Data Is the New Oil—But How Do We Drill It? Pathways to Access and Acquire Large Data Sets in Communication Science

DANIEL POSSLER
SOPHIE BRUNS
JULIA NIEMANN-LENZ
Hanover University of Music, Drama, and Media, Germany

In the Internet era, many forms of human communication leave massive digital traces that could serve as a revolutionary source of empirical information. However, researchers are often faced with extensive challenges when they try to access these "big communication data." These restrictions prevent scientific exploration and the application of new computational methods and, thus, ultimately constrain the development of the field of computational communication science (CCS). As a first step to improve scientific access to large communication data, this article systematizes the data acquisition options and outlines their specific advantages and barriers based on a qualitative interview study with 22 researchers in the field of CCS. Three overarching themes are identified that seem to determine the resulting data quality of all data collection methods. On the basis of these findings, we develop an agenda on how communication science could overcome the challenges of accessing large data sets in the future.

Keywords: computational communication science, data collection, big data, API, web scraping

The rapid and broad diffusion of Internet-based infrastructures, services, and platforms has fundamentally changed the way interpersonal, intragroup, and public communication is carried out. Today, not only are large parts of human communication mediated via the Internet, but also completely new forms of communication have emerged—social media represent a key example of this development (Castells, 2002). These changes in the communication landscape also provide new opportunities for communication science. Above all, the body of relevant data has grown rapidly as many forms of Internet-based communication leave massive digital traces. These new big communication data could serve as a revolutionary source of empirical information for the scientific study of communication (Lazer et al., 2009; Liang & Zhu, 2017; Lomborg & Bechmann, 2014). To name just a few examples: User comments can provide nonreactive insights into the development of public opinions (e.g., González-Bailón, Banchs, &

Daniel Possler: Daniel.Possler@ijk.hmtm-hannover.de
Sophie Bruns: Sophie.Bruns@ijk.hmtm-hannover.de
Julia Niemann-Lenz: Julia.Niemann-Lenz@ijk.hmtm-hannover.de
Date submitted: 2018–10–19

Kaltenbrunner, 2012), and logs of smartphones or Internet browsers can precisely measure individual media usage behaviors (e.g., Bodo et al., 2017).

However, although an enormous amount of "big communication data" is generated on online platforms, researchers are often faced with the challenge of accessing the data they need to answer substantial research questions (boyd & Crawford, 2012; Bruns, 2018a; Lazer et al., 2009). These restrictions may hamper scientific exploration as scholars are forced to focus on those questions for which data is available, even though these might not be the most pressing topics scientifically ("availability bias"; Mahrt & Scharkow, 2013). For example, although Facebook is by far the most popular social media platform worldwide (Hootsuite & We Are Social, 2018), research on social media users' activities mostly draws upon Twitter data as the platform has (initially) been quite generous in granting access to these data resources (Bruns, 2018a; Lomborg & Bechmann, 2014). In addition, limited access prevents researchers from realizing the potentials of analyzing large data sets such as the possibility of making inferences that go beyond individual cases, a single media content, or small groups of users (e.g., using global social media data to conduct cross-cultural comparisons; Golder & Macy, 2011). Finally, these limitations also prevent the application of new computational methods that are designed to deal with large data sets (e.g., machine learning; see Scharkow, 2013).

Hence, to strategically develop the field of computational communication science (CCS), it seems crucial to improve options for acquiring large communication data resources. This article makes a first step in this regard by systematizing the current data acquisition options and outlining their specific advantages and barriers. To do so, we conducted a qualitative interview study with 22 researchers in the field of CCS. Based on a systematic analysis of the experts' evaluations of the various acquisition options, we develop an agenda on how communication science could overcome the challenges of accessing online data sets in the future.

Literature Review

Data Acquisition Pathways

A first systemization of data acquisition pathways can be derived from previous research and prior discussions of large communication data sources (i.e., Liang & Zhu, 2017; Metaxas & Mustafaraj, 2014; Munzert, Rubba, Meißner, & Nyhuis, 2015). The literature suggests that researchers can gain access to big communication data via five routes:

(1) First, data can be directly acquired through cooperation with the data owners[1]—the provider, producer, or publisher of a website, social media platform, app, online content, or another form of digital communication service (Metaxas & Mustafaraj, 2014). These new gatekeepers of data resources not only provide access to large sets of publicly viewable data (e.g., user-generated online reviews; Emmert &

[1] Although it is questionable whether data can be property (Lohsse, Schulze, & Staudenmayer, 2017), the term is used here as a metaphor for individuals or organizations that are licensed to systematically store and share specific data.

Meier, 2013) but also to material that is only accessible for service providers (e.g., activity logs of a video game; Castronova et al., 2009). Cooperation between data owners and scholars can take different forms (boyd & Crawford, 2012; Metaxas & Mustafaraj, 2014): Data owners may, for example, provide data sets in the form of public competitions (e.g., on www.kaggle.com), fund research via financial or data grants (e.g., Twitter research grant; Twitter, 2014), or collaborate with scholars via research labs—most often with the intention to benefit from the result and improve their own services (Metaxas & Mustafaraj, 2014). Finally, some few data sets are published publicly to support science (e.g., Yahoo: https://webscope.sandbox.yahoo.com; see Metaxas & Mustafaraj, 2014).

(2) Another pathway to collect large communication data is to buy data sets from the data owners or specialized resellers. Selling digital traces such as social media data is a fast growing market (Puschmann & Burgess, 2014). Some data owners sell the right to access their resources directly or through a subsidiary reseller (e.g., Twitter and Gnip; Bruns, 2018a). Moreover, some third-party providers also collect and offer data (e.g., Datasift). Several researchers already made use of these resellers (e.g., Giglietto & Selva, 2014).

