The Quality of Big Data: Development, Problems, and Possibilities of Use of Process-Generated Data in the Digital Age

Baur, Nina; Graeff, Peter; Braunisch, Lilli; Schweia, Malte

The Quality of Big Data. Development, Problems, and Possibilities of Use of Process-Generated Data in the Digital Age

Nina Baur, Peter Graeff, Lilli Braunisch & Malte Schweia *

Abstract: *»Die Qualität von Big Data. Entwicklung, Probleme und Chancen der Nutzung von prozessgenerierten Daten im digitalen Zeitalter«*. The paper introduces the HSR Forum on digital data by discussing what big data are. The authors show that big data are not a new type of social science data but actually one of the oldest forms of social science data. In addition, big data are not necessarily digital data. Regardless, current methodological debates often assume that "big data" are "digital data." The authors thus also show that digital data have a big drawback concerning data quality because they do not cover the whole population – due to so-called digital divides, not everybody is on the internet, and who is on the internet, is socially structured. The result is a selection bias. Based on this analysis, the paper concludes that big data and digital data are data like any other type of data – they have both advantages and specific blind spots. So rather than glorifying or demonising them, it seems much more sensible to discuss which specific advantages and drawbacks they have as well as when and how they are better suited for answering specific research questions and when and how other types of data are better suited – these are the questions that are addressed in this HSR Forum.

Keywords: big data, mass data, process-generated data, process-produced data, digital data, digital methods, computational social sciences, historical sociology, survey methodology, corpus linguistics, social science methodology, data quality, social research.

* Nina Baur, Technische Universität Berlin, Institut für Soziologie, Professur für Methoden der empirischen Sozialforschung, Fraunhoferstraße 33-36 (Sekr. FH 9-1), 10587 Berlin, Germany; nina.baur@tu-berlin.de.
Peter Graeff, Christian-Albrechts-Universität zu Kiel, Institut für Sozialwissenschaften, Professur für Soziologie und empirische Sozialforschung, Westring 400, 24118 Kiel, Germany; pgraeff@soziologie.uni-kiel.de.
Lilli Braunisch, Technische Universität Berlin, Institut für Soziologie, Fachgebiet Methoden der empirischen Sozialforschung, Fraunhoferstraße 33-36 (Sekr. FH 9-1), 10587 Berlin, Germany; lilli.braunisch@innovation.tu-berlin.de.
Malte Schweia, Christian-Albrechts-Universität zu Kiel, Institut für Sozialwissenschaften, Professur für Soziologie und empirische Sozialforschung, Westring 400, 24118 Kiel, Germany; mschweia@soziologie.uni-kiel.de.

Historical Social Research 45 (2020) 3, 209-243 | published by GESIS
DOI: 10.12759/hsr.45.2020.3.209-243

1. What are "Big Data"?

In most current methodological debates, "big data" appear to be a relatively new phenomenon, arisen in the last two decades. Despite this seeming novelty, "big data" or, as they have been called in older debates, "process-produced mass data" (Baur 2009, 10) have been used both by scientists and non-scientists (such as politics, pubic administration, companies, or the media) for more than 200 years. Early forms of big data are, e.g., census data and other forms of public administrative data (also: public administrational data), archival data (such as church registers), or newspaper data. These data have been widely used especially in Germany and other European countries for more than 200 years, as there historically, public administrative data were developed as an instrument for modern administration, government, and exercising power during modernization (Wallgren and Wallgren 2014).

The methodological reflection on big data is as old as the use of big data for research. In fact, while, e.g., French sociology has just been discovering the use of statistics as a means of power in recent decades (e.g., Desrosières 2005, 2011; Thévenot 2011, 2016; Salais 2012; Whiteside 2015; Amossé 2016; Behrisch 2016; Diaz-Bone 2016; Diaz-Bone and Didier 2016; Speich Chassé 2016), the methodological critique of an unreflected and positivist use of big data in public administrative statistics (and of a positivist use of single-case case studies in historical sciences) was one of key incentives for founding German sociology: In Germany in the early 20th century, sociology was supposed to be a complementary science to statistics, economics, and historical sciences in order to provide methodological reflection and control of an unthoughtful use of big data and other data (Baur et al. 2018). In the last 120 years or so, there have been great methodological advances concerning methods of big data analysis (Baur 2009), and it cannot be stressed enough that – considering the amount of time invested in improving this methodology – it would be fruitful to ensure that this knowledge is not lost.

So all in all, neither big data nor the methodological reflection on them are new. Still, this does not mean that nothing has changed. However, the changes have occurred in different spheres than current methodological discourses often imply. Namely, what is new is not the *existence* of big data. Rather, what is new is that in the last decades, the *variety and accessibility* of big data as well as the possibilities of big data *analysis* have largely increased.

In order to understand these changes, it has to be acknowledged first that there are two *types of big data* with fluent boundaries:

- *Traditional-Type Big Data:* The oldest form of big data are *administrative data* which are a by-product of organizational and administrative processes, e.g., register data, other public administrative data, newspaper and other media data and companies' customer data. Since the 1970s, these data have increasingly become "digital data" in the sense that they

have been increasingly digitalized, i.e., either directly collected by digital means or first collected offline and then digitalized for storage and data management. Since the 1990s, these data have been increasingly made accessible for scientific analysis, for example in Germany by implementing Research Data Centres and the RatSWD. In parallel, since the 1970s, data have almost solely been analysed digitally, first by using classical statistical analysis techniques (Baur and Lamnek 2016) and recently also by using text mining (Manderscheid 2019a) and other big data analysis techniques (Riebling 2019).

- *New-Type Big Data:* In addition, big data can also be a *by-product of digital communication* in the Web 2.0, such as web server logs and log files (Schmitz and Yanenko 2019), websites (Schünzel and Traue 2019), blogs (Schmidt 2019), social media (Schrape and Siri 2019), web videos (Traue and Schünzel 2019), gaming data (Bottel and Kirschner 2019), digital self-measurement data (Koch 2019) or Twitter, chats and other user-generated data (Mayerl and Faas 2019; Nam 2019), geo data (Lakes 2019), and geo tracking (Kandt 2019) as well as data created by surveillance companies, wearables such as smartwatches, smart eyewear, and in smart houses in smart cities. In contrast to traditional types of big data, these new types of big data usually are a by-product of using these data in the internet, and data production is a lot faster than in the past – this property of new-type big data is called "*velocity*" (Lane 2001, Weichbold et al. 2020 – in this issue).

If one looks at common definitions of big data, it is obvious that it is not clear what big data are. A property that is seen as typical for big data both in the discussions on traditional-type big data (e.g., Baur 2009) and on new-type big data (Lane 2001, Weichbold et al. 2020, in this issue) is that big data have a large *volume*. That means both that there are a lot of data and that data sets are large. This property is very often combined with the *idea* that big data do not contain a sample but cover the whole *population* and therefore sampling and generalization strategies are not necessary, as one already covers the full population. Note that the latter is an idea which usually is not true in practice, and actually systematic exclusion of parts of the population are one of the biggest weaknesses of big data – a point we discuss in more detail below.

A second property of both traditional-type big data and new-type big data is *variety*, meaning that in a given big data set, different data types are usually mixed: Big data sets usually consist of both qualitative and quantitative components from a variety of data types (e.g., numerical, verbal, and visual data) and data sources. The result is that the data structure is very complex and increases the requirements for and challenges of data analysis – and big data analyses are always *mixed methods analyses* (Baur et al. 2017).

All in all, big data are not a new type of data and may or may not be digital. If one accepts this notion, the question arises as to what the specific properties of new-type big data are compared to traditional-type big data.

In this paper, we will approach this question by first discussing the history of big data and their various forms. We will show in more detail than in this brief introduction that big data are not new but actually one of the oldest forms of social science data – *the actual methodologically relevant difference is not between traditional-type and new-type big data but between process-produced data (which may or may not be) and research-elicited data (which are rarely big) – and the difference is not about data size but about data quality*, an issue that is discussed in more detail in the papers of this HSR Forum. Moreover, what counts as data of high quality depends on the data themselves and the specific application in the scientific fields (see Graeff and Baur 2020, in this issue).

We will also show that *big data are not necessarily digital data and that research-elicited data are not necessarily non-digital*. This is important because in current methodological debates, it is usually assumed that "big data" are "digital data" and that digital data have a better quality than non-digital data.

We will therefore show, second, that digital data have a big drawback in data quality because they do not cover the whole population – due to so-called *digital divides*, not everybody is on the internet, and who is on the internet, is socially structured. This is methodologically important because it thus means that when using digital data, specific *blind spots* are created or specific phenomena are noticed that actually do not exist (e.g., fake news, etc.). We explore the type of blind spots by discussing who is excluded from the internet and for what reasons. Since (deliberately) *biased or falsely created data* are hard to tackle and generate a huge research field on their own (Vargo et al. 2017), we leave these out of our considerations.

The result of our discussion is that *big data and digital data are data like any other type of data* – they have both advantages and specific blind spots. So rather than glorifying or demonising them, it seems much more sensible to discuss, which specific advantages and drawbacks they have as well as when and how they are better suited for answering specific research questions and when and how other types of data are better suited – these are the questions that are addressed in this HSR Forum. We thus conclude by introducing the papers of this forum.

2. The History of Big Data

2.1 The Invention of Big Data as a Tool of Power during Modernisation

The history of big data starts with public administrative data, especially census data. Depending on how much one wants to go back in time, one could start telling the history of big data beginning with the ancient world when ancient governments tried to do censuses for administrative planning, e.g., in order to assess how many taxes they could raise, how many people they could recruit for the military and how many crops to plant. For example, one of the first censuses was conducted in Babylon around 3,800 BC and in Egypt around 3,000 to 2,000 BC.

Alternatively, one could start the history of big data with the invention of modern bureaucracy between the 16th and 18th century during the times of absolutism and cameralistics (*Kameralistik*) in Europe and especially in Prussia (which was a part of the "Holy Roman Empire of the German Nation").

During this time, first, the *variety of data collected increased*: While governments had been continuously collecting census data since antiquity, now other government agencies started collecting data, too. For example, the police started collecting crime statistics. When social security systems were invented, each so-called social security agency started collecting data on their recipients. Companies were legally required to regularly report firm statistics in order to be able to calculate a country's productivity and so on. In parallel, economic actors and NGOs started collecting data as well; e.g., each company typically collects data on their customer, for every order that is placed. Another traditional type of big data are news, because each newspaper regularly produces whole volumes of data.

