studies have also linked surveys and data from Facebook and Twitter at the individual level. Guess, Munger, Nagler, & Tucker (2018) and Haenschen (2019, in this issue) reveal that self-reports on social media activity are (surprisingly) accurate at the aggregate level. At the same time, these correlations mask substantive individual-level biases correlated with demographics but with a lot of variation across different behaviors (such as sharing and following political accounts.).

The most robust results across validation studies and the most problematic finding for the field of political communication are that political interest is both correlated with misreporting (Guess et al., 2018; Haenschen, 2019; Vraga & Tully, 2018) and, as is well-known, with factors like voting or political knowledge. It is also noteworthy that the relationship between misreporting and political variables all come from the United States, an increasingly polarized political and media system. Further studies should investigate whether political considerations affect survey responses on (online) behavior in other contexts as well. These studies have also shown ways to improve survey items by testing various types of self-report items against each other (Guess, 2015; Guess et al., 2018; Haenschen, 2019).

Other studies linking surveys with digital traces have not focused on the validity of self-reports but rather on using these data for substantive analysis. One of the most influential research teams in this area collected information on more than 4 million users through their *myPersonality* Facebook application that gave participants feedback on their personality traits (Kosinski, Stillwell, & Graepel, 2013). The researchers gathered data on the Big Five personality traits, intelligence, satisfaction with life, substance use, and other attributes. The Facebook app also collected extensive metadata on the Facebook profiles, including age, gender, or number of friends, and behavioral variables, such as page likes. The various published studies demonstrate that Facebook behavior predicts personality traits and personal attributes (measured via surveys).

Wells and Thorson (2015) used a Facebook application to collect survey and digital trace data from a sample of young adults, following a process of informed consent. Their app collected all the posts that appeared in participants' newsfeeds in order to capture digital traces of news exposure. The authors found low levels of Facebook news exposure in their sample but demonstrated the importance of political interest and the habit of social media customization in

explaining levels of exposure. Substantively, Facebook news exposure was related to participation but not to political knowledge.

The Copenhagen Networks Study integrated even more data sources at a larger scale by collecting data on face-to-face interactions, mobile phone communication, and Facebook use from 1,000 university students. Their main source of information was data from mobile phones, but they also used surveys, tracking via campus Wi-Fi, and participant observation. They found that extraversion is positively correlated with the intensity of Facebook use and the size of communication networks (Stopczynski et al., 2014).

Hofstra, Corten, van Tubergen, and Ellison (2017) used data from Facebook profiles of fifth-grade students to model their social networks. Applying onomastic methods to detect the ethnic origin of names of the contacts, the authors were able to determine the level of segregation in students' social networks. This way of generating network data is much more precise than self-reports and saves costs as administrating modules on ego-centered networks is time consuming and usually limited to the three to five most important contacts. Other approaches (Kristensen et al., 2017) focus on predicting (in a statistical sense) voting intentions measured in surveys by data on Facebook likes. The results suggest that likes accurately express individuals' political views and only a few digital traces are required for an accurate measurement of political opinions. The study by Burke and Kraut (2016) is a special case where the authors had access to internal Facebook data. Linking data from a survey and server logs for $N = 1{,}910$ Facebook users they found that the impact of Facebook use on well-being depends on the type of communication and contact. They also revealed that only receiving direct messages from close contacts positively affected well-being among the participants.

In their contribution to this special issue, Möller, van de Velde, Merten, and Puschmann (2019) offer a different approach to tracking news exposure online. They recruited a panel to install a browser plug-in that collected data about the users' website visits to a series of white-listed media domains. Their method allowed them to track not only visits to news media, but also the journey users took to reach those sites. They distinguish between several modes of online news use: routine use, as when users visit the home page of a news site news use triggered by search, and social media news use. They found little routine news use online and substantially higher amounts of news use driven by search.

The combination of surveys and digital trace data is also a promising avenue for experimental research. In the intervention study by Munson, Lee, and Resnick (2013), a browser extension was created, which provided participants with information about the imbalance of their political news consumption (liberal vs. conservative) and found that receiving this feedback led to a slightly more balanced news consumption. The contribution by Vraga and Tully (2018) in this special issue simulated a news aggregator site that experimentally manipulated contextual and story cues. Participants were able to choose from news stories and were unobtrusively tracked and surveyed before and after the experiment. The results show that political beliefs and characteristics of news environments affect the accuracy of self-reported media use.