(3) Additionally, some platforms and services allow scholars to gather data via application programming interfaces (APIs). These interfaces were originally developed to enable third-party software to interact ("speak") with a platform or service in a standardized manner (Bruns, 2018a; Liang & Zhu, 2017; Lomborg & Bechmann, 2014). Some data owners (i.e., social media platforms) allow searching for and downloading specific data subsets via their API (Bruns, 2018a). Scholars can make use of these API functions to collect data for research purposes. Especially the Twitter API has been used intensively to study diverse topics (Bruns, 2018a). Moreover, several software solutions have been developed to enable researchers to download data from popular social media platforms without extensive programming skills by using the API (e.g., Netvizz for Facebook; Rieder, 2018).

(4) If data cannot be collected via API, researchers can also use various crawling and scraping techniques (Liang & Zhu, 2017; Metaxas & Mustafaraj, 2014; Munzert et al., 2015). Simply stated, this means to copy the relevant content from a large body of websites in an automatized fashion (Liang & Zhu, 2017). Most often scholars use a combination of crawling and scraping methods: During the crawling process, various websites will automatically be visited and downloaded. Afterward, web scraping is used to extract relevant information from the downloaded content (Liang & Zhu, 2017; for an extensive introduction see Munzert et al., 2015). These methods are particularly useful to collect texts in stable and publicly visible online environments such as news pages (e.g., Welbers, van Atteveldt, Kleinnijenhuis, & Ruigrok, 2018).

(5) Finally, researchers can build panels of users and track their communication behavior. The data collection could be based on specific software such as browser plugins that participants install on their computers and that, for example, record incoming and outgoing data packages (e.g., Bodo et al., 2017), as well as mobile apps that keep track of smartphone activities (e.g., Andone et al., 2016). Alternatively, scholars could simply ask their participants to donate chat or browser histories. In the most extreme version of this data acquisition option, scholars set up and monitor a whole communication

environment such as a social media platform by themselves by using open-source software like Social Lab (Garaizar & Reips, 2014).

Factors That Affect Data Quality

The systematization presented here shows that scholars have used various options to obtain large communication data. Their work has also highlighted some of the major limitations and strengths of the various pathways.[2] In the following, we synthesize those factors that have most often been discussed as affecting data quality.

One issue raised rather often in the literature is that providers, publishers, and producers of online services are not (legally) obliged to grant scientific access to their resources (e.g., boyd & Crawford, 2012; Metaxas & Mustafaraj, 2014). Hence, scholars often depend on data owners' willingness to share their data. This has been stressed particularly in the context of data gathering via APIs as the interfaces are completely controlled by the data owners. Hence, API providers are not only able to censor specific data—for example, to promote monetarization models—but can also change the technological specifications of their services at any time, which may harm ongoing data collection processes (Bechmann & Vahlstrup, 2015; Lomborg & Bechmann, 2014; Metaxas & Mustafaraj, 2014; Munzert et al., 2015). Moreover, in recent years scholars have reported that companies (i.e., social media providers) are less willing to grant access to their data via APIs. For example, Facebook Inc. just recently decided to tighten access restrictions for the APIs of all its services (Bruns, 2018b).

A second factor that affects data quality are restrictions in the completeness and structure of the data output. The data structure has been addressed mostly in the literature on APIs and web scraping. These works characterize APIs as tools to collect well-structured, clean data with accurate metadata (e.g., timestamps). In contrast, it is discussed that the output of web crawling lacks this additional information and that intensive data cleansing is needed (Liang & Zhu, 2017; Munzert et al., 2015). Sampling and data completeness have also been addressed most often in the literature on APIs. Although the interfaces of different platforms and services vary heavily in the size and structure of the data they provide (Lomborg & Bechmann, 2014), many seem to have some limitations in the completeness and representativeness of their data output. Most notably, many API providers restrict the amount of data that can be downloaded (Bruns, 2018a; Metaxas & Mustafaraj, 2014). As APIs are usually not well documented, scholars are not able to retrace how the sample was composed (boyd & Crawford, 2012; Lomborg & Bechmann, 2014).

Additionally, resource requirements including financial means and technological resources and skills have been discussed for almost all the methods outlined so far. For example, while it has been suggested that no or only limited technological resources are needed when scholars cooperate with data owners (Metaxas & Mustafaraj, 2014), equipment and methodological skills seem to be a key prerequisite for crawling and scraping (boyd & Crawford, 2012; Metaxas & Mustafaraj, 2014). Metaxas and Mustafaraj (2014) even characterize these methods as "technically the most challenging" (p. 234) approaches to gather

[2] Additionally, inherent disadvantages of working with large communication data sets have been identified (Mahrt & Scharkow, 2013)

big communication data. In contrast, financial resources seem to play a big role when buying data sets from resellers, and it has been assumed that scholars often lack the funding to buy relevant data sets, which might result in a "digital divide" (boyd & Crawford, 2012, p. 647) between well- and poorly funded universities (Bruns, 2013).

Moreover, several legal barriers to collecting large communication data have been discussed. In particular, it has been highlighted that data owners could restrict data access by making use of their copyright and the terms of use of their services. Most of these considerations focus on crawling and scraping methods as even legal experts discuss whether or not web crawling of publicly visible content without the consent of the platform owner is legal and which legal basis is decisive (Dreyer & Stockton, 2013; Jennings & Yates, 2009; Metaxas & Mustafaraj, 2014). Additionally, it has been pointed out that researchers have to make sure to comply with data protection principles of the law and secure users' anonymity (Liang & Zhu, 2017; Mahrt & Scharkow, 2013; Zimmer, 2010). This has particularly been discussed by scholars who designed data gathering tools in the context of user panels on their own (Bodo et al., 2017).

Finally, ethical considerations have been discussed for all data acquisition options—especially the question of whether users need to give informed consent for data gathering. Although there is often no legal issue, as many services, platforms, and apps state in their terms of service that user data can be incorporated for research, it is argued that this does not absolve researchers of the ethical obligation to obtain such consent (Liang & Zhu, 2017; Lomborg & Bechmann, 2014), particularly as users might often not be aware that their data is publicly available (Bodo et al., 2017; boyd & Crawford, 2012; Metaxas & Mustafaraj, 2014).