Second, *fiscal accounting* became an internal part of modern administration. According to Max Weber, the discovery of modern fiscal accounting and systematic use of data is an essential condition for the development of modern capitalism (Collins 1980). So, around 1800, modern states discovered public administration and started using not only census data but also other types of data for collecting information on both their citizens and their economy in order to assess how their societies worked. From now on, modern states had two main sources of power: the military and big data. Thus, big data were consolidated as a *tool of power and dominion* in the competition of nations. In addition to volume and variety, this is the third property that traditional-type and new-type big data share: They have *value*. While in the past this value became visible when big data were utilized by sovereign states as an instrument of power, the value of new-type big data is much more tangible in the form of companies' financial resources that can be sold.

Third, in the 1800s, governments not only started to collect big data more systematically but they always also invented *statistics* as a science in order to analyse these big data and to use this power source even more effectively. In contrast to other sciences (again especially in the German-language context), statisticians were typically closely related to the government. What was typical for statistics of that time was that many statisticians believed that these data were "real" in so far as they gave an "objective" image of society – a misconception that can be found in current debates on big data, too.

As can be seen from this historical overview, big data are a very old data type and one of the oldest data types used in the social sciences. For example, Friedrich Engels's (1845) study on the British working class, Émile Durkheim's (1897) suicide studies, and Max Weber's (1906-1922) study on the rise of modern capitalism all heavily relied on analysing big data. In the beginning, big data were always produced as analogous data, but since the 1970s, when computer sciences started first administering and analysing these data, big data have been increasingly digitalized, and in the recent decades these data are increasingly produced digitally, also. Thus, with traditional-type big data, one cannot always say clearly if they are analogous or digital data because there might have been a switch in data production in the course of their history. Moreover, as the case of newspapers illustrates, traditional-type big data might be simultaneously digital and analogous, as most modern newspapers still have a printed version but provide the same information online.

2.2 Social Theory and Social Science Methodology as Countermeasures against Big Data Quality Issues

What is important for social science methodology is that in the 19th century, big data were the most commonly used data type used in analyses of societies, and most statisticians believed that these data provide a "true" and "objective" image of reality. This belief in the objectivity of big data is shared by many current scholars in the field of new-type big data and is called "*validity*." The belief is that big data cover the full population without any distortions or errors which also means that sampling is not necessary.

However, as the long methodological discourse on big data has proven over and over again, this is a misconception (for a summary, see Baur 2009): In contrast to common belief, big data typically do not cover the full population. Instead, (a) depending on which institution collects data according to which rules, and also (b) depending on how the persons data are collected on react to data collection (e.g., nonresponse, purposefully wrong answers, etc.), big data are typically characterized both by biased samples and lack data quality (Bick and Müller 1984). The institutional selection bias especially often results in huge differences between the target population and the sampling population (Baur 2009). For example, what governments usually want to know about their

population is something about their whole population, but what they usually are actually able to collect data on is the resident population. In the case of Germany, several groups of people are not covered by census data, such as people illicitly living in Germany or Germans who live outside the country. Very often, only citizens are covered in census data (Wallgren and Wallgren 2014, 131ff.). Therefore, even if one uses census data, they will not contain data on all people living in a given territory. In addition, the groups covered by a country's census may change from census to census. Finally, data collection (especially of big data) always takes some time – and during this time, the population may change. Therefore, the actual meaning of "full population" must be defined in advance.

It is therefore no surprise that early social science research pointed out that all data (including big data) are socially constructed (Baur et al. 2018). Similar to surveys, when constructing sets of big data, by selecting a specific target population, specific persons are excluded from the data set. By asking specific questions, concepts are measured in a specific way. For example, Thorvaldsen (2009) shows for Scandinavia that there are about five different measurement concepts on how to measure ethnicity and, depending on which concept is used, the size of each ethnic group will vary widely.

Despite the obvious sampling and data quality issues of big data, similar to popular belief today especially in computational social sciences, 19th century statisticians typically believed that big data provided an objective and complete image of society – a notion that was strongly criticized by early German sociologists. More precisely, the criticism of an unreflected use of big data was one of the main reasons for the founding of German sociology (Baur et al. 2018). More specifically, the "Verein für Socialpolitik" was founded in 1873, followed by the "German Sociology Foundation" ("Deutsche Gesellschaft für Soziologie," DGS) in 1909. German sociologists criticised both statistics and historical science of their time. Their main argument was that data do not speak for themselves. Because data are constructed, they need to be reflected – if researchers do not reflect the construction process of data, they adopt and replicate existing power structures in the field in their research. Instead, researchers need to use both social theory and social science methodology for reflecting these data and for assessing how data are distorted (Baur 2005, 24-38). The newly-founded discipline of sociology aimed at being a complementary science which was supposed to provide both statistics and historical science with theory and methodology in order to learn how to assess data quality and handle data properly (Baur 2008; Baur et al. 2018).

Early social science methodology explored two paths in handling data quality issues of big data: one was developing a *data lore for big data* (Bick and Müller 1984) which aimed at assessing the advantages and disadvantages of big data (see Graeff and Baur 2020 in this issue for more detail).

One of the main disadvantages of process-generated data was that researchers could not control the process of data production and therefore had to live with whatever blind spots the data had (Baur 2009). Therefore, the second strategy of early social science methodology was encouraging social scientists to collect their own data – the so-called *research-elicited data*. Research-elicited data typically have errors, too. However, in contrast to process-produced data, researchers can control which types of errors occur and thus better reflect and handle them methodologically.

Building on this idea, there have been some methodological innovations in the early 20th century. For quantitative research, the two most important innovations were first of all, inventing *research-induced data collection*, such as "enquêtes" (*Enquêten*) and surveys (*standardisierte Befragung*).

Innovations on *sampling strategies* soon followed, the most important sampling innovation in quantitative research being the invention of random sampling in the 1930s – for this methodological innovation, the market research institution "Gallup" was especially historically important: In 1936, the US magazine, "The Literary Digest," tried to sample the full population of its about 2,000,000 readers. Using these big data, the magazine predicted that Alf Landon would win the presidential election. In contrast, George Gallup randomly sampled only 50,000 people and predicted that Franklin D. Roosevelt would win the election which proved to come true.

These methodological tools for research-elicited data collection and sampling were improved in the next decades, and both data collection and sampling techniques were more and more improved, but also the error lore was more and more refined. Since then, the discussion on research-elicited data has been dominating social science methodology. While for large parts of the 20th century quantitative social research was used to discuss various errors that could arise from research-induced data separately in various disciplines, the turn of the 21st century saw quantitative social science research having integrated all these ideas into the concept of the "survey life cycle." The different concepts of errors were integrated into the concept of the "total survey error" (TSE; Baur 2014).

What is important to stress in the context of debates on big data is that, in current debates, research-elicited data and sampling are sometimes stylized as a "problem." However, in fact, when they were developed, they were meant as a solution to the obvious drawbacks of big data and other process-generated data.

2.3 The Resurgence of Big Data since the 1960s

Because research-elicited data could be methodologically controlled and reflected, they had some obvious advantages concerning data quality and sampling. However, it was mostly due to the different kinds of research questions social scientists were interested in the early 20th century (Scheuch 1977) that

research-elicited data slowly replaced process-generated data as the dominant data type in social science research between the 1930s and 1950s. In the post-war period, social scientists hardly ever used process-generated big data any more (Baur 2005, 38-45; Baur 2009, 10). However, since the 1970s, big data started resurging as a source of social science research due to a number of developments (Baur 2009, 10-11):

1) *Paradigm Shift within Theory:* Within some research fields, research questions (Bick and Müller 1984, 125-126) and theoretical approaches shifted. While there was a time in the 1970s and 1980s when rational choice theory dominated social science discourse and individual persons were the typical unit of analysis in social science research, social scientists in recent decades have been increasingly interested in institutions, organizations, and long term social change. Especially for longitudinal questions, process-produced data often are the only option, as no research-elicited data exist for many research questions – simply, because no-one thought the question important enough 30 years ago to collect data on it (Baur 2004, 2005).
2) *Paradigm Shift within Methodology:* In some research fields there has been a paradigm change in methodology. Examples are sociology of deviance and economic sociology. In these fields, due to the nature of the research problem, it is hard to collect very good research-elicited data. For example, in sociology of deviance, usually people who are deviant will hesitate to admit this in surveys. In fields like economic sociology, the typical unit of analysis is not an individual person but a company or market, making it difficult to work quantitatively. Historically, these research fields had been dominated by qualitative research. However, since the 1960s, they, too, increasingly tried to work quantitatively (Bick and Müller 1984, 125-126). Due to the nature of the research question, in these fields, quantitative data are usually big data, resulting in a strong orientation towards either mixing different types of process-produced data (Baur 2011) or mixing process-produced data with qualitative research-elicited data (Baur and Hering 2017).
3) *Development of IT:* While shifts in social theory and social science methodology provided the need for using big data again, other developments provided the opportunity, the first one being the enormous advances in IT since the 1960s. The invention of computers facilitated data management and made preparation and data analyses of large-scale administrative data possible for a broad range of researchers, which would not even have been thinkable 60 years ago (Baur 2005). In addition, data that had to be originally collected analogously could now increasingly be collected digitally (Seysen 2009). This is true not only for public administrative data but also for survey data, as the development of computer-aided and mobile survey methods in the recent decades illustrate.

4) *Data Ownership and Reflexion of Data as a Power Source:* Due to the increasing democratization of society, the questions of whether governments (and other institutions) may use data as a tool of power and who actually owns data have received public attention. This is especially true for Germany, which has suffered from two regimes (Nazi Germany and the GDR) that misused big data for controlling the population, identifying persons who were "undesirable" from the point of view of the regime in order to imprison them and – in the case of Nazi Germany – killing them off (Korte 2004). Thus, since the 1960s, the main question in the (German) debate has been: Who owns the data? Since then, there have been two contrasting lines of argument.
   a. *Data Privacy:* The first line of argument is that the people whom data are collected on are the owners of the data, having a right to data privacy and informational self-determination (*informationelle Selbstbestimmung*; Mühlichen 2019). The idea is that, for example, it should be a person's own right to decide by themselves if anyone else can collect data on their age, income, and so on. Similarly, companies should be able to decide by themselves which business information they want to reveal and which they do not want to reveal. This results in the demand for strong data privacy, i.e., that each individual person and company should have the right to decide which data are collected and stored on them and which data they want to reveal. Note that there is an implicit conflict between individual persons' and companies' interests here, which is rarely discussed in data ownership discourses but has been becoming increasingly important with new-type big data: Individuals might have information on other persons or companies, e.g., when this information is relational, as is the case with friendships or customer data – it is unclear whom the data belong to in this case.
   b. *Open Data:* The second line of argument contradicts this viewpoint and argues that, first, the state needs to collect data for effective policymaking and public planning and that, second, the general public has a right to be informed, e.g., about dubious business practices or criminal individuals. Third, the social sciences and other disciplines need data, too, in order to do research – which is why there is a strong demand that all data should be open data.