The above examples illustrate that it is fruitful to combine survey experiments with digital trace data. When embedded in such field experiments, observing participants' behavior allows for a better measurement of prior behavior (how often did a participant engage in a behavior the researcher wants to change?), compliance with treatments, and the intensity of treatment effects (how often does a participant engage in the behavior triggered by the treatment?). If changes in attitudes rather than changes in behavior itself are the dependent variable, the researchers can additionally survey participants in a posttreatment survey. While only a few studies have combined surveys, digital trace data, and experiments so far, this approach holds great potential for research on topics like online information search or news consumption.

In summary, the studies discussed in this section demonstrate that the various linking types presented in Table 1 can be used to (1) improve substantive analysis, (2) cross-validate and improve

measurements, and (3) design survey experiments to generate meaningful digital trace data and directly measure treatment effects.

Challenges in Combining Surveys and Digital Trace Data

Even though scholars increasingly use digital traces in social research, a critical and comprehensive reflection about the limitations of such data is still needed. The next step for research with linked data is to create data sets suitable for a (broader) generalization of the results. In the following, we will outline a set of key challenges that emerged from existing research and from the contributions in the special issue.

Recruiting Respondents

From an ethical perspective, it is important to keep in mind that linking digital trace data and survey data at the individual level requires explicit and informed consent from the people whose data are collected (Menchen-Trevino, 2018). From a data privacy perspective, linking survey and digital trace data can also be problematic as this usually requires that account names and sometimes also the real names are known to the researchers. In addition, the possibilities for collecting, storing, and sharing digital trace data depend on the terms of services (ToS) of the services or platforms these data come from and, potentially, also the options and limitations of their respective APIs. From a more technical perspective, linking surveys and digital trace data can be challenging because not even full names are unique identifiers and many people do not use their real names online.

The introduction of the EUGDPR modified the rules for obtaining consent by specifying that scholars must inform participants about the scope of data storage, the research purpose, and the time data will be stored for. Moreover, participants have the right to withdraw consent. Researchers also need explicit (ideally written) consent to get an identifier for linking with information from social media platforms (e.g., Twitter handles) or for installing tracking apps on participants' devices (e.g., for web tracking or passive location data).

Obtaining consent, therefore, is a crucial step for research designs linking data at the individual level. Nevertheless, when reviewing existing literature, we noticed that many studies do not report information about how informed consent was obtained from study participants. Two papers in this special issue contribute to the study of factors influencing informed consent to the collection of digital trace data. Both provide extensive information about how respondents were recruited into the collection of Twitter data (Al Baghal, Sloan, Jessop, Williams, & Burnap, 2019) and extensive sensor and usage data collected with an app for Android mobile devices (Kreuter et al., 2019). They use data from high-quality large-scale survey programs from the United Kingdom and Germany that were sampled using probabilistic methods, which is still the exception in research with digital trace data. The results are in line with previous research, showing that consent rates are comparatively low, even if participants are recruited from an existing panel study. Especially when recruiting survey respondents for the tracking of social media data, the low consent rates are problematic. For example, Al Baghal, Sloan, Jessop, Williams, and Burnap (2019) report a Twitter penetration between 20% and 25% for Great Britain. With consent rates of 40%, only 10% of the net sample is available for data linking.

The most important advantage of Twitter for social research is that the data are public and easy to collect using various tools freely available to researchers. Future research, however, should focus more on platforms with higher penetration rates. Given recent restrictions of social media APIs, innovative ways for data collection must be explored, including the model of data donation by the users (Thorson, Medeiros, Cotter, & Pak, 2018), without violating the ToS of platforms.