This brief overview highlights several dimensions that could influence the quality of collecting large communication data—dependency on data owners, lack of data completeness and structure, resource requirements, and legal and ethical barriers. Many of these aspects have been derived from literature on a specific data collection option. So far, a systematic review that compares the options to gather large data sets within one framework of criteria is missing. However, it seems legitimate to assume that most if not all those dimensions can be applied to each of the outlined data collection methods. For example, issues of data sampling have been discussed mostly about APIs. However, as well-established sampling strategies are reaching their limits in the online context in general (Liang & Zhu, 2017; Mahrt & Scharkow, 2013), securing representativeness and completeness of data should be a challenge for all outlined methods. Legal requirements can serve as another example: Although data protection has been discussed mostly by scholars who designed tracking tools (Bodo et al., 2017), these laws also apply to other data acquisition options (Metaxas & Mustafaraj, 2014). Based on these considerations, we suggest that these dimensions could constitute a general framework to characterize the various acquisition options. In the present study we will thus build upon the identified categories to compare the potentials and barriers for all pathways systematically. This can—in a second step—help us identify overarching patterns that limit scientific data accessibility for all pathways. Identifying these "core problems" should provide a good starting point for deriving an agenda to improve scientific access to large communication data resources. Thus, we ask: What are the strengths and challenges of the individual data acquisition pathways? And what patterns can be identified that limit accessibility of large data sets across the data gathering options?

Method

As we were specifically interested in individual experiences and evaluations of various pathways that go beyond general descriptions covered in journal articles (e.g., required financial or technical resources), a guided interview study was carried out among 22 professors and postdoctoral researchers working in the field of CCS. To cover a broad range of perspectives on methodological challenges about data collection, acknowledged experts and newcomers were targeted in the recruitment process. We defined CCS experts as scholars who met two criteria: (1) They published at least one peer-reviewed journal article or peer-reviewed conference presentation in which they applied or studied one (or more) of the data gathering methods we identified in the literature or in data analysis (i.e., secondary analysis; see results). Table 1 shows which data gathering methods have been used or studied in the publications of the participants. (2) They expressed explicitly in their CV or in the interview that computational methods are one of their fields of expertise or main areas of research interest. In the following, all statements by experts are marked with an E, and all statements by newcomers (i.e., scholars not meeting the two criteria) are indicated with an N. Moreover, gender, international experience, and career status (postdoctoral students, junior professors, and professors) were considered relevant factors for a wide array of viewpoints so participants were recruited with regard to diversity in these categories.

Table 1. Participants That Applied or Studied the Respective Data Gathering Method in at Least One Peer-Reviewed Publication or Conference Presentation.

Method	**Participants**
Cooperation	E10, E12, E16, E17, E19
Buying Data	E2, E4, E7, E9, E11
API	E1, E4, E7, E9, E11 E12, E13, N14, N15, E16, E19, N20, E21, E22
Scraping	E2, E4, E6, E7, E11, E12, E16, E17, E19, E21
Tracking	N3, E10, E11, E12
Secondary Analysis*	N3, E10, E12, E18, E22

* Secondary analysis was not identified as an important CCS data gathering method in the literature review but emerged in the data analysis.

The interview guideline contained questions on the interviewees' professional backgrounds and their experiences in using computational methods as well as their evaluations of access and availability of big data for communication research in general and for the five data acquisition pathways identified in the literature in particular. Furthermore, questions about the integration of computational techniques with traditional methods used in social science, interdisciplinary collaboration, teaching of computational methods, and future perspectives in the field of CCS were included in the guideline. However, the guideline was also open for other topics and procedures to access and work with big data deemed relevant by the experts. The interviews were conducted between December 2017 and February 2018, lasted between 30 minutes and nearly two hours (majority: 45 to 60 minutes), and were fully transcribed.[3]

[3] Note that some interviews were conducted in German. Some of the statements presented in the next sections were therefore translated by the authors.

The interviews were analyzed in a qualitative content analysis by two trained coders. Since not all topics addressed in the interviews were relevant to the research purpose, all statements about the access and availability of big data were identified and extracted from each interview for further analysis in a first step (Gläser & Laudel, 2010). In a second step, these statements were coded in a deductive-inductive procedure. The initial code book was developed based on aspects discussed in the CCS literature and included the following categories: data completeness, data structure, resource requirements, and legal and ethical considerations. During the coding process, the two coders extended the code book with new categories and subcategories that emerged from the material (e.g., tools, personal contacts as a resource). By means of these categories, each pathway was characterized and overarching patterns were identified.

Results

The interviewed experts point out that the five methods identified in the literature review offer rich potentials for accessing large data resources. Additionally, some interviewees stress the importance of Open Science repositories for acquiring big communication data. Thus, secondary analysis can be regarded as an additional, sixth pathway. Although the experts acknowledge the potentials of all six methods, they also point out problems and barriers for each. Therefore, an overview of the individual strengths and weaknesses of the distinct methods can be derived from the data we present in the following. For each data acquisition pathway, we focus primarily on the statements of those experts who claimed to have worked with the respective method before. Based on these findings, we finally carve out three overarching themes that determine the resulting data quality and that apply to—more or less—all six methods.