As can be seen from these lines of argument, there is a clear and irreconcilable conflict of interests between these perspectives. In the German context, this conflict of interest has been framed as a conflict between "the state" and "the public" on the one hand and "research subjects," i.e., individual people, on the other hand. This conflict was mediated by introducing legislation as early as the 1980s: The interests of research subjects are typically protected by data protection laws (*Datenschutz*; RatSWD

2017a). In addition, there is a strong debate on research ethics in the social sciences (RatSWD 2017b). As data privacy does not only address data collection but also archiving, data management, and data analysis, the debate on archiving and on data privacy are closely linked, resulting in quantitative social science methodologists developing procedures for being able to store and re-analyse sensitive data without compromising data privacy as early as the 19670s (RatSWD 2018).

5) *Accessibility:* Public administrations started making data easily accessible for a broad range of researchers. Examples in Germany are data made available by institutions such as the Federal Statistical Office, the Regional Statistical Offices, and so-called Research Data Centres. In order to improve the availability of big data, even an own council, the RatSWD, was created, and a series of workshops was introduced in order to train students and researchers how to use these data, which today are often already prepared for data analysis by the data-providing institution.
6) *Methodological Reflexion of and Data Lore for Big Data:* In the 1970s, the increased accessibility and use of big data in social science research led to a resurgence of methodological debates on how big data can be used in social science research, resulting in a refined *data lore* by the 1980s (Bick and Müller 1984) in order to assess the quality of big data – an issue that Graeff and Baur explore further in this forum of *Historical Social Research*. Large parts of this ongoing debate have been discussed in this journal as well as in the publications of RatSWD.

3. Methodological Properties of Big Data, or: What is New in New-Type-Big Data?

To sum up, big data are not a new data type at all but instead probably the oldest type of data used in social science analysis. Therefore, there is also a long tradition of reflecting on the properties as well as strengths and weaknesses of big data in social science methodology. From this discourse, what do we know about big data?

First, *big data are just data* in the sense that – like any data – they have strengths and weaknesses. As not only big data but all data are faulty in some way, the question of whether or not big data are suitable for research cannot be clarified once and for all on a general level but depends both on the specific type of big data used and on the research question – and can only be answered in comparison to other data types available. Two further prerequisites for using specific types for data for answering a specific research question are that (a) data have to exist at all and (b) researchers can gain access to these data (Baur 2009).

Second, *big data are a subtype of process-produced data (also: process-generated data, natural data)*. Process-produced data are by-products of social processes and produced for specific purposes, e.g., public administrations handling their clients or companies managing their relationships to their customers. In contrast, *research-elicited data (also: research-induced data)* are specifically produced by social scientists for research purposes, and there are two subtypes of research-elicited data: Primary data are collected by the same researchers who analyse the data. Alternatively, researchers can re-analyse data that were collected by other researchers in secondary data analysis (Baur 2009). No data (regardless or whether they are process-produced or research-elicited) give a "true," "objective" image of social reality (Baur et al. 2018). Instead, there is always a difference between social reality and data concerning two aspects (Baur 2009):

- *Sampling:* The target population is usually never fully represented in the data – instead, the data set represents a sample. How well the sample represents the population depends on how the selection process was organized and biased – and knowing the selectivity of the data in turn influences if and how results drawn from specific data can be generalized. When producing research-elicited data, scholars typically define a target population and select cases by either random sampling or purposeful sampling. By controlling the process of case selection, scholars can assess which types of cases are selected and which are not, or in other words, which sampling errors might occur. In contrast, for process-produced data, cases are selected, too. However, the logic for case selection follows the institutional needs – when using these data, researchers have to live with whatever selectivities are created by this. For example, the German Institute for Employment Research (iab) possesses very large data sets which are a by-product of administering the social security data of employees. However, this data set does not contain any data on self-employed persons who account for about 10% of all employees in Germany (Maier and Ivanov 2018, 14) and are a very distinct group of the working population. Depending on the specific research question, this may or may not be a problem – the methodologically important point is that researchers have no control over the distortion of samples of process-generated data.
- *Data Collection:* The same is true for data collection. When collecting research-elicited data, researchers control the process of data collection and therefore not only can try to minimize measurement errors but – as some errors are unavoidable – also do usually know what types of errors are produced. In contrast, with process-produced data, the data producing institution decides what type of data to collect and also which measures of quality control to take. Again, scholars using these data have to live with whatever information is collected.

According to Bick and Müller (1984), the selectivities and blind spots on process-produced data are not errors because it is not the data-producing institutions' fault that their needs might not fit the needs of social scientists. Still, when using process-generated data, social scientists need to reflect the specific properties of these data – a process Bick and Müller (1984) called "data lore" (*Datenkunde*) in contrast to the "error lore" (*Fehlerkunde*) that is used for reflecting research-elicited data (see also Graeff and Baur 2020, in this issue).

Third, *big data can be distinguished from other types of process-produced data by their volume*. While there are many examples of small-volume process-generated data (e.g., historical documents, pieces of art, autobiographies, architecture etc.) big data are usually mass data, that is, similarly structured, process-produced data in very large quantities such as administrative records. Traditionally, both quantitative research-elicited data and big data have been analysed using classical statistical analysis techniques (Baur and Lamnek 2016). Since the 1970s, more and more complex statistical procedures as well as new techniques such as corpus linguistics and quantitative content analysis (which all relied on the heavy use of computers) could be used. In recent years, analysis techniques such as text mining (Manderscheid 2019) and other big data analysis techniques (Riebling 2019) became accessible for social scientists, too.

Fourth, *big data are usually mixed data – they typically combine a variety of data types*. For example, a public record may contain a written letter or a photograph, or a social media entry may contain log file data, comments, and pictures. Big data analysis therefore usually is mixed methods analysis.

Fifth, big data have always been a means of *power and dominion* – they always had *value*.

Sixth, *big data may or may not be digital*. This is important because in current methodological discourses, big data are often confused with digital data, a point we will come back to in the next chapter.

All these properties characterize both traditional and new-type big data. Therefore, the question arises what is actually new with new-type big data. We argue that – in contrast to common social science discourse – these are actually only four properties:

1) *Shift of Power Balances from the State to MNCs:* In comparison to traditional-type big data, the process of data production and data ownership have changed. While public administrative data were most often produced by government agencies, and citizens often had to be forced to provide data, the new-type big data are typically collected by large Multinational Corporations (MNCs). Currently, most of these companies are US-American companies (such as Facebook, Google, and Twitter). This has led to a shift of power balances between the state and companies, placing data production outside of the control of the nation state. This is in so far an unresolved problem because traditional data protection laws

and ethical rules are bound by the rules of the nation state and aim at keeping the state in check – and not companies.

2) *Shift of Power Balances from Citizens/Customers to MNCs:* At the same time, the power balances between individual people and companies have shifted. While in the past, citizens often needed to be forced to provide data to government agencies, today, as customers, they seem to be much more willing to provide even very sensitive personal data to companies. However, it is unclear, if this is really willingness or unawareness of how much information they actually reveal. For example, most smartphone apps automatically collect geo data and mobile data, meaning everybody who has a smartphone and does not disable the tracking function will continuously be monitored. Similarly, many new technologies in public and private spaces unobtrusively collect data. Examples are surveillance cameras, wearables such as smartwatches and smart eyewear, or the technologies engrained in smart houses in smart cities. In addition, customers often do not have a choice, i.e., they have to reveal their data, if they want to use the technology, even if these data are not necessary for the technology to function. As many of these technologies have become de-facto infrastructures, customers do not have a choice but to "agree" to reveal their data; e.g., in many contexts, Skype, Facebook, or What's App have become necessary for business and personal communication – if one refuses to use these technologies, one cannot do business or is excluded from social networks. All in all, while digitalization seemingly promises democratization, it in fact has decreased citizens' control over their own data (Traue 2020).
3) *Velocity:* Data production is a lot faster than in the past – this property of new-type big data is called "*velocity*" (Lane 2001; Weichbold et al. 2020, in this volume).
4) *Digital Data:* While traditional-type big data may or may not be digital, new-type big data are always digital data. From the point of view of social sciences, what is really new about the Web 2.0 and new-type big data is that society itself is changing due to digitalisation and mediatisation, but at the same time the patterns of digitalisation and mediatisation are structured by society. This is not a methodological issue but these new patterns of digitalisation and mediatisation might actually influence the way data are produced. Meaning, there might be a *recursive relation between digitalisation of society and analysing digital data*, which might be partly linked to the velocity of data production. What is more important from a methodological point of view is that – because they are digital data – all new-type big data (and the digital traditional-type big data) share the *specific selectivities of digital data*, which in social science discourse have been discussed via the label "*digital divides*" (*digitale Spaltungen*). We will examine these specific selectivities and how they influence data

quality of big data in the next section. Note that, in addition, specific forms of new-type big data (such as web videos vs. log files) may have additional selectivities.

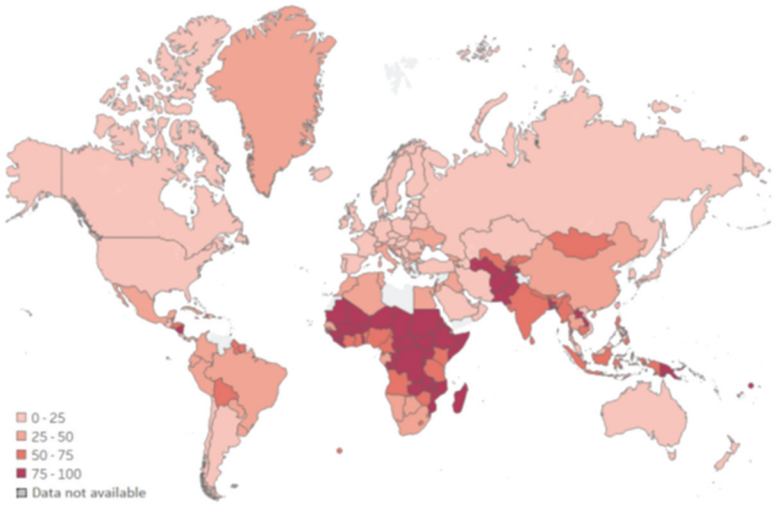
4. Digital Divides and Data Quality of New-Type Big Data

Some discourses in computational social sciences suggest that new-type big data are "better" than research-elicited data because they are supposed to cover the "full population" and thus make sampling obsolete. This assumption is simply false. What these debates forget is that even today, not everybody or everything is "on the internet" or using digital devices. There are many people who, and aspects of social life which, do not leave any digital traces at all and therefore are not reflected in new-type big data. Similar to coverage and nonresponse errors in survey research, this would be irrelevant, if it was random who does or does not use the internet. However, like with coverage and nonresponse errors, the contrary is true: Internet usage is socially structured, and some people are systematically excluded from internet usage. This social structure of internet (non)usage is called "digital divides" (*digitale Spaltungen*) in sociological discourse, and it implies that – like other process-generated and research-elicited data – new-type big data do not give a complete, "true," "objective," or "representative" image of society or the world but are actually systematically distorted. In order be able to assess how new-type big data are biased, we discuss in this chapter (1) the extent and patterns of digital divides in order to assess who is (not) using the internet in order to assess the extent of possible errors when using new-type big data as well as (2) possible reasons for exclusion from the internet and (3) different ways of using the internet in order to assess how and why data are biased.