Minimizing Potential Bias

Comparatively low consent rates raise the issue of potential selectivity and bias because the risk for bias accumulates with different types of nonresponse. The first concern is unit nonresponse in the survey, the second is nonuse of the online platform or service under study, the third is nonconsent to the tracking, and the fourth is nonresponse to the tracking. In their contribution to this issue, Jürgens et al. (2019) provide a detailed analysis of sampling biases that can emerge in digital tracking studies. They draw on a 14-day tracking panel that captured participants' data from computers, mobile phones, and tablets. The authors leverage these data to empirically demonstrate the relationships between sampling biases and bias in self-reports of time spent online. For example, participants who were unwilling to share their mobile data were also more likely to overreport media use. These findings remind readers that, given the sampling challenges we outlined above, tracking data themselves can also be biased measures of media exposure.

Kreuter, Haas, Keusch, Bähr, and Trappmann (2019) also identified technological issues for studies that want to link surveys and digital trace data. For example, developing apps covering all relevant operating systems is costly and the distribution must proceed through the app stores of the service providers. A key problem is that the operating systems for mobile devices do not grant access to the same data for all apps. Apple's iOS, for example, does not allow for the collection of location data. Therefore, there is an additional risk of technology-induced bias in tracking data, although the authors show that Android and iOS users differ only slightly regarding demographic characteristics.

Notably, recruiting participants from social media platforms will only allow for conclusions about a subset of the users on a given platform, which cannot be generalized to the population on social media (e.g., participants tend to be the most active and intrinsically motived users). Moreover, the techniques to target users are subject to changes by platforms. Twitter, for instance, has made sending automated messages via social bots (the approach used by Vaccari et al., 2015) much more difficult. Hence, researchers trying to invite thousands of Twitter users to a survey would have to create multiple accounts and carefully navigate API rate limits and Twitter's ToS. Tailored surveys can be used, however, to reach specific target populations that tend to be less responsive to traditional survey sampling strategies (Iannelli, Giglietto, Rossi, & Zurovac, 2018).

Survey companies like YouGov have created online panels, which include behavioral tracking. But these commercial panels are also built on the willingness of participants to opt in and are neither "truly" representative of the general nor even the online population. These data might also have additional biases a researcher cannot control for (Jürgens, Stark, & Magin, 2019). Taken together, recent methodological research reveals substantive biases in all potential sampling approaches that allow for a linking of surveys and digital trace data.

Accessing and Sharing Digital Trace Data

Due to changes to the Facebook API (Freelon, 2018), the approaches employed by Hofstra et al. (2017), Wells and Thorson (2015), and others cannot be used anymore. To collect data from Facebook, researchers have to develop new methods that respect the privacy of users and the ToS of platforms. As several services have limited access and APIs can be closed quickly for a variety of reasons, Freelon (2018) speaks of a "post-API age" for computational research. Despite all of the obvious advantages of digital trace data collections, the need for cooperation from platform companies to access data is an important caveat that researchers also need to keep in mind.

In addition to data access, another major concern is how the data can be made available for other researchers for replication or additional analyses while respecting the privacy of the people whose data are collected, as well as the ToS of the platforms on which the data were collected. While some publications provide guidance for the ethical sharing of social media data (e.g., Bishop, 2017;

Mannheimer & Hull, 2017; Williams et al., 2017), finding a good balance between privacy protection, ToS compliance on the one hand, and reproducibility and reusability on the other hand remains a challenging task and requires case-by-case decisions and consultation.[4]

Measurements Constructed from Digital Trace Data

All too often, CSS studies have focused on easily measurable online metrics without convincingly linking these to established social science theories (Jungherr, 2018). Integrating surveys and digital trace data opens up novel opportunities to explain human behavior with individual-level variables. However, the accelerated engagement of individuals with digital technology also requires the development and testing of new theoretical models for the study of human behavior. One such attempt is the "curated flows" model by Thorson and Wells (2016). It is based on the observation that the individual usage patterns for digital media tend to be unique and not only depend on personal preferences and habits but also on social and technological factors. Research questions about the influence of these different forces cannot be answered using isolated data from one platform or an ego-centered survey of media use. Correspondingly, Resnick, Adar, and Lampe (2015) argue that, from an epistemological standpoint, user behavior should be studied across time and across different platforms to allow for robust causal testing of social scientific theory using digital trace data. Especially longitudinal studies using digital traces have to find a way to control for changes in technologies to disentangle technology-induced change from changing behavioral patterns.