Cooperation

Cooperation with companies that own data (e.g., social network sites, publishers, market researchers) can be initiated by applying for data-related research grants, by participating in data analysis competitions held by the data owner, by using personal contacts in the respective company, as well as by cold acquisition of project partners. However, our experts argue that this form of data acquisition is often heavily limited by the rather low willingness of data owners to cooperate with research institutions (N3, E4, E6, E11, E16, E22). According to our experts, this can be attributed to the fact that many data owners have their own resources to analyze data and strive to protect themselves from breaches of their users' privacy as well as public debates about using private data for research (N3, E16, E18, E22). Cooperation is thus mostly established between single researchers or institutions and data owners and is highly dependent on personal contacts with employees of the companies (e.g., "knowing someone from college who has gone to Twitter or Facebook," E4) and regional proximity of the research institution (e.g., regional publishers; N3, E4).

If researchers are privileged to set up cooperation, data access is characterized by a close proximity to the data owner. On the one hand, this proximity can increase data quality as companies are able to provide well-structured data sets directly from their database (E19) that may include data points not accessible elsewhere (e.g., unpublished articles; E11, N14, E19). Moreover, data collection is mostly well-covered legally by the terms and conditions set by the data owner for their service or platform (E17). On the other hand, data access often comes with strings attached. These strings may range from consulting activities to nondisclosure agreements that prohibit data sharing (E17). Therefore, results often cannot be replicated or validated, which caused one expert to evaluate this method as "the most opaque at all" (E17). Moreover, scholars' dependency on the data owner may also have negative consequences for transparency: Since companies control sample generation, size, and composition and are not particularly transparent about these processes, researchers sometimes lack information to evaluate the sample quality (E2). Moreover, the dependency on the data-providing company may also interfere with research ethics (N20). Most notably, one of our experts reports that companies sometimes try to censor results: "And sometimes the companies themselves will have . . . gatekeeping activities, so for example, naturally they wouldn't want anything that is bad for their company to be revealed, so that they keep an eye on that" (E17).

Buying Data

Experts point out several options for buying data: Researchers themselves but also universities/research institutions (E1, E4) can buy specific data sets from data owners directly, from resellers, or from market research companies. Purchasing data is associated with a comparatively high expenditure of monetary resources (E4, N15, N20, E22), which implies that the possibility of using this data acquisition pathway highly depends on a researcher's financial situation (e.g., being part of a university with access to large data archives or a well-funded research project; E1, E4, E6, E17). However, some experts point out that, despite monetary strains, buying data might sometimes be the only (E2, E4) or the most cost-efficient option: Some data might be "so laborious to collect by myself that I believe the result of a cost-benefit calculation would be that I would buy the data" (E1).

The experts' comments about transparency imply differences between (1) buying data sets and (2) purchasing access to data archives. Buying data sets is described as a black box by some experts: "I don't have any control over the quality of the data. Everything has been preprocessed before the data is released to me" (E12). This means scholars often do not know how well the sample they bought from the reseller reflects the original full data set or whether it has been preprocessed or filtered (E4). Thus, scholars are compelled to "trust the quality management" (E2) of the providers. In terms of purchasing access to data archives, issues of sample structure are discussed. As archives sometimes only comprise a selection of data, "a systematic part of data is not available for research" (E2) through this pathway. Therefore, researchers need to identify potential sampling errors (E2).

The legal situation for buying data is evaluated as "good" (E4), which is attributed to the business relationship entered into by scholars and data providers. However, business contracts oftentimes limit the options to use and publicize the data. Particularly, some experts mention that archives sometimes explicitly forbid systematic search and storage of data or require previous request of permission (E1, N20). With these legal constraints come problems of ethical considerations: "The problem with legal issues is that it is occasionally not allowed to do things that are important to me as a scientist" (E2).

API

Large data sets are also accessible via APIs. The experts mention several interfaces to large communication platforms that vary substantially in the scope of their functions. For instance, at the time of the interviews,[4] the free APIs of Twitter and Facebook had different features and limits: "Twitter has time limits in terms of how far back you can go, Facebook has no time limits so you can go back as far as you want, but Twitter allows you to search with keywords when Facebook doesn't" (E7). To access these interfaces, most experts use R- and Python-packages or other open-source software that provide more elaborated graphical user interfaces. Accessing programming interfaces with these tools is often free of charge (E7, E17) and does not require exceptional technical skills, especially compared with scraping (E7). Even the newcomers in our sample agree that technical handling is not too demanding (N14): "More challenging than creating an online survey but still possible" (N3). Thus, the resource requirements for this data collection method can be described as relatively low. Moreover, the experts point out that the output of APIs often is well structured, is machine readable, and comes with a plethora of metadata (N3, E12, E13, E19).

As a major disadvantage of data access via APIs, the experts state they are "caught in this corset of the API" (N20): Data owners control the process of data generation and selection. They can set limits to time span and amount of data retrievable via API. In addition, public APIs are not necessarily the ones companies use for native development. Thus, the output is often only a small and not further specified segment of a larger and not further specified data set (E1, E7, E21). "I don't know how the data I actually access can relate to the data that exists. So how's the sample?" (E2). While there are well-documented public APIs, companies often have no interest in transparency about the procedures used to generate data output for API requests, and there are no possibilities to validate quality and completeness of data sets (E4, E12, E21). Moreover, platform owners may change the technical specifications and terms of use for their APIs. This enhances resource requirements for long-term projects because codes need to be adjusted continuously (E4). In a worst-case scenario, the data researchers are interested in may not be available anymore (N3, E19, E21). Finally, the type of data accessible via APIs is seen as a strength as well as a weakness. While experts consider the data quality to be equivalent to observational data, because users do not know their data is being used for empirical research, ethical and legal considerations of user consent are also mentioned: "It cannot necessarily be assumed that information or messages that are public were intended to be public" (E4, E11).