4.1 Extent and Patterns of Digital Divides: Who is (Not) Using the Internet?

Internet usage has increased fast in the last two decades: While only about 17% of the world population used the internet in 2005, in 2019, 54% of the world population were "online" (ITU 2019a, 1). However, this means that even today, in contrast to common sense, not everybody is on the internet. On the contrary: *Almost half of the world population is still offline, i.e., they do not leave any type of digital traces*.

Graph 1: Digital Divides on a Global Scale: Global Rates of Offliners (2019): 46%



Source: ITU 2019a, 2.

Moreover, the *rate of non-usage is not evenly distributed around the globe* (Graph 1): In the Anglo-Saxon states (U.S., Canada, Australia, New Zealand), Western and Northern Europe, and in Russia, somewhere between 75 and 100% of the overall population are online. The rates are slightly lower in Latin America. In Asia, the Pacific, and the Arab countries, only about half of the population uses the internet. In contrast, in Africa, more than 80% of the population are offline (ITU 2019a, 2; 2019b). This global pattern seems to be at least partly connected to economic development and postcolonial structures, as in 2019, 87% of the population of developed countries but only 50% of the population of developing countries had internet access (ITU 2019b).

However, even within countries, there are huge *regional disparities* in internet access as well as *urban-rural divides*. A good example is Germany, where as a rule of thumb, the urban population is well connected while residents in less densely populated rural areas are not so well connected (BMVI 2017). This is not a small matter, as about one third of the German population lives in communities with less than 20.000 inhabitants.

Moreover, digital divides are socially structured along the dimensions of social inequality we already know from other social phenomena. In general, one can say that younger people, better educated people, members of higher social classes, and – especially in countries with strong gender inequalities – men are much more likely to use the internet than other parts of the population (for Germany, see, e.g., Initiative D21 1919; 2020a; 2020b).

Rates of digitalisation become even lower, if one takes into account, that no one is using the whole internet. Instead, people might use different devices (computers, smartphones, etc.) and different parts of the internet, ranging, e.g., from email, simple searches, using social media, or complex usage patterns. For example, for Germany, different types of internet users can be identified. One user type is officially "online" because they typically own either a smartphone or a computer. However, their internet usage typically does not exceed making a phone call or conducting a simple search via a search engine. They typically do not use social media, online payment, or streaming. These minimal-users together with the offliners comprised about one third of the German population in 2019, which on a world scale is a highly digitalized country. In addition, while usage rates have been largely increasing in the last decade, the proportion of this group has been more or less stagnating for a couple of years (Initiative D21 1919; 2020b).

So all in all, internet usage is not only socially structured but mostly follows similar patterns we know from other social phenomena, meaning that those groups of the population who are typically disadvantaged in other spheres of social life are also typically excluded from the internet: The person most unlikely to be on the internet is an elderly uneducated lower-class woman in rural Africa. This is methodologically important because this means, that currently, new-type digital data can be used for analysing patterns of using the internet, but they are only partly useful for analysing the offline, non-digital world. Digital data are especially not suitable for any types of research question where social inequality may play a role because they do not cover the most disadvantaged groups of society. This also means that they are at not or only partly suitable for many sociological research questions: The whole point of sociology is that different social groups also differ in typical behaviour, i.e., one cannot simply assume that – if members from one social group behave in a typical way – members of other social groups will behave in a similar way. If the social sciences use digital data without reflecting this bias in the data, they will replicate global power structures and miss especially the lifeworlds of the Global South and most disadvantaged parts of the population. This also points to the need for learning more about how digital data are biased.

4.2 Reasons for Exclusion from the Internet

In order to properly assess how digital data are biased, it is not only important to know *who* is excluded from the internet but also *why* they are excluded. Similar to reasons for survey nonresponse (Engel et al. 2004, 1-7, 24-32, 87-96; Baur 2006; Baur and Florian 2008), the reasons for exclusion might be linked with other typical behaviours of the same person. For example, if the reason for being offline is poverty, the respective person might also not be able to afford many other, even more essential things in life – this would be especially im-

portant to know in a study on consumption. In contrast, if a person is offline because of data privacy issues, this might be especially important in studies on internet data privacy.

In addition to complete non-usage, it is important to know *how* the persons who are online use the internet because, as stated above, not everybody uses the internet in the same way – and thus does not leave the same type of digital traces. This in turn will mean that specific digital platforms will also only grasp very specific aspects of social reality.

So the first question to ask is who has *access to internet* or rather: *fast* internet, because a lot of services will only work if you have 3G or higher mobile standards. During the time from 2007 to 2019, the number of people with access to internet has increased from about 5.5 billion people to almost 7 billion people, which is currently a large part of the world population, who in theory would be able to access the internet. In addition, the number of people having access to at least 3G is almost as high (ICT 2019a, 8). In other words, while there is still some regional variation following the patterns described above (ICT 2019a, 9), today, almost the whole world population could access fast internet, so this cannot be the main reason for exclusion from the internet anymore.

Even if one in theory has access to the internet, one still needs *digital devices* like smartphones, notebooks, computers, and so on in order to get access to the internet. On a worldwide scale, little information is at hand because only worldwide averages are available which say little about the regional and social distribution of devices. However, what can be said on a worldwide scale is that smartphones and other cellular phones today are the main way of accessing the internet (ICT 2019b). When looking at specific countries, there are more detailed data available which help to differentiate the image. In Germany, a country with relative high accessibility on a worldwide scale, one can see that in 2019 – although the numbers have been decreasing in recent years – only four in five people had smartphones and two in three had a notebook or laptop. The proportion of desktop users has declined to about half of the population, the proportion of users of simple mobile phones has decreased to about 28% of the population, and the number of tablet users is stagnating at about a third of the population. Only 28% of the German population use Smart TV, and only 7% use wearables (Initiative D21 2020b, 20). This means that the type of appliances accessible to people varies widely. While some people might have several appliances, others have only a specific point of access (Initiative D21 2020a; 2020b). Because some digital platforms need a specific device for functioning, this in turn means that due to lack of this specific device, people will not have access to these specific services.

A common way of getting access to the internet is during work. However, not everybody is employed, and only some people will get access to these devices while they work. For example, in Germany in 2015, only about two in

five employers gave their employees access to collaboration tools, like working together with others in documents, telephone conferences, home offices, or VPN for homeworking or mobile working. Only one in three employers gave their employees access to a smartphone or the possibility of videoconferences, one in five employers allowed smartphones to be integrated into the company infrastructure, only one in six employees worked with tablet PCs and one in three employees did not have any of these usages of devices or systems at work (Initiative D21 2016, 44-49). Moreover, there is a strong gender division of labour, because only one in five men but two in five women did not have any of these ways of access to digital devices at work (Initiative D21 2016, 44-49; 2020a). A reason for this gender-division of internet usage at work could be that occupations themselves are gendered. One can only assume that occupations that are stereotyped as "male" occupations make use of the internet more strongly than "female" occupations. This is a problem for general internet usage because many people gain skills in using the internet at work. If people do not use the internet at work, they will not get this experience and might use the internet less in general (Kirchner 2015). In addition, the internet usage patterns differ depending on, if one uses the internet at work or at home.

The third barrier to accessibility to the internet are *literacy* and *language* – if one cannot read or understand the content, one cannot use it. It is therefore important to keep in mind that on a worldwide scale, 14% of the population are illiterate. Again, strong gender, age, and regional effects can be observed: 10% of men and 17 % of women as well as 32% of persons aged 65 years and older are illiterate. In many parts of mid-Africa, literacy rates are well below 50%, and in India, they are also lower than 70% (UIS 2017).

Even if they can read, many people cannot access internet content due to language barriers. As Table 1 reveals, while only 15% of the world population can speak English, 59% of the internet content is in English. The language barriers become most explicit in comparison to Mandarin Chinese which is also spoken by 15% of the world population but represents only 1% of the internet content. Other languages overrepresented are Russian, German, Japanese, Turkish, Persian, Italian, and Polish, while especially languages spoken in China and India as well as Arabic are underrepresented.

Table 1: Languages Spoken vs. Languages of Internet Content on a World Scale

Language	Speakers (% of World Population)	Internet Content
English	15%	59%
Mandarin Chinese	15%	1%
Hindi	8%	
Spanish	7%	4%
French	4%	3%
Arabic	4%	
Bengali	3%	
Russian	3%	8%
Portuguese	3%	2%
Indonesian	3%	
Urdu	2%	
German	2%	3%
Japanese	2%	3%
Swahili	1%	
Marathi	1%	
Telugi	1%	
Western Punjabi	1%	
Wu Chinese	1%	
Tamil	1%	
Turkish	1%	3%
Persian		3%
Vietnamese	1%	1%
Italian		1%
Polish		1%

Note: The languages spoken do not only include native speakers but total usage in 2019, i.e., also those persons who have learned the language as a second language. A person can speak several languages. The percentages are calculated based on Ethnologue (2020) divided by an estimated world population of 7.7 billion persons. Source for languages of the internet in 2020: W3Techs 2020. Empty cells = percentage less than 1%.