A key challenge that remains is to accurately measure a concept in both surveys and digital trace data. Varying the operationalization of a given online metric can impact its correlation with an offline indicator. For instance, Twitter mentions closely mirrored election results for one German election, but only when references to the Pirate Party—the dominant party in online discussions—were excluded (Jungherr, Jürgens, & Schoen, 2012). Future studies will have to develop sound theoretical models that link online behaviors to offline outcomes and include robustness tests to demonstrate that a given finding holds with different operationalizations of constructs based on digital trace data.

In terms of measurement, few studies linking digital traces to survey data have gone beyond simple metadata counts of digital traces. When researchers want to analyze the substance of human communication and behavior, they very quickly face constraints due to the sheer size and unstructured nature of the data, which makes hand-coding unfeasible in most cases (but see Haenschen, 2019). The paper by Hopp, Vargo, Dixon, and Thain (2018) in this issue stands out by applying the Google's Perspective API to classify incivility at a larger scale. The authors demonstrate that there is a considerable congruence between self-reports on online incivility and computational measures. They also outline ways how to improve measurements of complex social science concepts with a theory-driven integration of surveys and digital trace data. But in general, the methodological portfolio of CSS methods has so far not been used to its full potential, at least not in studies that combine these two data types.

Conclusion

This review has shown that the social sciences have already profited from innovative studies integrating survey data and digital trace data. Among the main advantages are the cross-validation and improvement of measurements, the explanation of human behavior at a large scale, and novel opportunities to improve causal inference in experimental settings. Nevertheless, there remain crucial methodological questions (e.g., regarding representativeness or informed consent) that have not been sufficiently addressed or even critically reflected yet. Researchers who want to exploit the advantages of linked data sets have to design their studies on thorough

theoretical grounds and be aware of the potential biases inherent to both underlying data generation processes. The contributions to the special issue show a way forward, so that this emerging field can further mature.

Authors' Note

We thank the Center for Advanced Internet Studies (CAIS) in Bochum for a generous conference grant that allowed us to bring together scholars working with surveys and digital trace data. We also thank Margarita Gutjar for her assistance with the literature review.

Declaration of Conflicting Interests

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

Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

Notes

1. It is important to note that our scope for this special issue excludes most of the recent advances in fields like medical research, human computing, mental health studies, and psychology, where surveys have also been combined with digital trace data in a number of studies. We also restrict our review of the literature to peer-reviewed published studies, even though we are aware of many relevant working papers.
2. Of course, it would be possible to add further dimensions to this typology, such as how the data are obtained (method and source) or what specific type of digital trace data is used, but we believe that the dimensions we present here are sufficient to characterize the main ways of linking surveys and digital trace data.
3. Another advantage of using this data is that it is less sensitive from an ethical perspective as it is public and comes from figures or institutions of public interest.
4. For this special issue, all data sets are available from the authors. We worked closely with authors to find data sharing solutions that strike this balance.

References

Al Baghal, T., Sloan, L., Jessop, C., Williams, M. L., & Burnap, P. (2019). Linking Twitter and survey data: The impact of survey mode and demographics on consent rates across three UK studies. *Social Science Computer Review*. doi:10.1177/0894439319828011

Araujo, T., Wonneberger, A., Neijens, P., & de Vreese, C. (2017). How much time do you spend online? Understanding and improving the accuracy of self-reported measures of Internetuse. *Communication Methods and Measures*, *11*, 173–190. doi: 10.1080/19312458.2017.1317337

Beauchamp, N. (2017). Predicting and interpolating state-level polls using Twitter textual data. *American Journal of Political Science*, *61*, 490–503. doi:10.1111/ajps.12274

Bishop, E. L. (2017). Big data and data sharing: Ethical issues. *UK Data Service, UK Data Archive*. Retrieved from https://www.ukdataservice.ac.uk/media/604711/big-data-and-data-sharing_ethical-issues.pdf

Boase, J., & Ling, R. (2013). Measuring mobile phone use: Self-report versus log data. *Journal of Computer-Mediated Communication*, *18*, 508–519. doi: 10.1111/jcc4.12021