Scraping

More than any other data acquisition pathway, scraping depends on researchers' technical skills and resources (E18). This can be attributed to the underlying technical procedure: Scraping most often relies on individually developed code scripts (E11, E13, E19), which have to be adapted to the individual construction of each website or platform from which data should be collected ("I have to retailor this for each [page], because HTML code doesn't scale," N20, E19). The process of creating and adapting the required code is described as technically demanding (especially compared with using APIs) but also inefficient, time-consuming, and sensitive to error (E7, E16, E19). Moreover, the experts mention that the data output of the scraping process often lacks important metainformation and is incomplete (E2, E4, E19). The latter occurs in particular if the code does not cover specific technical characteristics of a website in the sample (E19), if the crawler cannot access certain parts of the website (E2), or if the website is personalized for the researcher's profile (E4). Researchers have to keep in mind that crawling means looking at a website from a user's point of view (which might also be beneficial for scientific research at the same time; E4, E19).

[4] Note that interviews were carried out before Facebook restricted access via API (Bruns, 2018b).

Although crawling and scraping methods might often result in failures, one of our interviewees points out that an advantage of this method is that sometimes sampling errors can be evaluated: "I know exactly which errors I have inserted. So I can look exactly for the sampling errors in my own scripts, I can understand what I did, I can optimize it myself, and I can maintain my own quality criteria" (E2). However, the options to evaluate sampling errors seem to be limited in personalized online environments (E4).

In addition to these technological and resource-related challenges, the experts also mention legal problems related to scraping techniques such as infringements of copyright or a website's terms of use (E2, E16, E21): Several providers of online services (e.g., social network sites; E6, E7, E17, N20) prohibit scraping in their terms of use or the usage of data obtained via scraping for commercial and sometimes also noncommercial purposes. Other data owners try to prevent web crawling with technical measures (N3, E19, E21). As the legal situation is often unspecified, the experts characterize it as a "grey area" (E19, E2, E11). These legal uncertainties render ethical considerations on the part of researchers all the more relevant for scraping: Do users know their data is publicly available and therefore subject to scientific research (E1, N8)? Does the content researchers want to scrape fall under freedom of research (E11)? Or do researchers have to adhere to the terms of service (N20)?

Tracking

Tracking covers many different techniques to observe individual behavior: The experts list smartphone tracking via apps, browser tracking (plug-ins, browser history), and tracking via specialized apps on social media platforms in this category. Tracking is characterized as a resource-intensive data gathering method in the interviews. First, if no tracking tool is available that meets the demands of a research project, scholars need to develop solutions on their own. However, the interviewees state that programming a tracking tool requires complex technical skills and is time-consuming, especially since tools often have to run on different forms of users' hardware and software (E10, N15, N20). Hiring someone to do this requires the investment of large financial resources (E10, N15, E21). Even if a suitable tool is available, recruiting a sample of users requires time and persistence as well because—unlike data access via APIs, via cooperation, or via buying—researchers approach users directly without intermediaries. Often complex measures for recruitment are necessary as the experts note that users are very protective of their data and often are not willing to participate in a tracking study (E10, E16, N20). Moreover, to be ethically sound, researchers have to get users' consent and inform them in detail about the research purpose, create awareness "about the scope of data they're giving to researchers" (E22), and outline options to disable tracking on their devices (E1, E2, E10, E13, E16). As getting users' consent is the main challenge for tracking, researchers often have to work with small, nonrepresentative samples (E4, N5, E10, N20). Moreover, due to self-selection, tracking samples often consist of special users who are either willing to help scientific research, have little privacy concerns, or seek help from interventions (N5, E10). Furthermore, one expert expresses the concern that users might adapt their behavior to the research situation (E4).

However, the interviewees also mention several benefits of tracking that contrast the limitations and resource requirements. First, it is argued that tracking tools allow scholars to directly access user behavior via their devices and thus measure their actions very accurately (E10). In addition, one expert argues that especially self-developed apps provide scholars with a strong degree of control over the data collection process: Researchers have knowledge about and influence on data quality and potential biases, user consent, and technical details of the process of data generation:

> That means when I build an app, I get exactly what I want and because of the structure and the content of the app I know who I am investigating or what kind of people are giving me their data and if they are basically aware of where the boundaries of these data are. (E2)

Secondary Analysis

Another way to obtain data that was not identified in the literature review, but often mentioned in the interviews, is the secondary analysis of already existing data sets. Access to these data sets can be provided by publicly accessible data archives or by other researchers upon request. As several of these data sets exist (E16, E22), this option is at first sight available to every researcher without the use of technical or financial resources. However, "data sets are mostly not systematically stored in a central archive but by the people who collected them at some point, that means somewhere decentralized" (E16). This implies that being able to access already existing data sets depends on the "mercy" of other scholars. For example, one expert mentions that he "was very lucky being part of a big group that achieves a large amount of data" (E22).

Regarding legal, ethical, and quality issues, secondary analyses depend on restrictions of the primary data collection (E4, E17). The experts discuss this aspect of ethical questions about user rights and anonymity: "So it really depends on where that data is initially collected. Have people given consent? Do people sign off their rights when they sign up for any kind of online service?" (E17, E16). The output quality of secondary analyses depends especially on the documentation of these data sets: "If data sets are not well documented then the reuse value is not very high" (N3). But if these data sets are well documented and publicly available, they enhance transparency in scientific research because results can be replicated (E4, E16).

Overarching Themes

Beyond comparable descriptions of potentials and challenges for each pathway, overarching themes that determine the resulting data quality can be identified. Limitations of control, resources, and ethical issues apply—more or less—to all six methods and provide a starting point to develop an agenda to promote the acquisition of large data sets in the future. These overarching themes are summarized in the following.