Another barrier to use the internet is the *lack of competence and the knowledge to use the internet.* On a worldwide scale, a considerable lack of knowledge in even basic ITC skills can be observed:

> In 40 out of 84 countries for which data are available, less than half the population possesses basic computer skills such as copying a file or sending an e-mail with an attachment. For more complex activities (classified as 'standard skills'), such as using basic arithmetic formulae in a spreadsheet or downloading and installing new software, the proportions are even lower. In 60 of the countries for which data are available, these proportions are below 50 per cent. With respect to advanced computer skills, in only two countries (United Arab Emirates and Brunei Darussalam) do more than 15 per cent of people report having written a computer programme using a specialized programming

> language in the last three months. In only 10 other countries is that proportion above 10 per cent. Although more data need to be collected, these results show that there is a strong need to develop digital skills. (ICT 2019a, 10)

Even in a highly developed country such as Germany, these skills are often lacking; e.g., in Germany in 2016, while most people were well aware what "apps" meant, only half of the population had a concept of what a "cloud" or "cookie" might be. Other concepts like "industry for zero shared economy," "internet of things," "wearables," and "big data" were known by less than 20% of population and concepts like "smart meters" and "eHealth" by even less than 10% of the population. Other concepts not listed here were also not so much known by the general population. All in all, if one asks specifically what kind of knowledge people had, one realises that basic ICT skills do not suffice but instead, you need to know a lot of things when using the internet (competently). Knowledge varies highly concerning not only basic concepts but also concerning skills such as: How to transfer data between devices, how to get information in the internet? How to search for information using several data sources or even not only looking at the hits on the first page? How to transfer money online? How to use cloud appliances? How to set content into the settings (e.g., in social networks how to use netiquette)? How to write texts, do calculations, presentations, and use web appliances or even how to do programming on the computer? How to do installations both of devices and the network and if one is competent to do others? All these knowledge types are important to be a competent user of the internet and, again, this knowledge is socially structured along the lines of gender, age, and education (Initiative D21 2016).

Even if people are knowledgeable about using the internet, there might be other reasons why they cannot use the internet. One issue would be that different *governments* use the internet for surveillance or actually for controlling content and knowledge offered by the internet. For example, according to the Web-Index 2014 (WWWF 2016), between 2013 and 2014, government requests for using data has increased by 78% on Twitter, 76% in Yahoo, 30% in Facebook, and 40% on Google. Again, some people might take legal safe courts against surveillance, and these legal cases have in fact increased in recent years. However, this is only an option in democratic countries. In many countries, governments use different measures to restrict access to the internet. According to the latest "Freedom on the Net Report," 71% of internet users live in countries where they might face arrest or imprisonment for posting content on religious or political content (Freedom House 2019, 1ff.).

Depending on the country, not only is it possible that the government might deny people access to the internet, but people may choose not to use the internet out of fear that they could be surveyed by their government. This is especially an issue in countries with a strong history of government surveillance and dictatorships. A good example is Germany, which had, both during Na-

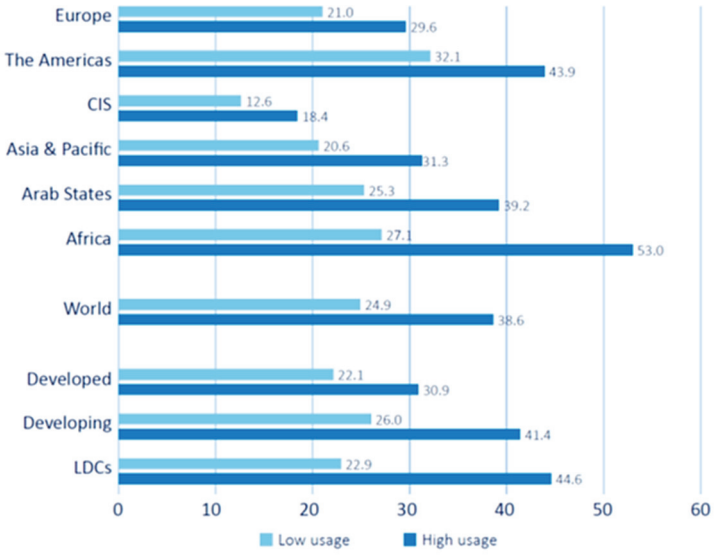
tional Socialism and in East Germany during the Socialist era, a strong history of surveillance. This is reflected in current patterns of internet usage: Many Germans are concerned about security when using the internet and thus try not to put any personal data on the internet. They regularly change passwords and update anti-viral software all while being aware that appliances and apps will collect and reveal data (to companies or governments). Some people are also concerned about where the data servers are hosted (Initiative D21 2016). Now this might result in different types of reactions: If people are competent, they might take special measures in order to secure their data. For example, in most social media, one can disable many standard activities and thus stay mostly anonymous, if one knows how to do this. However, an alternative way of handling this issue is by just simply not using social media. This becomes an issue, if a digital service becomes a standard, thus forcing a person to actually use this, regardless of whether they want to or not.

Even if people have no concerns about data privacy, possess the knowledge to use the internet, and have access to it, they still need to *want to use the internet*. In general, people will only use technologies (including the internet) if they find using them helpful in their everyday life – if they do *not* find them helpful, why should they spend time and money on the technology? Data available for Germany reveal why people might or might not want to use the internet: In 2016, only a third of the German population believed that the internet helps people to stay in contact with other people and also helps being more flexible in keeping up a good work-life balance. In contrast, 26% of the population believed that if there would not be internet tomorrow it would not have any negative consequences on their lives – for lower-income groups, the rate was even 40%. Only about a third of the German population sometimes use the internet longer than they originally wanted to, while one in four persons actually plan to stay consciously offline for longer in the future. Only one in five persons is interested in the new trends in the digital world. In addition to lack of positive incentives for using the internet, there might also be negative side-effects of using the internet, such as becoming a victim of cyber-bullying or other social hardships. However, in 2016, this was only a problem for 8% of the German population. More importantly, there might be a lot of people, who do not see any additional benefit for their daily life in using the internet, and this again depends on the user type (Initiative D21 2016). All these might be reasons for staying offline. This might all be changing now in the course of the Corona crisis where the internet might be the only way of staying in contact with persons one cannot meet face-to-face.

However, a further issue for not using the internet might actually be the *costs* of using the internet. In this context, concerning mobile broadband costs, when they are adjusted by GNI per capita, one finds that it is not the richest countries, but actually the poorest, that have the highest internet costs, with internet being most expensive in Africa (WWWF 2016, ICT 2019a, 11, see

also Graph 2); this is why decreasing the cost of internet access has become a World Development Goal (ICT 2019a, 11).

Graph 2: Costs of (Fast) Internet (2019): Bundled Mobile Broadband Prices, PPP$, 2019



Source: ICT 2019a, 11. Note: Simple averages, based on the economies for which data on mobile-broadband prices were available. High usage refers to a bundle including 140 minutes of voice, 70 SMS, and 1.5 GB of data. Low usage refers to a bundle including 70 minutes of voice, 20 SMS, and 500 MB of data.

Looking at the global data, in many ways *postcolonial structures* are replicated in the internet, meaning that the poorer and less developed a country and the weaker the state infrastructure, the more difficult it is for its residents to get free access to information. This also means that residents of these countries hardly leave any digital traces, which is why big data are not a suitable source of information for these countries.

However, the overall pattern of digital use on a global scale is not that clear, and it is not only income and GNP alone that influence access to the internet. In contrast, governments do strongly influence internet usage by *internet policies*. For example, Kirchner and Wolf (2015) show that in Europe, between early 1990 and 2010, the digitalization paths were strongly structured by welfare regimes. The concept of welfare regimes argues that specific welfare states have specific institutional structures, which both are deeply grounded in history and at the same time linked to concepts of equality and freedom. This also results in how, and what kind of, welfare provisions are provided as well as to

what extent. For Western Europe, one can distinguish (a) a Liberal or Anglo-Saxon, (b) a Social-Democratic or a Scandinavian, (c) a Conservative or Continental, and (d) a Southern-European Model. Kirchner and Wolf (2015) show that the Scandinavian countries were actually the most successful in promoting digitalization in the general population, and the Southern-European countries were the least successful. In the Scandinavian countries in 2010, 80% of the population were actually at least using computers and were on the internet, while in Southern-European countries just a little over half of the population were. Also interestingly, while the liberal countries had a strong head start, internet usage has stagnated since the mid-1990s. A similar pattern can be found when one distinguishes between basic uses and intensive uses. For basic uses, one can see again the Scandinavian countries have the highest percentage of basic uses, which is about 40% of the population, while with the other countries it is less than 25% of the population. In contrast to the liberal countries, Scandinavian and Continental European countries have very high rates of intensive uses, while here the Mediterranean countries lag behind. Kirchner and Wolf (2015) use more advanced analyses to conclude that government policies might actually shape the way of digitalization and in fact therefore influence what kind of data are used.

This does not only hold true for welfare states, but also for *colonial structures*. For example, if one looks on a worldwide scale (WWWF 2016) on the uses per population and compares them with the internet penetration per capita, one can see that the Western countries of the Global North have the highest internet usage of the world, while the countries of the Global South have very low rates of usage of the internet.

4.3 Digital Divides within the Internet

In addition to complete non-usage, it is important to know how the persons who are online use the internet because not everybody uses the internet in the same way – and thus does not leave the same type of digital traces. This in turn will mean that specific digital platforms will also only grasp very specific aspects of social reality.

First, *not everybody uses all kinds of technology*. While some people might use smartphones and mobile apps, others might use wearables, and others might still use their desktop computers and so on. There is also a difference between people using multiple devices or a single one. In addition, which is rarely reflected in methodological discourse, there might be issues like having several smartphones, sharing a smartphone, and so on.

Secondly, there might be *different reasons for usage*, the most important ones being people using the internet for work or during their leisure time. During work, usage patterns strongly differ between occupations and strata of the population (Kirchner 2015).

If people do not use the internet for work, but during their leisure time, there are also many *ways for using the internet*. For example, one might use the internet for gathering information, which in Germany in 2016, 17% of the population did, or simply for fun. Only one in five persons would use the internet for learning, e.g., languages with the help of online courses. Another way of using the internet is for communication, but again the usage is widely overestimated. In 2016 in Germany, about three in five people used the internet for instant messaging like WhatsApp, but this is especially true for the population below 60; in the population 60+, it was only one in three persons. Only two in five persons would actually use social networks such as Facebook, Xing, and Google+. One in five persons would use cloud services, like Dropbox and Google Drive, or forums, blogs, and websites, meaning both asking and answering questions there. Only about 13% of the population use devices like Google Docs and Microsoft SharePoint, in order to share information with other people (Initiative D21 2016).