Burke, M., & Kraut, R. E. (2016). The relationship between Facebook use and well-being depends on communication type and tie strength. *Journal of Computer-Mediated Communication*, *21*, 265–281. doi: 10.1111/jcc4.12162

de Vreese, C. H., Boukes, M., Schuck, A., Vliegenthart, R., Bos, L., & Lelkes, Y. (2017). Linking survey and media content data: Opportunities, considerations, and pitfalls. *Communication Methods and Measures*, *11*, 221–244. doi:10.1080/19312458.2017.1380175

de Vreese, C. H., & Neijens, P. (2016). Measuring media exposure in a changing communications environment. *Communication Methods and Measures*, *10*, 69–80. doi:10.1080/19312458.2016.1150441

Diaz, F., Gamon, M., Hofman, J. M., Kiciman, E., & Rothschild, D. (2016). Online and social media data as an imperfect continuous panel survey. *PLOS ONE*, *11*. doi:10.1371/journal.pone.0145406

Freelon, D. (2018). Computational research in the post-API age. *Political Communication*, *35*, 665–668. doi:10.1080/10584609.2018.1477506

Génois, M., Zens, M., Lechner, C., Rammstedt, B., & Strohmaier, M. (2019). Building connections: How scientists meet each other during a conference. arXiv preprint: 1901.01182

Guess, A. M. (2015). Measure for measure: An experimental test of online political media exposure. *Political Analysis*, *23*, 59–75. doi:10.1093/pan/mpu010

Guess, A. M., Munger, K., Nagler, J., & Tucker, J. (2018). How Accurate Are Survey Responses on Social Media and Politics? *Political Communication*, 1–18. doi:10.1080/10584609.2018.1504840

Haenschen, K. (2019). Self-reported versus digitally recorded: Measuring political activity on Facebook. *Social Science Computer Review*. doi:10.1177/0894439318813586

Hofstra, B., Corten, R., van Tubergen, F., & Ellison, N. B. (2017). Sources of segregation in social networks: A novel approach using Facebook. *American Sociological Review*, *82*, 625–656. doi:10.1177/0003122417705656

Hopp, T., Vargo, C. J., Dixon, L., & Thain, N. (2018). Correlating self-report and trace data measures of incivility: A proof of concept. *Social Science Computer Review*. doi:10.1177/0894439318814241

Howison, J., Wiggins, A., & Crowston, K. G. (2011). Validity issues in the use of social network analysis with digital trace data. *Journal of the Association of Information Systems*, *12*, 767–797.

Iannelli, L., Giglietto, F., Rossi, L., & Zurovac, E. (2018). Facebook digital traces for survey research: Assessing the efficiency and effectiveness of a Facebook Ad–based procedure for recruiting online survey respondents in niche and difficult-to-reach populations. *Social Science Computer Review*. doi:10.1177/0894439318816638

Japec, L., Kreuter, F., Berg, M., Biemer, P., Decker, P., Lampe, C., . . . Usher, A. (2015). Big data in survey research. *Public Opinion Quarterly*, *79*, 839–880. doi:10.1093/poq/nfv039

Jungherr, A. (2018). Normalizing digital trace data. In Natalie Jomini Stroud & Shannon McGregor (Eds.), *Digital discussions* (pp. 9–35). New York, NY: Routledge.

Jungherr, A., Jürgens, P., & Schoen, H. (2012). Why the pirate party won the German election of 2009 or the trouble with predictions. *Social Science Computer Review*, *30*, 229–234. doi:10.1177/0894439311404119

Jürgens, P., Stark, B., & Magin, M. (2019). Two half-truths make a whole? On bias in self-reports and tracking data. *Social Science Computer Review*. doi:10.1177/0894439319831643

Karlsen, R., & Enjolras, B. (2016). Styles of social media campaigning and influence in a hybrid political communication system: Linking candidate survey data with Twitter data. *The International Journal of Press/Politics*, *21*, 338–357. doi:10.1177/1940161216645335

Kosinski, M., Stillwell, D., & Graepel, T. (2013). Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences*, *110*. doi:10.1073/pnas.1218772110