Limitations of control: For the process of data collection, the interviews reveal a similar pattern for most acquisition pathways: Control over data access determines researchers' scope for action to achieve satisfactory data quality (E11, E13). Categorizing the pathways by control results in two scenarios: Either scholars collect data by themselves (e.g., tracking, scraping) and are therefore in control of the process or they outsource data collection (e.g., API, buying, cooperation, secondary analysis) and sacrifice control to data owners. In the latter scenario, limited control is established mostly by legal and technical barriers. Legal barriers are regulations set up by data owners that limit access to or usage of data (e.g., terms of service, buying contracts, nondisclosure agreements). Technical barriers are the actual process of data collection, which is governed by data owners (e.g., specification of the API). These barriers interfere with researchers' ability to identify, disclose, and address data quality issues (e.g., sampling bias). Therefore, scholars' degree of control over the data collection process directly determines the transparency of their research. Scraping and tracking, on the other hand, provide scholars with higher degrees of control and thus transparency. However, these pathways require higher expenditure of resources (i.e., technical skills, time, and financial means) and more careful ethical considerations on the part of researchers (e.g., user consent). Thus, while a high degree of control over the data collection process enables researchers to meet data quality standards more easily, it also comes with higher resource requirements.

Limitations of resources: The interviews revealed that all six data gathering methods directly depend on scholars' resources, including technical skills (e.g., writing code for web scraping), financial resources (e.g., buying data sets), and personal contacts (e.g., acquiring data sets for secondary analysis). However, as access to most data sets and use of most acquisition pathways do not hinge on one specific type of resource, researchers who lack one resource (e.g., financial means) are not excluded from data access. Rather, other resources (e.g., technical skills) can be used to compensate for the shortage. Additionally, the interviews showed that although limited resources may impair data quality (e.g., small or biased samples), the investment of resources alone does not ensure data quality. Rather, researchers must use their available resources in a way that meets quality and transparency standards (E18). Considering the crucial role of control, expenditure of resources is more likely to improve data quality if it helps researchers control data collection.

Limitations of ethical considerations: Finally, the interviewees point out that acquiring large data sets poses new ethical challenges. First, the experts address a dilemma between the ideals of Open Science to publish original data and promote reproducibility[5] and legal requirements related to many data acquisition pathways (E13, E18). Access to large data sets is often tied to nondisclosure agreements that prohibit data sharing (API, scraping, cooperation, buying). These restrictions limit researchers' ability to uphold ethical standards (replicability, transparency). Using methods that allow researchers to control the process of data gathering (e.g., tracking) could be a solution to this dilemma. However, these methods are often resource intensive—especially when researchers adhere to ethical standards (e.g., clearly explain a tracking procedure and obtain user consent). Second, many interviewees and in particular the newcomers were uncertain about how to ensure the protection of users' rights (E1, E6, N15). Most pathways are not designed to obtain users' consent for data collection (e.g., API, scraping, cooperation). Although users, by agreeing to the terms of service (e.g., of a social media platform), often allow data owners to share their data, ethical questions need to be considered by researchers (e.g., do users know that by agreeing to the terms of use they give consent for scientific use of their data?).

Discussion

In the Internet age, many forms of human communication leave massive digital traces. Analyzing these traces can provide new insights into media content, user behavior, or the structure of social networks in an unobtrusive, real time, or global fashion (Lazer et al., 2009; Liang & Zhu, 2017). These new possibilities

[5] Note that Open Science is not limited to sharing original data sets but also includes other practices (see Nosek et al., 2015).

have triggered a veritable "data rush" (Mahrt & Scharkow, 2013, p. 21) in the past years. The present study among 22 scholars from the field of CCS illustrates the diversity of methods that have been developed to capture the "data gold." However, the experts in the present study also highlight the limitations and hurdles of data collection for CCS. In particular, our empirical analysis reveals three overarching themes that determine the quality of data acquisition: (1) scholars' restricted control over the data gathering process, (2) limitations in the resources required for accessing large data sets, and (3) ethical obstacles. These barriers result in the paradoxical situation that large amounts of digital communication data are created every day, but the availability of scientifically relevant data has not improved alike. We believe this limitation in data access is the weak spot in the development of the field of CCS since it affects both research quality and the kind of questions that can be addressed. To develop an agenda for CCS to overcome these challenges, two perspectives have to be taken into account.

Data Is Power

Our study reveals that scientifically relevant big communication data are mostly in the hands of (a few) individual companies. These data owners decide for which research purposes specific data can be used and, thus, have the power to determine the research questions communication science is able to study. In line with past reasoning, our experts describe this situation as a veritable power imbalance between companies with economic aims on the one hand and scientific interests on the other (boyd & Crawford, 2012; Metaxas & Mustafaraj, 2014). Our findings also show that scholars can only overcome this imbalance by investing enormous amounts of resources. For example, large financial means are required to build scientifically owned panels and track online user behavior. Thus, our experts expressed concerns that this situation may also result in a two-tier academic society in which well-equipped universities can access large communication data while others are not able to do relevant research. Hence, our study supports concerns of a "new digital divide" in academia (boyd & Crawford, 2012, p. 674; Bruns, 2013). Taken together, access to large communication data should not only be conceptualized as a chance for new scientific exploration but rather as an important means of power in the generation of knowledge.

Great Power Comes With Great Responsibility

Paralleling past research, the present study suggests that the mere availability of large data resources does not release scholars from the obligation to apply to well-established quality criteria and deal with ethical considerations (Mahrt & Scharkow, 2013; Metaxas & Mustafaraj, 2014). User consent and securing data protection are issues of great importance to our experts. Moreover, our study shows that all data acquisition pathways come with certain distortions and restrictions about data completeness. Researchers are therefore responsible for investigating and transparently acknowledging the distortions and errors in their data. This is true for small data sets as well. However, since established procedures to ensure sampling quality and ethical principles are much harder to realize in the context of large data resources (e.g., Lomborg & Bechmann, 2014; Mahrt & Scharkow, 2013), scholars have a greater responsibility for their work.

How Can We Empower Science?

Based on the interviews, the aforementioned two perspectives, and the three overarching themes, we can derive five claims that scholars should enhance and demand on the way to a strategic roadmap of CCS.