Another way of usage is for collecting data in the sense of *quantification* (Mau 2017) in order to better control one's life (for example, counting steps and calories to collect data for oneself, which only one in ten persons would do, or using smartphone appliances like intelligent heat control at home, which only one in twenty persons would do, or even using eHealth to exchange health and fitness data, which again only one in twenty persons would do). Therefore, the typical data associated with wearables in the smart home are not only currently a tiny part of the general population, but they are also especially used by the higher income strata, and typically by men, hinting this might not only be an issue of money and knowledge but also of wanting to use them, because these imply a degree of self-control of one's everyday life by the internet, which a lot of people might not want.

Another possibility of using the internet is digital commerce, which is also strongly overestimated: In 2016, only about half of the German population used online shopping; only two in five persons used online services like booking holidays, car sharing, and so on. Only one in four persons used demand streaming services like Spotify, Netflix, and Amazon, and only one in five persons sold anything via the internet (Initiative D21 2016).

5. Aim and Content of this Forum

To sum up, big data need not be digital data, and the boundary between traditional-type and new-type big data is very fluent. While sometimes in methodological discourses, traditional-type and new-type big data seem to be completely different types of data, in research practice, their boundaries are fluent: Not only research data centres, but also historical data archives and a large number of university projects are working on making both types of data (i.e., previous-

ly untapped old data collections as well as new digital data) increasingly accessible for research. This is expanding the possibilities for the social sciences to analyse social phenomena from both a current and historical perspective. These possibilities increase even further, when data sets are linked or fused (Cielebak and Rässler 2019), and big data are combined with each other and/or with research-induced data such as survey data, qualitative interviews, or ethnography, resulting in mixed-mode or mixed-methods data analysis.

At the same time, the power of new-type big data should not be overestimated, as only about half the world population is using the internet and thus leaving digital traces, and exclusion from the internet is strongly structured along lines of social and spatial inequality. The inequalities on a global scale show that there are many reasons for keeping people from using the internet, and so today, we are far away from saying that the internet grasps the complete social area of reality. Like any process-generated data, digital data are distorted and can by no means substitute per se for other data. This does not mean that digital data cannot be used for social science research, but instead means that digital data are just like other data: They might be useful for some questions but not useful for others. In order to find out which types of people and situations one can actually analyse by using digital data, one should rather ask "Who is excluded from the internet and why?" and "What do we already know about digital divides?"

However, currently, there is a strange and unfruitful contradiction between research practice and the possibilities of handling big data and methodological discourse on big data, which is a barrier to asking these questions. This contradiction is increased by a division within the methodological community:

1) *Social science methodology* focusses its discussion on research-elicited and traditional-type big data and stresses that these data are socially constructed and each have their own methodical problems. As we have shown in this paper, this debate is not new but – at least for German-language sociology – is almost 150 years old and was a main reason for founding sociology as a discipline. This long-lasting debate has revealed the specific strengths and weaknesses of both big data and research-elicited data and also resulted in recommendations on how to handle each data type.
2) *Computational social sciences* primarily focus on new-type big data. In methodological research, computational social sciences are increasingly turning to new analysis techniques and algorithms for evaluating big data. Here, too, a debate on methods is emerging, which primarily addresses pragmatic feasibility problems as well as structuring through technology.

So while one line of discourse mostly focusses on data quality, the other one mostly focusses on data analysis. Both these research lines are hardly connected, and both have mutual blind spots. This HSR Forum therefore aims at dis-

cussing how these research lines can complement each other and can be improved by using the findings of each other. For example, the potentials of new analysis procedures and algorithms of computational science are promising also for analysing traditional-type big data. On the other hand, critically reflecting on possible errors and distortions of new-type big data is necessary to consider because data quality will strongly influence results. A critical examination of the erroneousness and internal distortion of new-type big data seems more than necessary.

This HSR Forum aims at contributing to such an exchange and at reflecting on the conditions under which big data are created as well as discussing the challenges of using them. This includes the question of whether and how the concept of "data science" needs to be expanded or updated. Thus, in addition to measurement-related characteristics, social, political, and economic conditions come into consideration, which make the interpretation of analysis results meaningful. The Forum also wants to pose these methodological and theoretical questions with the intention of pointing out possibilities for increasing the significance of big data analyses in social science studies.

Peter Graeff and Nina Baur start the discussion by applying the concept of "data lore" to three data types about corruption: research-elicited data (such as survey data), traditional-type big data, and new-type big data. When these different data sources about corruption are compared with each other according to their quality, the immediate question arises as to which is the best one for measuring corruption and answering questions about this topic. Since data quality is necessarily connected to the amount and types of errors within the data, classical ideas about public administrative and survey data are applicable for answering this question. Graeff and Baur pick up the so-called "Bick-Mueller-Model" (Bick and Müller 1984) that was developed in the 1980s in order to regard the special features and particularities of administrative mass data. They show how and why errors can occur within the administrative process of registering corruption data and juxtapose those in opposition to errors which belong to research-elicited data. They also describe new trends in data generation and application and show the progress made since Bick and Müller (1984) introduced their model and discuss new features of digitalism and new technologies with particular reference to the new-type big data. One of their conclusions is that the question about the "best" data cannot be answered according to their quality features. They exemplify this by referring to corruption data. No data source or data type measures corruption directly and all data are flawed in some way (such as a lack in population coverage or generalizability). However, data from different sources could be – sometimes – combined in order to answer specific research questions. From this point of view, new-type big data have the advantage of complementing other equally flawed data sources – they reveal their strength in triangulation with these other data

sources. However, fruitful triangulation is only possible if researchers are aware of the specific advantages and disadvantages of using specific data types.

In the second paper, *Gertraud Koch and Katharina Kinder-Kurlanda* argue that "source criticism" (which resembles the concept of "data lore" introduced by Graeff and Baur) is as an epistemological practice in social and cultural studies, which is crucial for specifying the range and scope of the findings, or in other words, their validity and reliability. In the context of big data, source criticism is not yet established in the fashion as it is known in other areas of social and cultural research. Currently emerging discussions in historical research emphasize the relevance of source criticism of digital objects respective data. In the context of these discussions, Koch and Kinder-Kurlanda suggest exploring the potentials of source criticism for platform logics. They focus on big data sourced from the internet. Nevertheless, their results aim at being transferrable to other sources of big data. The inclusion of source criticism into big data analysis may in turn foster the integration of data-driven analyses into social and cultural studies research approaches. For an integration of source criticism, Koch and Kinder-Kurlanda propose source critical analyses of information systems respective internet platforms in big data analysis with regard to (a) types of big data platforms, (b) researchers as data producers, and (c) mixed realities of platform usage practices. In analogy to source repertoires (*Quellentypen*), Koch and Kinder-Kurlanda suggest classifying internet platforms as providers of particular types of big data sources depending on their infrastructural materiality and ontologies for tracing the key issues of (external) source criticism: provenance, authenticity, and integrity.

In the third contribution, *Martin Weichbold, Alexander Seymer, Wolfgang Aschauer, and Thomas Herdin* highlight the potentials and limits of big-data analyses of media sources compared to conventional, quantitative content analysis: In an multidisciplinary project in Austria (based on the KIRAS security research program), the software tool WebLyzard was used for an automated analysis of online news and social media sources (comments on articles, Facebook postings, and Twitter statements) in order to analyse the media representation of pressing societal issues and citizens' perceptions of security. Frequency and sentiment analyses were carried out by two independent observers in parallel to the automated WebLyzard results. Specific articles on selected key topics like technology or Muslims in two major online newspapers in Austria (*Der Standard* and *Kronen Zeitung*) were counted, as were user comments, and both were evaluated according to different sentiment categories. The results indicate various weaknesses of the software leading to misinterpretations, and the automated analyses yield substantially different results compared to the sentiment analysis carried out by the two raters, especially for cynical or irrelevant statements. From a social science methodology perspective, the results clearly show that methodology needs to promote theory-based research, should counteract the attraction of superficial analyses of complex social issues, and

should emphasize not only the potentials but also the dangers and risks associated with big data.

In their contribution, *Rainer Diaz-Bone, Kenneth Horvarth, and Valeska Cappel* argue that the phenomenon of big data does not only deeply affect current societies but also poses crucial challenges to social research. The authors argue for moving towards a sociology of social research in order to characterize the new qualities of big data and their deficiencies. They draw on the neopragmatist approach of economics of convention (EC) as a conceptual basis for such a sociological perspective. This framework suggests investigating processes of quantification in their interplay with orders of justifications and logics of evaluation. Accordingly, methodological issues such as the question of the "quality of big data" must be discussed in their deep entanglement with epistemic values, institutional forms, and historical contexts and as necessarily implying political issues such as who controls and has access to data infrastructures. Using this conceptual basis, the paper uses the example of health to discuss the challenges of big data analysis for social research. Phenomena such as the rise of new and massive privately owned data infrastructures, the economic valuation of huge amounts of connected data, or the movement of "quantified self" are presented as indications of a profound transformation compared to established forms of doing social research. Methodological, epistemological, but also institutional and political strategies are presented to face the risk of being "outperformed" and "replaced" by big data analysis as they are already done in big US American and Chinese Internet enterprises. In conclusion, the authors argue that the sketched developments have important implications both for research practices and methods teaching in the era of big data.

Michael Weinhardt concludes the focus with his contribution on the "Ethical Issues in the Use of Big Data for Social Research." He argues that with the advent of big data, vast amounts of data have become available. This happens faster than the development of according ethical and legal standards for the analysis of this type of data. Weinhardt highlights that researchers face moral dilemmas concerning privacy and autonomy and asks which ethical and legal aspects need to be considered when collecting and analysing data from the web. In order to answer these questions, the paper provides an overview over existing research ethic regulations like the right to be protected from harmful conduct and specific laws like the EU-GDPR, which came into effect in May 2018. Weinhardt then points out different ethical problems that arise when collecting big data. For example, in order to link different data sources, individuals have to be identifiable. Weinhardt concludes with a set of recommendations (e.g., using a risk benefit analysis or the development of ethical guidelines) that are aiming to stimulate further scientific discussion.

All in all, from a methodological point of view, all the critique of traditional-type big data applies to digital data and all the knowledge accumulated for analysing these data applies for digital data as well. This in turn means that the

social sciences have already explicit knowledge on how to handle these data methodologically. Still new-type big data are not the same as traditional-type big data, because they have some extra properties, but these are not the ones that are typically discussed in the big data discourse. An important new aspect of digital data is velocity, meaning the high speed with which data are generated, transferred, and linked. In addition, what is relatively new, but which could also be applied to classical process-generated data, are new analysis techniques like text mining. Another important aspect to be discussed is the shift in the balance of power (*Machtbalance*) between the state, companies, and citizen/customers.