Kreuter, F., Haas, G. -C., Keusch, F., Bähr, S., & Trappmann, M. (2019). Collecting survey and smartphone sensor data with an app: Opportunities and challenges around privacy and informed consent. *Social Science Computer Review*. doi:10.1177/0894439318816389

Kristensen, J. B., Albrechtsen, T., Dahl-Nielsen, E., Jensen, M., Skovrind, M., & Bornakke, T. (2017). Parsimonious data: How a single Facebook like predicts voting behavior in multiparty systems. *PLOS ONE*, *12*. doi:10.1371/journal.pone.0184562

Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabasi, A.-L., Brewer, D., . . . Van Alstyne, M. (2009). Computational social science. *Science*, *323*, 721–723. doi:10.1126/science.1167742

Mannheimer, S., & Hull, E. (2017). *Sharing selves: Developing an ethical framework for curating social media data*. Paper presented at the twelfth International Digital Curation Conference (IDCC), Edinburgh, Scotland.

Retrieved from https://scholarworks.montana.edu/xmlui/bitstream/handle/1/12661/Mannheimer-Hull-Sharing-Selves-2017.pdf

Mellon, J. (2014). Internet search data and issue salience: The properties of Google Trends as a measure of issue salience. *Journal of Elections, Public Opinion and Parties*, *24*, 45–72. doi:10.1080/17457289.2013.846346

Menchen-Trevino, E. (2016). Web Historian: Enabling multi-method and independent research with real-world web browsing history data. In *iConference 2016 Proceedings*. Grandville, MI: iSchools.

Menchen-Trevino, E. (2018). Digital trace data and social research: A proactive research ethics. In B. Foucault Welles & S. González-Bailón (Eds.), *The Oxford handbook of networked communication*. Oxford: Oxford University Press.

Möller, J., van de Velde, R. N., Merten, L., & Puschmann, C. (2019). Explaining online news engagement based on browsing behavior: Creatures of habit? *Social Science Computer Review*. doi:10.1177/0894439319828012

Mukerjee, S., Majó-Vázquez, S., & González-Bailón, S. (2018). Networks of audience overlap in the consumption of digital news. *Journal of Communication*, *68*, 26–50. doi:10.1093/joc/jqx007

Munson, S. A., Lee, S. Y., & Resnick, P. (2013). Encouraging reading of diverse political viewpoints with a browser widget. In *Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media* (pp. 419–428). Palo Alto, CA: AAAI Press.

Nelson, J. L., & Webster, J. G. (2017). The myth of partisan selective exposure: A portrait of the online political news audience. *Social Media + Society*, *3*. doi:10.1177/2056305117729314

O'Connor, B., Balasubramanyan, R., Smith, N. A., & Routledge, B. R. (2010). From tweets to polls: Linking text sentiment to public opinion time series. In *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media* (pp. 122–129). Palo Alto, CA: AAAI Press.

Pasek, J., McClain, C. A., Newport, F., & Marken, S. (2019). Who's tweeting about the president? What big survey data can tell us about digital traces. *Social Science Computer Review*. doi:10.1177/0894439318822007

Prior, M. (2009). The immensely inflated news audience: Assessing bias in self-reported news exposure. *Public Opinion Quarterly*, *73*, 130–143. doi:10.1093/poq/nfp002

Quinlan, S., Gummer, T., Roßmann, J., & Wolf, C. (2018). "Show me the money and the party!"—Variation in Facebook and Twitter adoption by politicians. *Information, Communication & Society*, *21*, 1031–1049. doi:10.1080/1369118X.2017.1301521

Resnick, P., Adar, E., & Lampe, C. (2015). What social media data we are missing and how to get it. *The ANNALS of the American Academy of Political and Social Science*, *659*, 192–206. doi:10.1177/0002716215570006

Revilla, M., Ochoa, C., & Loewe, G. (2017). Using passive data from a meter to complement survey data in order to study online behavior. *Social Science Computer Review*, *35*, 521–536. doi:10.1177/0894439316638457

Roßmann, J., & Gummer, T. (2015). Using paradata to predict and correct for panel attrition. *Social Science Computer Review*, *34*, 312–332. doi:10.1177/0894439315587258