First, data owners should provide researchers with a legal way to access data if they can demonstrate that their work follows noncommercial interests and is in line with ethical principles (Bruns, 2018b). For example, data owners could provide special APIs that suit scientific research and ensure data protection at the same time (Bruns, 2018b). Alternatively, the legislation could establish a fair use policy that, for example, allows scholars to crawl online data for scientific purposes as long as these practices do not overtax the servers of a service (Dreyer & Stockton, 2013). Moreover, one of the interviewees suggests that fair use policies should not be limited to data access, but should also include data sharing (E13). Such regulations could not only ensure that scientific research is less bound to the financial or institutional resources of individuals but also could support Open Science. However, enforcing these ideas most likely requires communication scholars to organize their demands to have a powerful position in negotiation with companies and policy makers (N3). National and international academic associations such as International Communication Association (ICA) or Association of Internet Researchers (AoIR) could take a leading role in organizing demands, developing ideas, and allowing scholars to speak with one voice.

Second, politicians, funding agencies, and researchers themselves should promote Open Science and data repositories. Along those initiatives, funding lines should also be established to promote the development of panels and science's own tracking software (e.g., Bodo et al., 2017). This is considered crucial for scientific research to maintain a certain independence from the rapidly growing involvement of market research companies in software development for tracking purposes (E10). In other words, scholars need to be equipped with sufficient resources to develop tools that help them overcome the aforementioned power imbalance. Again, academic associations could take a leading role by (1) creating transparency for digital data archives or developing such archives for their members, (2) promoting data sharing in their journals, and (3) assisting in organizing collaborative research projects to collect data that can serve multiple research projects (E4, E11).

Third, there was a great deal of uncertainty among the interviewed experts as to whether data could be obtained and shared ethically without hesitation. Scientific associations and ethic committees could resolve this uncertainty by developing ethical standards as guidelines for future studies. First attempts in this regard have already been undertaken, such as the Association of Internet Researchers' ethics guidelines (see Bruns, 2018a). Because the legal conditions vary greatly from one country to another, it is important that legal pitfalls are identified by legal experts and made available in an easily understandable form to researchers—regional research associations could play an integral role in this context.

Fourth, communication science needs to increase method research to better understand what biases result from the various data gathering methods (e.g., Morstatter, Pfeffer, & Liu, 2014) and how to overcome methodological limitations (e.g., random sampling; Zhu, Mo, Wang, & Lu, 2011). This should ultimately lead to the development of quality criteria and standard methods to secure them. A content analysis of studies using computational methods could furthermore provide a first overview over the data collection methods used in CCS and identify their popularity.

Fifth and finally, technical training should be improved at all career levels to equip scholars with sufficient methodological knowledge. In particular, computational methods should be integrated into existing curricular and new study programs. Taking into account that these methods developed rapidly in the last few years, a useful starting point could be to offer training courses for lecturers.

Limitations and Directions for Future Research

The methodological choices made in this study clearly leave space for improvements and replications. First, we interviewed a convenience sample of experts. Although we undertook several measures to enhance sample diversity—such as recruiting males and females from diverse countries and various scientific career steps—we cannot make sure they covered all possible acquisition pathways or all related potentials and challenges. Especially our focus on researchers at a higher career level (i.e., doctoral degree required) excludes the perspective of younger scholars. This can be justified by the fact that researchers need some scientific experience to serve as valuable informants for our purpose. On the other hand, PhD students might be able to provide valuable perspectives as they might have more time resources to learn and apply new computational methods for data gathering.

Furthermore, the research interests vary heavily among the interviewed experts. Thus, our results should include experiences of many researchers in our cross-sectional discipline. However, as shown, the importance of challenges and potentials differs depending on the research question and type of data. For example, researchers interested in journalistic online content need to pay particular attention to sampling biases while studies tracking behavioral data should be most concerned with user consent. Hence, a more fine-grained analysis divided by data type (see Liang & Zhu, 2017) or research area might be a valuable endeavor for future research.

Additionally, our study focused on the experience of scholars. However, it would be fruitful to also interview data owners and research associations in the future. Their perspectives should help us better understand potential conflicting interests and systemic dependencies. Moreover, we suggest that future studies test the identified potentials and challenges of the acquisition pathways empirically in a systematic manner. For instance, data quality should be compared for data sets obtained via different pathways. As such an endeavor should consider various data types and data sources to make reliable statements, communication science is in demand to give rise to a broad empirical research program that illuminates the methodological challenges computational data collection methods hold for our discipline.

References

Andone, I., Blaszkiewicz, K., Eibes, M., Trendafilov, B., Montag, C., & Markowetz, A. (2016). Menthal: A framework for mobile data collection and analysis. In P. Lukowicz, A. Krüger, A. Bulling, Y.-K. Lim, & S. N. Patel (Eds.), *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing Adjunct—UbiComp '16* (pp. 624–629). New York, NY: ACM Press. https://doi.org/10.1145/2968219.2971591

Bechmann, A., & Vahlstrup, P. B. (2015). Studying Facebook and Instagram data: The Digital Footprints software. *First Monday*, *20*, 12.

Bodo, B., Helberger, N., Irion, K., Zuiderveen, B., Moller, J., van de Velde, B., . . . de Vreese, C. (2017). Tackling the algorithmic control crisis—the technical, legal, and ethical challenges of research into algorithmic agents. *Yale Journal of Law & Technology*, *19*, 133–180.

boyd, d., & Crawford, K. (2012). Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, Communication & Society*, *15*, 662–679. https://doi.org/10.1080/1369118X.2012.678878

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

Bruns, A. (2018a). Big social data approaches in Internet studies: The case of Twitter. In J. Hunsinger, L. Klastrup, & M. M. Allen (Eds.), *Second international handbook of Internet research* (pp. 1–17). Dordrecht, The Netherlands: Springer Netherlands. https://doi.org/10.1007/978-94-024-1202-4_3-1

Bruns, A. (2018b). *Facebook shuts the gate after the horse has bolted, and hurts real research in the process.* Retrieved from https://policyreview.info/articles/news/facebook-shuts-gate-after-horse-has-bolted-and-hurts-real-research-process/786

Castells, M. (2002). *The Internet galaxy: Reflections on the Internet, business, and society*. Oxford, UK: Oxford University Press on Demand.

Castronova, E., Williams, D., Shen, C., Ratan, R., Xiong, L., Huang, Y., & Keegan, B. (2009). As real as real? Macroeconomic behavior in a large-scale virtual world. *New Media & Society*, *11*, 685–707. https://doi.org/10.1177/1461444809105346

Dreyer, A., & Stockton, J. (2013). *Internet "data scraping": A primer for counseling clients.* Retrieved from https://www.law.com/newyorklawjournal/almID/1202610687621/

Emmert, M., & Meier, F. (2013). An analysis of online evaluations on a physician rating website: Evidence from a German public reporting instrument. *Journal of Medical Internet Research*, *15*. https://doi.org/10.2196/jmir.2655

Garaizar, P., & Reips, U.-D. (2014). Build your own social network laboratory with Social Lab: A tool for research in social media. *Behavior Research Methods*, *46*, 430–438. https://doi.org/10.3758/s13428-013-0385-3

Giglietto, F., & Selva, D. (2014). Second screen and participation: A content analysis on a full season dataset of tweets. *Journal of Communication*, *64*, 260–277. https://doi.org/10.1111/jcom.12085

Gläser, J., & Laudel, G. (2010). *Experteninterviews und qualitative Inhaltsanalyse als Instrumente rekonstruierender Untersuchungen* [Expert interviews and qualitative content analysis as instruments for reconstructing studies] (4th ed.). Wiesbaden, Germany: VS Verlag für Sozialwissenschaften.

Golder, S. A., & Macy, M. W. (2011). Diurnal and seasonal mood vary with work, sleep, and daylength across diverse cultures. *Science*, *333*, 1878–1881. https://doi.org/10.1126/science.1202775

González-Bailón, S., Banchs, R. E., & Kaltenbrunner, A. (2012). Emotions, public opinion, and U.S. presidential approval rates: A 5-year analysis of online political discussions. *Human Communication Research*, *38*, 121–143. https://doi.org/10.1111/j.1468-2958.2011.01423.x

Hootsuite & We Are Social. (2018). *Digital in 2018.* Retrieved from https://digitalreport.wearesocial.com

Jennings, F., & Yates, J. (2009). Scrapping over data: Are the data scrapers' days numbered? *Journal of Intellectual Property Law & Practice*, *4*, 120–129. https://doi.org/10.1093/jiplp/jpn232

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

Liang, H., & Zhu, J. J. H. (2017). Big data, collection of (social media, harvesting). In J. Matthes, C. S. Davis, & R. F. Potter (Eds.), *The international encyclopedia of communication research methods* (pp. 1–18). Hoboken, NJ: John Wiley & Sons. https://doi.org/10.1002/9781118901731.iecrm0015

Lohsse, S., Schulze, R., & Staudenmayer, D. (Eds.). (2017). *Trading data in the digital economy*. Baden-Baden, Germany: Nomos.

Lomborg, S., & Bechmann, A. (2014). Using APIs for data collection on social media. *The Information Society*, *30*, 256–265. https://doi.org/10.1080/01972243.2014.915276

Mahrt, M., & Scharkow, M. (2013). The value of big data in digital media research. *Journal of Broadcasting & Electronic Media*, *57*, 20–33. https://doi.org/10.1080/08838151.2012.761700

Metaxas, P., & Mustafaraj, E. (2014). Sifting the sand on the river bank: Social media as a source for research data. *it - Information Technology*, *56*. https://doi.org/10.1515/itit-2014-1047

Morstatter, F., Pfeffer, J., & Liu, H. (2014). *When is it biased? Assessing the representativeness of Twitter's streaming API*. Retrieved from http://arxiv.org/pdf/1401.7909v1

Munzert, S., Rubba, C., Meißner, P., & Nyhuis, D. (2015). *Automated data collection with R: A practical guide to web scraping and text mining*. Chichester, UK: Wiley.

Nosek, B., Alter, G., Banks, G., Borsboom, D., Bowman, S., Breckler, S., . . . Yarkoni, T. (2015). Promoting an open research culture. *Science*, *348*(6242), 1422–1425. doi:10.1126/science.aab2374

Puschmann, C., & Burgess, J. (2014). The politics of Twitter data. In K. Weller, A. Bruns, J. Burgess, & M. Mahrt (Eds.), *Digital formations: Vol. 89. Twitter and Society* (pp. 43–54). New York, NY: Lang.

Rieder, B. (2018). *The politics of systems: Thoughts on software, power, and digital method.* Retrieved from http://thepoliticsofsystems.net/

Scharkow, M. (2013). Thematic content analysis using supervised machine learning: An empirical evaluation using German online news. *Quality & Quantity*, *47*, 761–773. https://doi.org/10.1007/s11135-011-9545-7

Twitter. (2014). *Introducing Twitter data grants.* Retrieved from https://blog.twitter.com/engineering/en_us/a/2014/introducing-twitter-data-grants.html

Welbers, K., van Atteveldt, W., Kleinnijenhuis, J., & Ruigrok, N. (2018). A gatekeeper among gatekeepers. *Journalism Studies*, *19*, 315–333. https://doi.org/10.1080/1461670X.2016.1190663

Zhu, J. J. H., Mo, Q., Wang, F., & Lu, H. (2011). A random digit search (RDS) method for sampling of blogs and other user-generated content. *Social Science Computer Review*, *29*, 327–339. https://doi.org/10.1177/0894439310382512

Zimmer, M. (2010). “But the data is already public”: On the ethics of research in Facebook. *Ethics and Information Technology*, *12*, 313–325. https://doi.org/10.1007/s10676-010-9227-5