So what are the questions we should ask for future research? We should ask first, how do the shifts in balances of power, both from transferring data ownership from the state to companies as well as the increasing willingness of people to provide data, inflect in data quality? Second, if one agrees that there are different types of data within the internet, how do different types of data in the internet differ in data quality and how do these data differ from classical process-produced data such as public administrative data, media data, and company data and how do they differ from research elicited data? How can one use the classical tools of social science methodology for assessing data quality in big data and how do they have to be modified? If one assumes, that all data have some types of distortions, which types of data are suitable for which types of questions? Finally, how can these data be analysed? This means, firstly, how can we use techniques like text mining for analysing these data, but also, secondly, how can we use other classical methods of social science data analysis for analysing digital data? Things that come to mind are new methods like "webnography" or the "sociology of knowledge approach to discourse."

Special References

Contributions within this HSR Forum
"Challenges for Big Data Analysis. Data Quality and
Data Analysis of Analogous and Digital Mass Data"

Diaz-Bone, Rainer, Kenneth Horvath, and Valeska Cappel. 2020. Social Research in Times of Big Data. The Challenges of New Data Worlds and the Need for a Sociology of Social Research. *Historical Social Research* 45 (3): 314-341. doi: 10.12759/hsr.45.2020.3.314-341.

Graeff, Peter, and Nina Baur. 2020. Digital Data, Administrative Data, and Survey Compared: Updating the Classical Toolbox for Assessing Data Quality of Big Data, Exemplified by the Generation of Corruption Data. *Historical Social Research* 45 (3): 244-269. doi: 10.12759/hsr.45.2020.3.244-269.

Koch, Gertraud, and Katharina Kinder-Kurlanda. 2020. Source Criticism of Data Platform Logics on the Internet. *Historical Social Research* 45 (3): 270-287. doi: 10.12759/hsr.45.2020.3.270-287.

Weichbold, Martin, Alexander Seymer, Wolfgang Aschauer, and Thomas Herdin. 2020. Potential and Limits of Automated Classification of Big Data – A Case Study. *Historical Social Research* 45 (3): 288-313. doi: 10.12759/hsr.45.2020.3.288-313.

Weinhardt, Michael. 2020. Ethical Issues in the Use of Big Data for Social Research. *Historical Social Research* 45 (3): 342-368. doi: 10.12759/hsr.45.2020.3.342-368.

References

Amossé, Thomas. 2016. The Centre d'Etudes de l'Emploi (1970-2015). Statistics – On the Cusp of Social Sciences and the State. *Historical Social Research* 41 (2): 72-95. doi: 10.12759/hsr.41.2016.2.72-95.

Baur, Nina. 2004. Wo liegen die Grenzen quantitativer Längsschnittanalysen? *Bamberger Beiträge zur empirischen Sozialforschung* 23. Bamberg.

Baur, Nina. 2005. *Verlaufsmusteranalyse. Methodologische Konsequenzen der Zeitlichkeit sozialen Handelns.* Wiesbaden: VS-Verlag für Sozialwissenschaften. doi: 10.1007/978-3-322-90815-5.

Baur, Nina. 2008. Taking Perspectivity Seriously. A Suggestion of a Conceptual Framework for Linking Theory and Methods in Longitudinal and Comparative Research. *Historical Social Research* 33 (4): 191-213. doi: 10.12759/hsr.33.2008.4.191-213.

Baur, Nina. 2009. Measurement and Selection Bias in Longitudinal Data. A Framework for Re-Opening the Discussion on Data Quality and Generalizability of Social Bookkeeping Data. *Historical Social Research* 34 (3): 9-50. doi: 10.12759/hsr.34.2009.3.9-50.

Baur, Nina. 2011. Mixing Process-Generated Data in Market Sociology. *Quality & Quantity* 45 (6): 1233-1251. doi: 10.1007/s11135-009-9288-x.

Baur, Nina. 2014. Comparing Societies and Cultures. Challenges of Cross-Cultural Survey Research as an Approach to Spatial Analysis. *Historical Social Research* 39 (2): 257-291. doi: 10.12759/hsr.39.2014.2.257-291.

Baur, Nina, and Linda Hering. 2017. Die Kombination von ethnografischer Beobachtung und standardisierter Befragung. Mixed-Methods-Designs jenseits der Kombination von qualitativen Interviews mit quantitativen Surveys. *Kölner Zeitschrift für Soziologie und Sozialpsychologie* 69: 387-414. doi: 10.1007/s11577-017-0468-8.

Baur, Nina, and Michael Florian. 2008. Stichprobenprobleme bei Online-Umfragen. In: *Sozialforschung im Internet. Methodologie und Praxis der Online-Befragung,* eds. Jackob, Nikolaus, Schoen, Harald, and Thomas Zerback, 106-125. Wiesbaden: VS-Verlag. doi: 10.1007/978-3-531-91791-7_7.

Baur, Nina, and Siegfried Lamnek. 2016. Multivariate Analysis. Ritzer, George, ed. *The Blackwell Encyclopedia of Sociology*. Oxford: Blackwell Publishing Ltd. doi: 10.1002/9781405165518.wbeosm133.pub2. Blackwell Reference Online

Baur, Nina, Kelle, Udo and Kuckartz, Udo. 2017. Mixed Methods – Stand der Debatte und aktuelle Problemlagen. *Kölner Zeitschrift für Soziologie und Sozialpsychologie* 69: 1-37. DOI: 10.1007/s11577-017-0450-5.

Baur, Nina, Knoblauch, Hubert, Akremi, Leila, and Boris Traue. 2018. Qualitativ – quantitativ – interpretativ. Zum Verhältnis methodologischer Paradigmen in der empirischen Sozialforschung. In *Handbuch Interpretativ forschen*, eds. Akremi, Leila, Baur, Nina, Knoblauch, Hubert, and Boris Traue, 246-284. Weinheim: Beltz Juventa.

Baur, Nina. 2006. Ausfallgründe bei zufallsgenerierten Telefonstichproben am Beispiel des Gabler-Häder-Designs. In *Stichprobenqualität in Bevölkerungsstichproben*, eds. Faulbaum, Frank, and Christof Wolf, 159-184. Bonn: IZ.

BBSR (Bundesinstitut für Bau-, Stadt- und Raumforschung). 2020. Laufende Stadtbeobachtung – Raumabgrenzungen. Stadt- und Gemeindetypen in Deutschland. <https://www.bbsr.bund.de/BBSR/DE/Raumbeobachtung/Raumabgrenzungen/deutschland/gemeinden/StadtGemeindetyp/StadtGemeindetyp_node.html> (Accessed March 02, 2020).

Behrisch, Lars. 2016. Statistics and Politics in the 18th Century. *Historical Social Research* 41 (2): 238-257. doi: 10.12759/hsr.41.2016.2.238-257.

Bick, Wolfgang, and Paul J. Müller. 1984. Sozialwissenschaftliche Datenkunde für prozeßproduzierte Daten. Entstehungsbedingungen und Indikatorenqualität. In. *Sozialforschung und Verwaltungsdaten. (Historisch-Sozialwissenschaftliche Forschungen Volume 17)*, eds. Bick, Wolfgang, Mann, Reinhard, and Paul J. Müller, 123-159. Stuttgart: Klett-Cotta.

BMVI (Bundesministerium für Verkehr und digitale Infrastruktur). 2017. *Der Breitbandatlas*. <http://www.bmvi.de/DE/Themen/Digitales/Breitbandausbau/Breitbandatlas-Karte/start.html> (Accessed December 01, 2017).

Bottel, Matthias, and Heiko Kirschner. 2019. Digitale Spiele. In *Handbuch Methoden der empirischen Sozialforschung*, eds. Baur, Nina, and Jörg Blasius, 1085-1099. Wiesbaden: Springer Fachmedien.

Cielebak, Julia, and Susanne Rässler. 2019. Data Fusion, Record Linkage und Data Mining. In *Handbuch Methoden der empirischen Sozialforschung*, eds. Baur, Nina, and Jörg Blasius, 423-440. Wiesbaden: Springer Fachmedien.

Collins, Randall. 2001. Weber's Last Theory of Capitalism. In *The Sociology of Economic Life*, eds. Granovetter, Mark, and Richard Swedberg, 379-400. Cambridge (MA): Westview.

Desrosières, Alain. 2005. *Die Politik der großen Zahlen. Eine Geschichte der statistischen Denkweise*. Berlin: Springer.

Desrosières, Alain. 2011. The Economics of Convention and Statistics: The Paradox of Origins. *Historical Social Research* 36 (4): 64-81. doi: 10.12759/hsr.36.2011.4.64-81.

Diaz-Bone, Rainer. 2016. Convention Theory, Classification and Quantification. *Historical Social Research* 41 (2): 48-71. doi: 10.12759/hsr.41.2016.2.48-71.

Diaz-Bone, Rainer., and Didier, Emmanuel. 2016. The Sociology of Quantification – Perspectives on an Emerging Field in the Social Sciences. *Historical Social Research* 41 (2): 7-26. doi: 10.12759/hsr.41.2016.2.7-26.

Durkheim, Émile. 1897. *Le Suicide*. Paris: Félix Alcan.

Engel, Uwe, Pötschke, Manuela, Schnabel, Christiane, and Julia Simonson. 2004. *Nonresponse und Stichprobenqualität. Ausschöpfung in Umfragen der Markt- und Sozialforschung*. Frankfurt a. M.: Deutscher Fachverlag.