Scharkow, M. (2016). The accuracy of self-reported Internet use—A validation study using client log data. *Communication Methods and Measures*, *10*, 13–27. doi:10.1080/19312458.2015.1118446

Schnell, R. (2014). An efficient privacy-preserving record linkage technique for administrative data and censuses. *Statistical Journal of the IAOS*, *30*, 263–270. doi:10.3233/SJI-140833

Schober, M. F., Pasek, J., Guggenheim, L., Lampe, C., & Conrad, F. G. (2016). Social media analyses for social measurement. *Public Opinion Quarterly*, *80*, 180–211. doi:10.1093/poq/nfv048

Schweers, S., Kinder-Kurlanda, K., Müller, S., & Siegers, P. (2016). Conceptualizing a spatial data infrastructure for the social sciences: An example from Germany. *Journal of Map & Geography Libraries*, *12*, 100–126. doi:15420353.2015.1100152

Stier, S., Bleier, A., Lietz, H., & Strohmaier, M. (2018). Election campaigning on social media: Politicians, audiences, and the mediation of political communication on Facebook and Twitter. *Political Communication*, *35*, 50–74. doi:10.1080/10584609.2017.1334728

Stopczynski, A., Sekara, V., Sapiezynski, P., Cuttone, A., Madsen, M. M., Larsen, J. E., & Lehmann, S. (2014). Measuring large-scale social networks with high resolution. *PLOS ONE*, *9*. doi:10.1371/journal.pone.0095978

Thorson, K., Medeiros, M., Cotter, K., & Pak, C. (2018). *Advertising categories as clues about political content exposure on Facebook*. Paper presented to the American Political Science Association Conference, Boston, MA.

Thorson, K., & Wells, C. (2016). Curated flows: A framework for mapping media exposure in the digital age. *Communication Theory*, *26*, 309–328. doi:10.1111/comt.12087

Vaccari, C., Valeriani, A., Barberá, P., Bonneau, R., Jost, J. T., Nagler, J., & Tucker, J. A. (2015). Political expression and action on social media: Exploring the relationship between lower- and higher-threshold political activities among Twitter users in Italy. *Journal of Computer-Mediated Communication*, *20*, 221–239. doi:10.1111/jcc4.12108

Vraga, E., Bode, L., & Troller-Renfree, S. (2016). Beyond self-reports: Using eye tracking to measure topic and style differences in attention to social media content. *Communication Methods and Measures*, *10*, 149–164. doi:10.1080/19312458.2016.1150443

Vraga, E. K., & Tully, M. (2018). Who is exposed to news? It depends on how you measure: Examining self-reported versus behavioral news exposure measures. *Social Science Computer Review*. doi:10.1177/0894439318812050

Wells, C., & Thorson, K. (2015). Combining big data and survey techniques to model effects of political content flows in Facebook. *Social Science Computer Review*, *35*, 33–52. doi:10.1177/0894439315609528

Williams, M. L., Burnap, P., & 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*, 1149–1168. doi:10.1177/0038038517708140

Author Biographies

Sebastian Stier is a senior researcher in the Department Computational Social Science at GESIS—Leibniz Institute for the Social Sciences in Cologne, Germany. His research focuses on political communication, party politics, populism, and the use of digital trace data and computational methods in the social sciences. Email: sebastian.stier@gesis.org

Johannes Breuer is a senior researcher in the Department Data Archive at GESIS in Cologne. His main research interests are the use and effects of digital media, the methods of media (effects) research, data linking, data management, and open science. Email: johannes.breuer@gesis.org

Pascal Siegers is the head of the research data center "German General Social Survey" in the Department Data Archive at GESIS in Cologne. His primary research interests are religious influences on moral and political attitudes in a longitudinal perspective and the discovery of new data types for social research including spatial data and digital trace data. Email: pascal.siegers@gesis.org

Kjerstin Thorson is an associate professor in the College of Communication Arts & Sciences at Michigan State University. Her research explores how people use digital and social media to learn about and participate in politics, and how social media platforms are reshaping the visibility of news and politics. Email: thorsonk@msu.edu