Engels, Friedrich. 1845. *The Condition of the Working Class in England*. Leipzig. <https://www.marxists.org/archive/marx/works/1845/condition-working-class/index.htm> (Accessed March 02, 2020).
Ethnologue. 2020. *What are the top 200 most spoken languages?* <https://www.ethnologue.com/guides/ethnologue200> (Accessed March 02, 2020)
Freedom House. 2019. *Freedom on the Net. The Crisis of Social Media*. Freedom House. <https://freedomhouse.org/sites/default/files/2019-11/11042019_Report_FH_FOTN_2019_final_Public_Download.pdf> (Accessed March 02, 2020)
Initiative D21. 2016. *D21-Digital-Index 2016. Jährliches Lagebild zur Digitalen Gesellschaft*. <http://initiatived21.de/app/uploads/2017/01/studie-d21-digital-index-2016.pdf> (Accessed March 02, 2020).
Initiative D21. 2019. *D21 Digital Index 2018/2019. Jährliches Lagebild zur Digitalen Gesellschaft*. <https://initiatived21.de/app/uploads/2019/01/d21_index2018_2019.pdf> (Accessed March 02, 2020).
Initiative D21. 2020a. *Digital Gender Gap. Lagebild zu Gender(un)gleichheiten in der digitalisierten Welt*. <https://initiatived21.de/app/uploads/2020/01/d21_digitalgendergap.pdf> (Accessed March 02, 2020).
Initiative D21. *2020b. D21 Digital Index 2019/2020. Jährliches Lagebild zur Digitalen Gesellschaft*. <https://initiatived21.de/app/uploads/2020/02/d21_index2019_2020.pdf> (Accessed March 02, 2020).
ITU (International Telecommunication Union). 2019a. *Measuring Digital Development. Facts and Figures 2019*. Geneva: ITU. <https://www.itu.int/en/ITU-D/Statistics/Documents/facts/FactsFigures2019.pdf> (Accessed March 02, 2020).
ITU (International Telecommunication Union) 2019b. *ITU Statistics*. Geneva: ITU. <https://www.itu.int/en/ITU-D/Statistics/Documents/statistics/2019/Stat_page_all_charts_2019.xls> (Accessed March 02, 2020).
Kandt, Jens. 2019. Geotracking. In *Handbuch Methoden der empirischen Sozialforschung*, eds. Baur, Nina, and Jörg Blasius, 1347-1354. Wiesbaden: Springer Fachmedien.
Kirchner, Stefan. 2015. Konturen der digitalen Arbeitswelt. Eine Untersuchung der Einflussfaktoren beruflicher Computer- und Internetnutzung und der Zusammenhänge zu Arbeitsqualität. *KZfSS* 67: 763-791. doi: 10.1007/s11577-015-0344-3.
Kirchner, Stefan., and Markus Wolf. 2015. Digitale Arbeitswelten im europäischen Vergleich. *wsi mitteilungen* 4/2015: 253-262.
Koch, Gertraud. 2019. Digitale Selbstvermessung. In *Handbuch Methoden der empirischen Sozialforschung*, eds. Baur, Nina, and Jörg Blasius, 1075-1084. Wiesbaden: Springer Fachmedien.
Korte, Hermann. 2004. *Soziologie im Nebenfach*. Konstanz: UVK.
Lakes, Tobia. 2019. Geodaten. In *Handbuch Methoden der empirischen Sozialforschung*, eds. Baur, Nina, and Jörg Blasius, 1339-1346. Wiesbaden: Springer Fachmedien.
Laney, Doug. 2001. 3D Data Management: Controlling Data Volume, Velocity and Variety. *META Group Research Note 6.*
Maier, Michael F., and Boris Ivanov. 2018. Selbstständige Erwerbstätigkeit in Deutschland. *Forschungsbericht 154 des Bundesministerums für Arbeit und Soziales*. Mannheim: ZEW. <https://www.bmas.de/SharedDocs/Downloads/DE/PDF-Publikationen/Forschungsberichte/fb514-selbststaendige-erwerbstaetigkeit-in-deutschland.pdf?__blob=publicationFile&v=1> (Accessed March 02, 2020).

Manderscheid, Katharina. 2019. Text Mining. In *Handbuch Methoden der empirischen Sozialforschung*, eds. Baur, Nina, and Jörg Blasius, 1099-1112. Wiesbaden: Springer Fachmedien.

Mau, Steffen. 2017. *Das metrische Wir: Über die Quantifizierung des Sozialen*. Frankfurt am Main: Suhrkamp.

Mayerl, Jochen, and Thorsten Faas. 2019. Quantitative Analyse von Twitter und anderer usergenerierter Kommunikation. In: *Handbuch Methoden der empirischen Sozialforschung*, eds. Baur, Nina, and Jörg Blasius, 1025-1038. Wiesbaden: Springer Fachmedien.

Mühlichen, Andreas. 2019. Informationelle Selbstbestimmung. In *Handbuch Methoden der empirischen Sozialforschung*, eds. Baur, Nina, and Jörg Blasius, 77-89. Wiesbaden: Springer Fachmedien.

Nam, Sang-hui. 2019. Qualitative Analyse von Chats und anderer usergenerierter Kommunikation. In *Handbuch Methoden der empirischen Sozialforschung*, eds. Baur, Nina, and Jörg Blasius, 1039-1050. Wiesbaden: Springer Fachmedien.

RatSWD [Rat für Sozial- und Wirtschaftsdaten] 2017a. Handreichung Datenschutz. *RatSWD Output 5 (5).* Berlin, Rat für Sozial- und Wirtschaftsdaten (RatSWD). https://doi.org/10.17620/02671.6 (Accessed March 02, 2020).

RatSWD [Rat für Sozial- und Wirtschaftsdaten] 2017b. Forschungsethische Grundsätze und Prüfverfahren in den Sozial- und Wirtschaftswissenschaften. *RatSWD Output 9 (5).* Berlin, Rat für Sozial- und Wirtschaftsdaten (RatSWD). doi: 10.17620/02671.1 (Accessed March 02, 2020).

RatSWD [Rat für Sozial- und Wirtschaftsdaten] 2018. Forschungsdatenmanagement in den Sozial-, Verhaltens- und Wirtschaftswissenschaften – Orientierungshilfen für die Beantragungund Begutachtung datengenerierender und datennutzender Forschungsprojekte. *RatSWD Output* 3 (5). Berlin, Rat für Sozial- und Wirtschaftsdaten (RatSWD). doi: 10.17620/02671.7 (Accessed March 02, 2020).

Riebling, Jan. 2019. *Methode und Methodologie quantitativer Textanalyse*. Bamberg: University of Bamberg Press.

Salais, Robert. 2016. Quantification and Objectivity: From Statistical Conventions to Social Conventions. *Historical Social Research* 41 (2): 118–134. doi: 10.12759/hsr.41.2016.2.118-134.

Scheuch, Erwin. K. 1977. Die wechselnde Datenbasis der Soziologie: Zur Interaktion zwischen Theorie und Empirie. In *Die Analyse prozeß-produzierter Daten*, eds. Müller, Paul J., 5-41. Stuttgart: Klett-Cotta. <https://nbn-resolving.org/urn:nbn:de:0168-ssoar-325047>

Schmidt, Jan.-Hinrik. 2019. Blogs. In *Handbuch Methoden der empirischen Sozialforschung*, eds. Baur, Nina, and Jörg Blasius, 1013-1024. Wiesbaden: Springer Fachmedien.

Schmitz, Andreas, and Olga Yanenko. 2019. Web Server Logs und Logfiles. In *Handbuch Methoden der empirischen Sozialforschung*, eds. Baur, Nina, and Jörg Blasius, 989-998. Wiesbaden: Springer Fachmedien.

Schrape, Jan-Felix, and Jasmin Siri. 2019. Facebook und andere soziale Medien. In *Handbuch Methoden der empirischen Sozialforschung*, eds. Baur, Nina, and Jörg Blasius, 1051-1062. Wiesbaden: Springer Fachmedien.

Schünzel, Anja, and Boris Traue. 2019. Websites. In *Handbuch Methoden der empirischen Sozialforschung*, eds. Baur, Nina, and Jörg Blasius, 999-1012. Wiesbaden: Springer Fachmedien.

Seysen, Christian. 2009. Effects of Changes in Data Collection Mode on Data Quality in Administrative Data. The Case of Participation in Programmes Offered by the German Employment Agency. *Historical Social Research* 34 (3): 191-203. doi: 10.12759/hsr.34.2009.3.191-203.

Speich Chassé, Daniel. 2016. The Roots of the Millennium Development Goals: A Framework for Studying the History of Global Statistics. *Historical Social Research* 41 (2). 218–237. doi: 10.12759/hsr.41.2016.2.218-237.

Thévenot, Laurent. 2011. Conventions for Measuring and Questioning Policies: The Case of 50 Years of Policy Evaluations Through a Statistical Survey. *Historical Social Research* 36 (4): 192–217. doi: 10.12759/hsr.36.2011.4.192-217.

Thévenot, Laurent. 2016. From Social Coding to Economics of Convention: a Thirty-Year Perspective on the Analysis of Qualification and Quantification Investments. *Historical Social Research* 41 (2): 96–117. doi: 10.12759/hsr.41.2016.2.96-117.

Thorvaldsen, Gunnar. 2009. Changes in data collection procedures for process-generated data and methodologicalimplications: the case of ethnicity variables in 19th century Norwegian censuses. *Historical Social Research* 34 (3): 168-190. doi: 10.12759/hsr.34.2009.3.168-190.

Traue, Boris, and Anja Schünzel. 2019. YouTube und andere Webvideos. In *Handbuch Methoden der empirischen Sozialforschung*, eds. Baur, Nina, and Jörg Blasius, 1063-1074. Wiesbaden: Springer Fachmedien.

Traue, Boris. 2020. *Selbstautorisierungen. Die Transformation des Wissens in der Kommunikationsgesellschaft*. Habilitationsschrift. TU Berlin

UIS (UNESCO Institute for Statistics). 2017. Literacy Rates Continue to Rise from One Generation to the Next. *Fact Sheet No. 45*. FS/2017/LIT/45. <http://uis.unesco.org/sites/default/files/documents/fs45-literacy-rates-continue-rise-generation-to-next-en-2017_0.pdf> (Accessed March 02, 2020).

Vargo, Chris J., Guo, Lei, and Michaelle A. Amazeen. 2017. The agenda-setting power of fake news: A big data analysis of the online media landscape from 2014 to 2016. *New Media & Society* 20(5): 2028-2049.

W3Techs. 2020. *Historical yearly trends in the usage statistics of content languages for websites*. <https://w3techs.com/technologies/history_overview/content_language/ms/y> (Accessed March 02, 2020).

Wallgren, Anders, and Brit Wallgren. 2014. *Register-based Statistics. Statistical Methods for Administrative Data. Second Edition*. Chichester: John Wiley & Sons.

Weber, Max. 1906-1922. *Die Wirtschaftsethik der Weltreligionen*. Reprinted in: Gesammelte Aufsätze zur Religionssoziologie. 3 Volumes. UTB: Stuttgart.

Whiteside, Noel 2015. Who Were the Unemployed? Conventions, Classifications and Social Security Law in Britain (1911-1934). *Historical Social Research* 40 (1): 150-169. doi: 10.12759/hsr.40.2015.1.150-169.

WWWF (World Wide Web Foundation). 2016. *Webindex. Report 2014 – 2015*. <http://thewebindex.org/wp-content/uploads/2014/12/Web_Index_24pp_November 2014.pdf> (Accessed December 01, 2017)

Historical Social Research
Historische Sozialforschung

All articles published in this Forum: