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

Graeff, Peter; Baur, Nina

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

Peter Graeff & Nina Baur [*]

Abstract: *»Digitale Daten, Verwaltungsdaten und standardisierte Befragungen im Vergleich. Update des klassischen Werkzeugkastens zur Ermittlung der Datenqualität von Massendaten, am Beispiel von Korruptionsdaten«*. In the digital age, new data types have become available that can, potentially, be used in social science research. Besides data that were originally created for scientific purposes (research-elicited data), administrative mass data (traditional-type big data) and data from digital devices (new-type big data) have become more and more relevant for research processes. Both data types can be subsumed under the term "big data." In this paper, we scrutinize the quality of administrative mass data on corruption in contrast to research-elicited data (e.g., survey data). Since data quality is crucial for the measurement of a social phenomenon such as corruption, we pose the question of how a social phenomenon can be measured by means of data from these different sources. As a first step, we refer to the so-called Bick-Mueller-Model. It was developed in the 1980s for observing the special features and particularities of administrative mass data (traditional-type big data). We contrast this model with the so-called Error-Approach that is typically applied in survey research. In order to account for new trends in data generation and application, we show the progress that has been made since Bick and Mueller introduced their model and discuss new features of digitalism and new technologies. We conclude that the features of the so-called Bick-Mueller-model are useful for tackling the particularities of administrative data and also – to some degree – new-type big data. The "error" perspective that is inherent both in the classical survey research and in the so-called Bick-Mueller model also applies to new-type big data when it comes to assessing their quality. Moreover, it is possible that the data from these different sources can complement each other. For this, researchers must be aware of the fact that neither data source actually measures corruption directly. For answering specific re-

[*] 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.
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.

Historical Social Research 45 (2020) 3, 244-269 │ published by GESIS
DOI: 10.12759/hsr.45.2020.3.244-269

search questions, it is crucial to consider the advantages and disadvantages of using specific data types.

Keywords: data quality, measurement, big data, mass data, process-generated data, process-produced data, digital data, survey data, digital methods, computational social sciences, survey methodology, total survey error, corruption.

1. Introduction

In the social sciences, “data” can be considered as bits of information that represent or are positioned in relation to a social phenomenon of interest. In order to describe or explain a social phenomenon, scientists usually apply research-elicited data which researchers collect themselves for research purposes. That is, they use data generated from sampling units (such as people or firms) in a research process that was specifically designed to explore and measure the social phenomenon of interest. Consider crime, e.g., corruption, as an example: If social scientists explore such a phenomenon, typically self-reporting surveys are conducted in order to gather information about corrupt practices (such as in Graeff et al. 2014; Dickel and Graeff 2018). The questionnaires or interviews are designed to investigate the prevalence or volatility of corruption and are applied under conditions which exacerbate issues in its measurement, such as socially desirable answering or sensitive topics (Dickel and Graeff 2016).

There are, however, other data that were not originally generated for research purposes but could also be applied for scientific research – the so-called “process-produced” or “process-generated data” (Baur 2009; Baur et al. 2020, in this HSR Forum). For example, corruption as a social phenomenon is not only of interest to social scientists but at the same time happens in every society and is thus typically registered by authorities such as the police, the prosecution departments, and the courts. Take, for example, the case of bribery when a firm’s employee conducts a corrupt transaction with a public official in order to get a public contract for his firm. When public procurement is involved, such a phenomenon frequently occurs in Western countries (Charron et al. 2017). If such a crime is discovered, the number of cases within a region (or a society) – gathered by the prosecution authorities – are related to the phenomenon of corruption but are not produced by a scientific process of data generation. Official figures of corruption-crimes can be of interest for social scientists if they are usable as (official) indicators for societal conditions (Skogan 1974).

The methodological question that arises from this problem is: Which of these data types is better? In some methodological discourses, the answer seems obvious. For example, survey methodologists tend to favor research-elicited data because researchers can at least estimate control over the measurement and sampling errors (Baur 2009). In contrast, computational social scientists tend to favor process-generated data because these are often “big data." The fact alone that different research fields disagree on what should be the preferred data type reveals that the answer which is the best data type is not clear at all (Baur et al. 2020, in this HSR Forum), even if in theory, one could choose any data one wanted. The situation becomes more complicated, if one does not reflect the ideal research world but looks at actual research practice because very often, not all data types are available for all research questions (Baur 2009). The topic of "corruption" is a good example: There simply are no single good data sources for measuring corruption.

Therefore, using the example of "corruption," in this paper, we pose the question of how a social phenomenon can be measured by means of data from different sources which are obviously, or at least indirectly, related to the phenomenon of interest. While this question has always been a key question of social science methodological discourses, it has become increasingly relevant in recent years because in the wake of the digital turn, new data sources have become available. New sources, such as media or regional (or spatial) data, augment the array of classical sources, such as statements (e.g., intentions of actions) by persons about their own or others' corrupt practices (survey data) or the number of cases registered by prosecuting authorities (administrative data or administrational data).

The paper is organized as follows: In section 2, we elaborate on the characteristics of administrative (mass) data (also: traditional-type big data) and survey data, using the example of corruption data in Germany. In the past, several attempts have been made to compare these types of data. In sections 3 and 4, we refer to the comparison made by the so-called Bick-Mueller-Model and contrast it with the so-called Error-Approach typically applied in survey research. We also highlight some progress made since Bick and Müller introduced their model in the 1980s and, in section 5, we ultimately turn to new features of digitalism and new technologies.

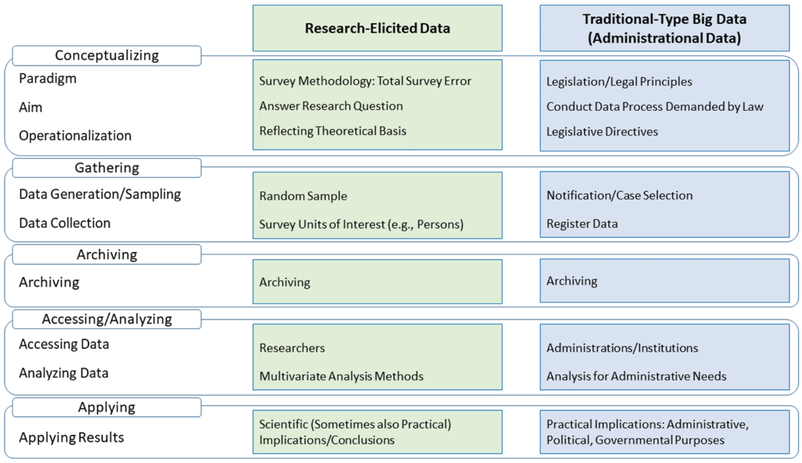
It is impossible to discuss the full array of social science data in one single paper (for an overview, see Baur and Blasius 2019). We therefore focus on quantitative data and those data types which are currently most often used in social science research on corruption. Specifically, we are going to use survey data as an example for research-elicited data and public administrative data as an example for process-produced data. Obviously, focusing on corrupt practices data in a country limits the generalizability of our ideas because different social phenomena show other specific data characteristics. This is, however, the call Bick and Müller (1984; and also Baur 2009) made when they suggested a framework for analyzing process-produced data such as public administrative and other big data: If process-produced data are to be used for social science research, it is necessary to regard their *specific characteristics*, which refers, in particular, to the circumstance that they are generated and applied within a legislative framework.

2. Aspects that Demarcate Survey Data from Traditional-Type Big Data

In Germany, the criticism of process-generated data, such as historical sources (*Quellen*), and traditional-type big data (also: social bookkeeping data, process-generated mass data), such as government data and other public administrational data (*Verwaltungsdaten*), dates back to the 19th century and lies at the heart of the founding of German sociology as a discipline (Baur et al. 2020, in this HSR Forum): In early German sociology, social science methodology was considered as one of the key elements of reflecting the advantages and drawbacks of different data sources (Baur et al. 2018). The criticism of process-produced data resulted in social scientists developing techniques for research-elicited data, and by the 1920s, survey data were an important social science data source (Baur 2014, 257). In the consequent decades, methodologists not only continuously improved methods for collecting research-elicited data but also discussed the advantages and disadvantages of various data types. This debate got a new impetus in the 1960s and 1970s, when the new methods of digital data storage and computation made many types of process-produced data more easily available (Baur et al. 2020, in this HSR Forum). In the 1980s, the debate about the differences and similarities between data generated by a scientific research process and data generated by administrative/institutional processes culminated in Bick and Müller (1980, 1984) suggesting a model for assessing the data quality of traditional-type big data (also: process-produced mass data; e.g., Baur 2009).

Referring in particular to this debate, we point out the most important features of each data type. As means of illustration, we link these ideas to the phenomenon of corruption. The ways in which survey and traditional-type big data (administrative data) are conceptualized, generated, collected, archived, analyzed, and applied can be described in similar steps (see Figure 1).

Figure 1: Characteristics of Data



	Research-Elicited Data	Traditional-Type Big Data (Administrational Data)
Conceptualizing		
Paradigm	Survey Methodology: Total Survey Error	Legislation/Legal Principles
Aim	Answer Research Question	Conduct Data Process Demanded by Law
Operationalization	Reflecting Theoretical Basis	Legislative Directives
Gathering		
Data Generation/Sampling	Random Sample	Notification/Case Selection
Data Collection	Survey Units of Interest (e.g., Persons)	Register Data
Archiving		
Archiving	Archiving	Archiving
Accessing/Analyzing		
Accessing Data	Researchers	Administrations/Institutions
Analyzing Data	Multivariate Analysis Methods	Analysis for Administrative Needs
Applying		
Applying Results	Scientific (Sometimes also Practical) Implications/Conclusions	Practical Implications: Administrative, Political, Governmental Purposes

Source: Modified from Baur 2009 and Bick and Müller 1984.

The two *types of data* are related to different *paradigms* in the literature:

1) *Research-elicited data*, such as survey data, are collected specifically for answering social science research questions – which is why we call the process of generating these data "scientific" in the context of this paper. As the scientific research process starts with a sound aim, research questions, and theoretically expected results, any difficulties that jeopardize reaching the research goals are considered (within survey methodology) to be errors. The conceptual framework of the "total survey error" (TSE) comprises all errors that could possibly occur during generating, archiving, or analyzing survey data (Biemer 2010; Groves and Lyberg 2010).[1]
2) The paradigm of *process-produced data* (such as public administrative data) refers to the legislation or rules given by the authority, as government institutions/administrations usually determine what kind of data are generated (Wallgren and Wallgren 2014, 8). Legislation provides authorities and statistical offices with the rights to notify, register, and use data of persons or firms/institutions. Unlike researchers, administrations (which generate data from persons most frequently) have the right to ascertain their identity and get information about personal characteristics, such as income and taxation.

The *overall aim* of the survey process is answering research questions. The survey design is tailored to getting the data necessary to answer the research

[1] There are a lot of different possible errors in regard to each step in the survey process (Biemer 2010). We will introduce the relevant error concepts in the next section.

question. Administrative data are generated because they are needed in the public sphere (such as information on the prevalence of diseases), for administrative/bureaucratic functions (such as conducting taxation or criminal prosecution), or for political/governmental reasons (such as making a decision on the tax rate or ways of prosecuting criminals) – they are gathered for the delivery of a service (Woollard 2014).

In most cases, these objectives are regulated by laws, so the aim behind the data generating process is to conduct it in accordance with both the objectives and the laws (Bick and Müller 1984). While the aim of research is oriented towards knowledge (or explaining), the administrative conduction is based on practical reasons (Baur 2004).

For *operationalization*, data generated in scientific research are aligned with theoretical propositions or explorative assumptions. Ideally, theoretical ideas drive any form of operationalization, suggesting which research designs and survey questions should be included in the survey instruments (Baur 2009). Data that are not research-elicited (e.g., process-produced data from public institutions or administrations) are operationalized according to the legislative directives or principles that are applied in administrative processes. The operationalization of non-research data is typically not done while accounting for forthcoming (statistical) analysis.

The process of data generation in scientific research is conducted by *sampling* the units of interest, such as persons, firms, or others entities. While there are several ways of case selection and sampling in a scientific way, if one would like to draw conclusions on the basis of inferential statistical grounds, a random sample needs to be selected from the target population. Under the conditions of random sampling, error probabilities can be produced. Large population surveys usually contain features of representativeness or the opportunity of correcting for it. In contrast, traditional-type big data are generated by notifications: The responsible authority in an administration/institution receives relevant information on a certain case (such as a tax payer or a criminal suspect). The *data collection* is conducted by registering this case together with all the information that is demanded by the law/administrative directives. For scientific purposes, randomly sampled persons or other sampling units of interest are surveyed to gather the information that is necessary to answer the research questions. In this way, metadata (such as information about time or interviewer) are usually also stored and eventually used. All data gathered in the research context refer to the population of interest, which is different to non-research-elicited data. Administrative metadata are only stored/registered if there is a directive for doing so. Not all data collected for administrative purposes refer to the population of interest, particularly not those that are part of the data exchange process between administrations/institutions.

Pointing out these differences between research-elicited data and traditional-type big data (e.g., administrative data) shows that the paradigms, ways of conducting, and intentions to use data differ between researchers and data-generating administrations/institutions. The idea of avoiding or minimizing errors in the data generating process does not apply to non-research-elicited data. Therefore, Bick and Müller (1984) suggested to substitute the paradigm of "error lore" applied to social science surveys with the paradigm of "data lore" which functions as a general framework to account for the characteristics of non-research-elicited types of data. Within this framework, typical "errors" could be pinpointed, such as mistakes in notification that the police make when witnesses report crimes.

After being gathered, data need to be archived, which usually requires that the data have to be prepared, to fit the demands of being stored. For research-elicited data, *archiving* necessitates removing unnecessary or false information from the data set (such as wrong entries), adding metadata, and matching the data with the chosen data format. In recent years, the directives for handling and storing data according to data protection laws have moved more into focus – both for research-elicited and process-generated data. As traditional-type big data from administrations/institutions most often contain person-specific sensitive information (such as financial information), data protection is paramount.

Access to process-generated data from administrations/institutions is usually restricted, as these data are not intended to be used for scientific analysis or to be used by persons outside the administration. While data derived from research processes are – if first anonymized – usually available for secondary analysis by other researchers (e.g., to reanalyze the data and approve/disapprove the results gained from them), the handing over of administrative data (to scientific researchers or third persons) crucially depends on the reasons for analysis. Data which are used to produce official figures on social conditions (such as crime rates in a region) have become available in the recent decades but access to data about individual cases requires a permit from the authorities. Moreover, data exchange between administrations also requires such legislative permits (Baumann 2015).

Furthermore, there are differences between research-elicited data and traditional-type big data when it comes to *data analysis*. While answering the research question usually involves multivariate analysis methods since research questions pertaining to only two variables usually do not fit the complexity of social reality, for practical reasons, administrative data are typically analyzed by means of descriptive statistics only (also due to the fact that they do not fulfill the criteria for statistical inference drawing). Since the legislation principle is primary in administrations/institutions and their needs for conclusions drawn from the data are often different, it is rare that authorities want to know the likelihood that they are drawing false or correct conclusions from a data set. The latter is of primary importance for the scientific research process, as suggested by the "total survey error" paradigm (Groves 2004).

When both data types are used for secondary data analysis, similar methods can be applied if the administrative data are preprocessed to be used in statistical analysis (Baur 2009, 17). Wallgren and Wallgren (2014, 3) point out that administrative registers need to meet the demands of four principles, namely:

1) the registers "should be transformed into statistical registers" by referring to all relevant sources,
2) "the statistical registers should be included in a coordinated register system" in order to make sure that the integration will work,
3) the consistency of population and variables is warranted in order to get coherent "estimates from different register surveys,"
4) a quality assessment should be applied by comparing the statistical surveys "with other surveys in the production system."

In their book on "Register-Based Statistics," Wallgren and Wallgren (2014) refer mostly to Swedish administrations. It is not clear yet whether these principles can be applied to administrations in other countries in the same way, due to differences in legislation and administrative conductions. However, if these principles can be applied, the coherence of the population and variables allow for coherent estimates from different surveys.

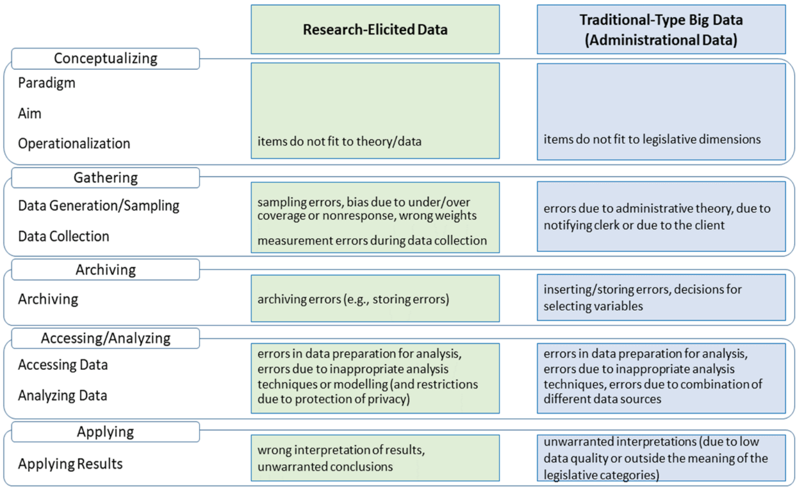
Finally, the *application of results*, for example the conclusions drawn from the data analysis, are usually also different: Results in the scientific area can, but do not need to, have practical implications, while this is the main criterion when it comes to conclusions from administrative data analysis. The latter are most often used in the media or in politics to lay foundations for policies (such as a report on crime prevalence for increasing the number of patrolling police officers, Neumann, and Graeff 2015).

3. A Comparison of Problems and Errors of Survey Data and Traditional-Type Big Data

Having compared the research processes between research-elicited and traditional-type big data, the next question is, which problems and errors occur during data generation, both in the scientific and non-scientific field? From a social science point of view, corruption, as a form of crime, is registered by prosecuting authorities (and the public prosecution departments and courts). This generates a special type of administrative data which has to be examined according to its characteristics and features in order to learn about its errors, biases, and problems (Bick and Müller 1980, 1984). Before we turn to the suggestions that Bick and Müller make regarding "data lore" (which refers to finding data distortions rather similar to errors), we will briefly summarize the problems of survey data production to provide grounds for a comparison with administrative data.

Errors within the scientific *survey process* can occur at any stage in Figure 1; that is to say, they can occur while considering the aim and operationalizing within the specific paradigm, while gathering, archiving or accessing/analyzing data, or applying the results practically. For operationalizing either theoretical ideas or legislative directives, one could produce questions or items for surveying that do not fit to theory or to the conditions of producible data (Hox 1997, 66). Detailed possibilities for errors are provided in the literature (e.g., Baur 2009). When it comes to the first stage in sampling, errors in defining and covering the target populations might occur. If one wants to use inferential statistics later, the selection of cases needs to follow random procedures which leave ample room for errors (such as those produced through item- or unit-nonresponses as well as erroneous adjustments or sample weights). Measurement errors pertaining to issues of validity or reliability are linked to processes of survey designing and data collection procedures. There are also particular problems that cause errors when using specific modes of data collection (such as memory problems of interviewees confronted with retrospective questions; see Baur 2009, 29 and Baur 2014 for an overview). Ultimately, errors might also occur during data analysis through choosing inappropriate analysis techniques (e.g., inappropriate models) or falsely interpreting results (see Figure 2).

Figure 2: Errors Specific to the Data Types

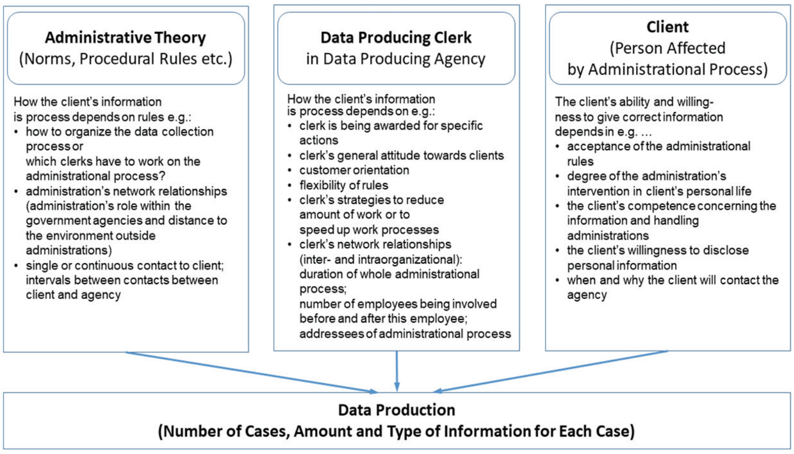


	Research-Elicited Data	Traditional-Type Big Data (Administrational Data)
Conceptualizing Paradigm Aim Operationalization	items do not fit to theory/data	items do not fit to legislative dimensions
Gathering Data Generation/Sampling Data Collection	sampling errors, bias due to under/over coverage or nonresponse, wrong weights measurement errors during data collection	errors due to administrative theory, due to notifying clerk or due to the client
Archiving Archiving	archiving errors (e.g., storing errors)	inserting/storing errors, decisions for selecting variables
Accessing/Analyzing Accessing Data Analyzing Data	errors in data preparation for analysis, errors due to inappropriate analysis techniques or modelling (and restrictions due to protection of privacy)	errors in data preparation for analysis, errors due to inappropriate analysis techniques, errors due to combination of different data sources
Applying Applying Results	wrong interpretation of results, unwarranted conclusions	unwarranted interpretations (due to low data quality or outside the meaning of the legislative categories)

Source: Modified from Baur 2009 and Bick and Müller 1984.

For *traditional-type big data*, problems and errors occur at the same stages as for the survey process. Bick and Müller (1980, 1984) consider the conceptualization of administrative data in close relation to survey data as official figures can be taken as indicators for a quantification of a topic of interest. This also leads to the possibility of taking these indicators (such as the number of crimes in a region) as an indicator for other topics within social science (such as anomie), as long as there are grounds for the suggestion that these indicators are related to the topics. The paradigm, however, under which these data are operationalized is strictly focused on legislation, which narrows the scope of reality and the amount of features that are of interest for administrative purposes. A first type of error occurs if items do not fit the legislative dimensions (for an example, see Rasner et al. 2007). The procedural rules and (legal) norms of conducting (mandatory in the "administrative theory"; Baur 2009, 25) impact the techniques of data collection, the work, and leeway of administrative staff (that deals with the data). Similarly, the role and functions of the administration within the governmental/political structure and its "distance to the environment outside administration" are also of importance.

Figure 3: Factors Possibly Contributing to Distortion of Data Production in Process-Produced Data such as Public Administration Data



Administrative Theory
(Norms, Procedural Rules etc.)

How the client's information is process depends on rules e.g.:
- how to organize the data collection process or which clerks have to work on the administrational process?
- administration's network relationships (administration's role within the government agencies and distance to the environment outside administrations)
- single or continuous contact to client; intervals between contacts between client and agency

Data Producing Clerk
in Data Producing Agency

How the client's information is process depends on e.g.:
- clerk is being awarded for specific actions
- clerk's general attitude towards clients
- customer orientation
- flexibility of rules
- clerk's strategies to reduce amount of work or to speed up work processes
- clerk's network relationships (inter- and intraorganizational): duration of whole administrational process; number of employees being involved before and after this employee; addressees of administrational process

Client
(Person Affected by Administrational Process)

The client's ability and willingness to give correct information depends in e.g. ...
- acceptance of the administrational rules
- degree of the administration's intervention in client's personal life
- the client's competence concerning the information and handling administrations
- the client's willingness to disclose personal information
- when and why the client will contact the agency

Data Production
(Number of Cases, Amount and Type of Information for Each Case)

Source: Bick and Müller 1984a, 138; translated, adjusted, and reprinted in Baur 2009, 25, and slightly modified for this publication.

The body of quantitative administrative data contains cases such as persons or objects for each of which the same information is gathered (Bick and Müller 1984, 123). Persons or objects need to be registered at the data producing institution (such as the police) before information about this case can be stored (implying that non-registered information leads to a missing data problem; Brame et al. 2010, 274). Different types of information are stored in a manner depending on the function of the administration/institution; the amount and content of information about a case is stored according the functional division of the administration/institution.

The data generating and "sampling" processes reveal clear differences to the survey process of research-elicited data. Bick and Müller (1984, 138-139; see also Baur 2009, 25) identify three *main factors influencing the particular way in which administrative data are generated*: The aforementioned rules and directives ("*administrative theory*"), the *data producing clerk* (in the administration/institution), and the person who is affected by the administrative process ("*the client*"). Administrative service regulations and the way in which the notifying administrative clerk (such as a policeman or a public official in a bureaucracy) applies them depends on her/his procedural leeway, attitude, customer orientation, and behavior towards clients. The notifying process changes if the clerk wants to skip paperwork or is awarded for registering specific information. This might produce distortions in the administrative data during the data generating process as do changes in intraorganizational features of the notifying/registering process, such as the duration of the process or the involved administrative positions. Inserting and coding processes are general sources of errors in administrative data, reducing data accuracy and reliability (for an application with US data, see Abowd and Vilhuber 2005). In general, any relevant information for an administration/institution that is reported by a client enters the administrative database as a combination of the report itself and the administration's documentation practice (Mac Donald 2002, 103). As the administrative data generating process needs a person who reports the information (such as a witness for crimes or a whistleblower for corrupt practices), this report may potentially lead to distortions. Delivering information to the administration's/institution's representatives presupposes that the reporting person accepts the administrative rules and the way its clerks conduct their work. The information is merely the description of a situation, by the reporting person, which could severely differ from actual situational characteristics (Kersting and Erdmann 2014, 16), in particular if crime or deviant behavior is involved. Usually, personal information of the reporting person is also revealed and requires the person's prior acceptance.

A good deal of the potential data distortions stem from individual behavior during interactions between people inside and outside of the administration/institution. Unlike the interaction between interviewer and interviewee in survey research, this interaction is not standardized – although administrative rules and directives exist.

Archiving the administrative data is also guided by legislative directives; the information being stored is selected by means of directives and normative assessments on what is worth archiving. Since the administrative data are not assembled in order to answer specific questions (as opposed to survey data), there might also be data distortions in the archiving process. Therefore, all rules of selection need to be known in order to assess these distortions (Bick and Müller 1984, 137).

Accessing and analyzing administrative data at the micro-level is limited (sometimes even prohibited) due to the protection of privacy (e.g., if the data refer to criminal offenses; Baumann 2015, 75), even if researchers are aware that suitable administrative data are available (Baur 2009, 31). Since the administrative data are not designed for answering specific (research) questions, the directives which determine the way in which the data are stored (for instance the categories for storing the age of a person) could complicate a detailed statistical analysis (Kersting and Erdmann 2014, 15). Moreover, sometimes the data body in administrations is not generated within a determined time span (as research-elicited data usually are) but consists of information from different sources assembled at different (past) points in time (Bick and Müller 1984; Wallgren and Wallgren 2014), which suggests an interpretation error if these data were perceived as a homogenous entity. Another analysis problem occurs if the directives for generating the administrative data change so that data become incomparable across time (Herrmann 2009, 649). All these difficulties suggest that administrative data cannot be analyzed without additional effort in data preparation, as compared to survey data (Wallgren and Wallgren 2014, 182).

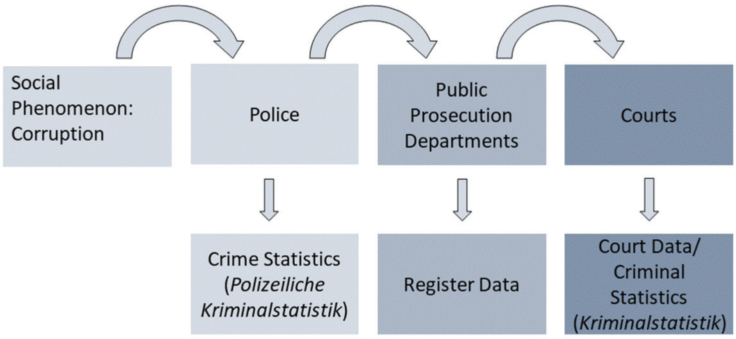
How the results derived from the data are used also differs between research-elicited and non-research-elicited data. Answers to research questions with research-elicited data are usually considered as valid, given a chosen significance level – the probability of rejecting a hypothesis which is actually true. Empirical research results based on the foundations of inferential statistics need to be approved or rejected by other studies' results and are contested by nature. They are sometimes utilized, however, as grounds for recommendations on societal/policy issues. Results derived from administrative data are also applied more and more in governmental or political matters (see "Commission on Evidence-Based Policymaking" 2017; Wallgren and Wallgren 2014, 28; Oberski et al. 2017) but are only rarely judged by the same (measurement-related) standards as research results are. Errors in applying the results occur if conclusions are derived which are unwarranted given the quality and the amount of errors within the data itself. For example, consider the interpretation of crime rates for a region without knowing the legal definitions and data keys that were used for coding the data and without any written documentation regarding the data sets. Increases in crime rates across years must not necessarily be due to an increase in committed crimes but rather may be based on improvement in police work. For using the results, data need to be interpretable in a meaningful sense. To achieve this, a data dictionary that guides data users is usually necessary. Even if this is available, it is questionable whether the interpretation of quantifications of administrative/legislative categories are conducted in the way that was suggested by the law. In the following section we will illustrate this with crime data from Germany.

4. Example: Characteristics of German Corruption Data

If one wants to use different data sources for examining the social phenomenon of corruption, several potential sources are available. As mentioned previously, some survey items exist in large German population surveys such as the General German Social Survey (ALLBUS). Also, there are administrative data from the police, public prosecution departments, and courts on numbers of suspected, processed, and sentenced cases of corruption. These data reflect discovered cases of corruption with a clear jurisdictional meaning.

In addition to these data sources, there are spatial/structural data about corruption (which measure the number and quality of anti-corruption institutions, public prosecution, and prevention measures within a German county or region). Moreover – already revealing typical qualities of big data – media data (Williams, Burnap, and Sloan 2017) exist that can be taken as an indicator for corrupt practices.

Figure 4: Processing Administrative Corruption Data


Social Phenomenon: Corruption
Police
Public Prosecution Departments
Courts
Crime Statistics (*Polizeiliche Kriminalstatistik*)
Register Data
Court Data/ Criminal Statistics (*Kriminalstatistik*)

In this section, we will apply the idea of "data lore" by Bick and Müller (1980, 1984) to administrative data on corruption in order to identify distortions or errors within an existing dataset. In the first step, we turn to the way in which data are produced by the prosecution authorities. Take again the example of a firm's employee who bribes a public official to get a public contract for her/his firm. Someone (such as a whistleblower) has to press charges against the persons who are suspected of having committed this type of bribery (see Figure 4).[2]

[2] Note that our scheme is stylized in the sense that not all crimes are primarily registered by the police. Depending on institutional criteria, such as the severity or type of crime, sometimes the public prosecution department is the authority which first registers the crime (such as cases of fraud and embezzlement). Sometimes (such as in cases of theft or misappropriation) the police are the registering authority (Baumann 2015, 84).

The “crime statistics,” created from cases registered by the police, comprises the number of corruption cases that are *suspected* to have been committed. These statistics are the point of departure for the next jurisdictional steps: The cases are handed over to the responsible prosecution department where a decision is made on whether or not to take a case to court. When a case goes to court, it is decided (in court) whether or not the suspicion is confirmed in a juristic sense. As a result, the number of suspects who are convicted for corruption appears in the final “criminal statistics” (built up of data from the courts). Typically, the number of cases registered (and suspected) by the police is far higher than the number of convicted persons in the criminal statistics (from the courts). This is partly due to the fact that during the administrative process the charge can be reassessed if new aspects regarding the suspected crime appear or the charge is not forwarded to the next jurisdictional stages because of formal reasons (Enzmann 2017), such as too much time elapsing since the crime. It is also partly due to erroneous reporting (which is an error similar to false answers in surveys; Pepper et al. 2010). In addition, the lower number of convicted persons (compared to suspected persons) results from the fact that not all charges are actually valid.

The statistics derived by means of this procedure include data “known to the police,” pointing to the major problem in criminology that official figures reflect only registered crimes and leave an unknown amount of crime (the “dark figures”) out of consideration (Brame et al. 2010, 274). The final statistics from court data preserve this mistake but must be considered as different data, pointing to slightly different phenomena: Both statistics deal with corruption but reflect either the number of possibly corrupt offenders or the number of persons convicted based on charges of corruption (Heinz 2003).

Applying the idea of “data lore” (Bick and Müller 1984; Baur 2009) to these administrative data allows for examining the administrative *processes for potential distortions or errors*.

First of all, “*administrative theories*” on what counts as corruption differ historically and culturally. For example, in Germany giving a gift above 25 € to a public servant is considered a crime. In effect, this means that even inviting a politician or another public servant to a dinner or gifting them with an expensive bottle of wine might be considered corruption in Germany, if it is not formally approved in advance. In other countries, these acts might be considered as acts of common courtesy. These differences in legislation will have an effect, which reported cases are classified as corruption and which are not. A good example for Germany is the case of former German President Christian Wulff who had to resign and was persecuted for accepting gifts that might have been considered insignificant in other countries.

Second, the “the clients” are important in process-produced data on corruption because not all cases of corruption are reported to the police. Corruption, e.g., might not be observed by anyone apart from the persons involved in the act (who naturally do not have any interest in reporting it). Even if corruption is observed, the persons observing the act might not consider the act as an act of corruption. Even if a person observes an act of corruption and wants to report it, they might be afraid to do so due to social or legal ramifications and thus never report it. Therefore, whistleblowers are quite rare and the number of possible cases of corruption reported are much lower than the actual cases due to "*client logic*."

Thirdly, even if a case of corruption is actually reported, there are many possible errors in the administrative processes in which the police are involved as the notifying authority. These potential errors due to the "*data producing clerk in the data producing agency*" (Bick and Müller 1984, 138) occur because of practical issues in the notifying/registering process, the "characteristics" of the police as data producing clerks, and general conditions leading to erroneous information.

In the practical work of the *police*, strategies for evaluating a report of a potential crime, assessing the reporting person, and the knowledge for classifying a potential crime play major roles (Rüping 2009). When faced with accumulating numbers of reported (potential) crimes, pressure on police officers from key management suggests short cuts for settling the work amount (Eterno et al. 2014). For corrupt incidences, the reporting of the suspected crimes (usually reported as a form of whistleblowing) depends on the way that the notifying authorities treat this report (cf. Koster 2016; Slocum 2018). Notifying police officers request the information which is necessary, from their point of view, in order to keep their work load within reasonable limits (Herrmann 2016, 44). For removing any disincentives, the whistleblowing of corrupt practices needs to be fully covered by legislative rules (Carr and Lewis 2010) and potential whistleblowers need to be sure of their (legal) protection. Not meeting these conditions results in a larger number of dark figures of crime, due to less reporting. There are also potential distortions suggested by the Bick and Müller's (1984) "data lore" in conjunction with the manner of processing pictured in Figure 4: If there are inconsistencies in registering the demanded information and this deficient information package is handed over to the next authority, errors can also be persevered and transferred to the next jurisdictional stage.

The "characteristics" of the police as data generating authority are tightly linked to their practical issues in registration, for example the police officers' "attitudes" against particular reporting groups (Slocum 2018) and the conditions of work conduction (low level of staff and equipment or specific office functions). A shortage of police personnel results in an "institutional inertia" in regard to the reporting of crimes (van Dijk 2009). Also, a shortage in equipment is perceived as an obstacle for completing office work (Bonewasser 2000, 235).

Further errors occur in the reporting situation if the registering police officer is not able to get proper information about the notified crime – either because the reporting person states the offense in an unclear way or because there is a lack of valid information in her/his statements (Brusten 1984). Under these work conditions given for registering crimes, the data derived from the first authority, which registers the crime (corrupt practices), are compiled in the police crime statistics (*Polizeiliche Kriminalstatistik*, PKS) that includes the registered number of suspected cases for a particular type of crime.

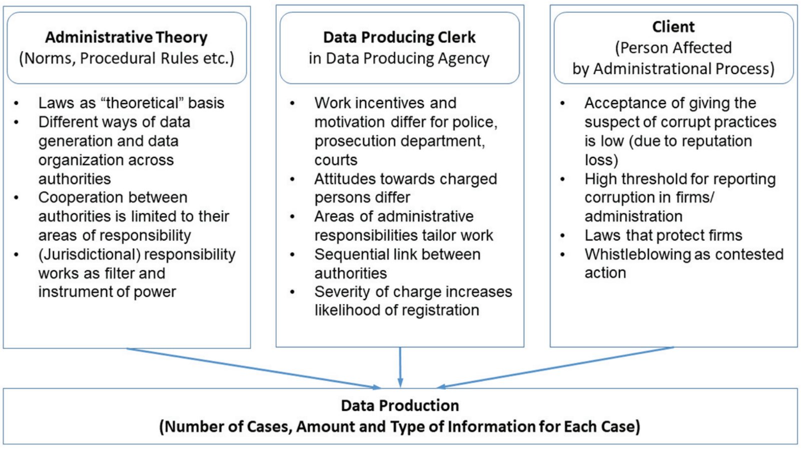
The "data lore" concept also applies to the stage of the public prosecution departments and for the courts. The *prosecution departments* aim at contributing to the jurisdictional conducting of the criminal cases. They consider and compile information that point to rebutting or confirming the charge. In this process, strategies for managing the amount of work that typically appear are similar to those of the police. There are empirical hints that the number of cases that need be taken care of are too high for (some) public prosecution departments (Elsner 2008).[3] The prosecution departments work as a de facto filter because only a fraction of cases is passed on to the courts. Terminating a charge is among the most frequent methods for tackling a case (Heinz 2004; Baumann 2015, 80). Even if legislative directives guide the departments' work, there is (jurisdictional) leeway in the consideration of the cases (Mayntz 1985, 213). An authority's (functional) responsibility is an important aspect (in practice) for settling, shortening, or processing a case (Göhler 2018). On the one hand, this reflects the application of legislative principles par excellence; on the other hand, it is sometimes an instrument (of power) for delaying the processing of a case, or its termination. This is, however, not visible in the data generated at the level of the prosecution department.

Data generated in the *courts stage* enter the criminal statistics (*Kriminal- oder Strafverfolgungsstatistik*), such as the number of people convicted due to a specific charge. The data represent the decisions of courts according the legal assessment of a charge. Courts are free (unlike the police and the prosecution departments) in their manner of deciding a case, that is they do not have to regard the opinions or directives of others (such as the ministry), as long as they comply with the law (Schmidt 2018).

If the three major factors of the "data lore" concept by Bick and Müller (1984, 138-139) are applied to administrative corruption data in Germany, the three factors (the "administrative theory," the data producing clerk [in the administration/institution], and the person who is affected by the administrative process ["the client"]) can be identified as "error-generating" (see Figure 5).

[3] Each federal state (*Bundesland*) in Germany has its own public prosecution department implying that the amount of work, the working conditions and resources, and also some legislative directives differ across counties.

Figure 5: Application of the Bick and Müller (1984) Concept of "Data Lore" to Corruption Data in Germany


Administrative Theory
(Norms, Procedural Rules etc.)
• Laws as "theoretical" basis
• Different ways of data generation and data organization across authorities
• Cooperation between authorities is limited to their areas of responsibility
• (Jurisdictional) responsibility works as filter and instrument of power
Data Producing Clerk
in Data Producing Agency
• Work incentives and motivation differ for police, prosecution department, courts
• Attitudes towards charged persons differ
• Areas of administrative responsibilities tailor work
• Sequential link between authorities
• Severity of charge increases likelihood of registration
Client
(Person Affected
by Administrational Process)
• Acceptance of giving the suspect of corrupt practices is low (due to reputation loss)
• High threshold for reporting corruption in firms/ administration
• Laws that protect firms
• Whistleblowing as contested action
Data Production
(Number of Cases, Amount and Type of Information for Each Case)

The administrative work of criminal prosecution is driven by laws and directives and produces different data on cases of suspected persons or convicted criminals at each stage of processing. An important element in the cooperation of the authorities is their functional responsibility that limits their work competence and capacity. Cases are usually filtered across the stages (e.g., during the recording process by the police; Enzmann 2017, 67) and selected due to the administrations responsibility.

Work-saving strategies for settling charges exist across all authorities, such as applying certain rules in order to expand or shorten a jurisdictional process (Göhler 2018). They are tightly linked to the specific areas of administrative responsibilities. The capacity of an authority tailors its work and influences the work motivation of the public officials. All of these factors are embedded in the link between the authorities, influencing a case when it proceeds from one authority to the next.

A reporting client who points to corrupt practices in firms or administrations usually has to overcome high social and regulative thresholds, as the acceptance of colleagues and superiors hearing news about corrupt practices in the organization is low. There are also laws in Germany stating that employees do not have to report lesser crimes or irregularities. As whistleblowing may put a firm or administration under general suspicion, it is sometimes perceived as a contested action (Perry 1998).

Summing up, *the "data lore" perspective allows for a detailed and structured analysis of the errors that occur in traditional-type big data such as administrative crime data*. A part of the errors in these data can be considered "measurement errors" as they concern distortions in capturing the corruption phenomenon (such as false reporting). Other errors correspond to sampling/coverage errors because of the fact that social and administrative obstacles exist that prevent the reporting of corrupt practices.

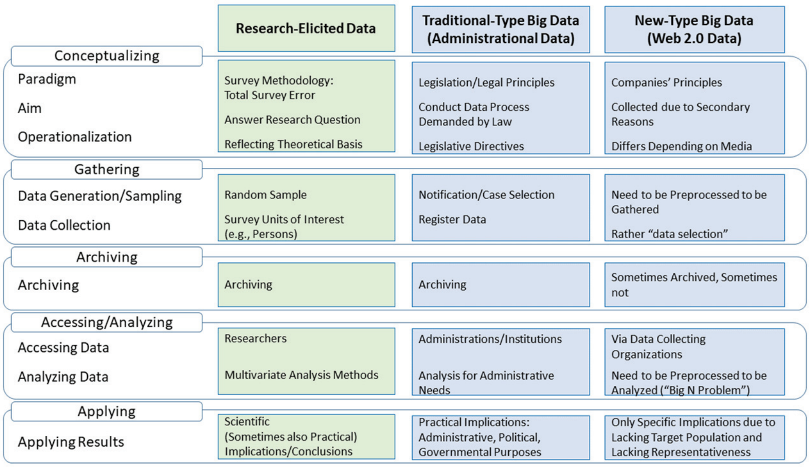
While survey data and traditional-type big data (such as administrational data) are reconciled with each other from the perspective of errors when using the concept of "data lore," new-type big data have opened up new opportunities for social research (Baur et al. 2020, in this HSR Forum). The next section succinctly addresses the question if this new "data paradigm" actually adds to the already existing data types.

5. Using New-Type Big Data for Exploring Corrupt Practices

The digital turn of the recent years and its new opportunities for getting and analyzing data which are (at least indirectly) linked to a social phenomenon of interest is prevalent in social science discourse. As Doorn and Tjalsma (2007, 7-8) put it, data sources and archives that were rather separate in the past are starting to converge in the digital world, leading to new challenges for survey data and administrative record data. Since administrative data always touch dimensions of legality, digital documents are different from "paper documents" in their relation to the "original document." The "authenticity" of information contained in data becomes crucial, both for survey and administrative data.

It is important to keep in mind that public administrative data (Wallgren and Wallgren 2014, 299; Conelly et al. 2016) also are "big data." In fact, the original German name for them was *Massendaten*, which translates as "big data" or "mass data" – in the context of this paper, we have been using the term "traditional-type big data" in order to distinguish these data from new-type big data created in the context of Web 2.0 (for a detailed discussion on terminology, see Baur et al. 2009). This is important in the current debate on big data because it reveals that big data are a rather old data type dating back at least to the 18th century when modern bureaucracy was invented. However, in recent years, new types of big data have arisen which denote large volumes of data that usually (but not only) occur as a result of online activities (such as ecommerce) or GPS tracking by digital gadgets (Baur et al. 2020, in this volume). The defining features of (both traditional-type and new-type) big data provided by Laney (2001) are prominent in the literature (see also Weichbold et al. 2020, in this volume). One feature is the fact that big data do not come as single cases. They occur in huge amounts of data points, typically so big that usual means of storing, accessing, and analyzing are difficult. In contrast to survey data, new-type big data come in a lot of different formats (ranging from numbers over picture- or video-data to text-strings). In contrast to survey data and traditional-type big data such as administrative data, new-type big data are closely related to new technologies, for example mobile phones. In addition, new-type big data often allude to exchanges in social networks (such as Twitter) about real-time events (Tinati et al. 2014). Two other major differences between traditional-type big data and new-type big data are, first, that traditional-type big data are usually collected by government agencies and therefore the administrative theory follows legal principles while new-type big data are often produced by large multinational companies (MNCs) who engrain their administrative theories in the software algorithms (Traue 2020). Thus, MNCs administrational logic of data production and data use follows business interests and is not democratically controlled. Second, the client logic has changed: Clients seem to be much more willing to reveal data to companies and to government agencies.

Figure 6: Features of Data (including New-Type Big Data)

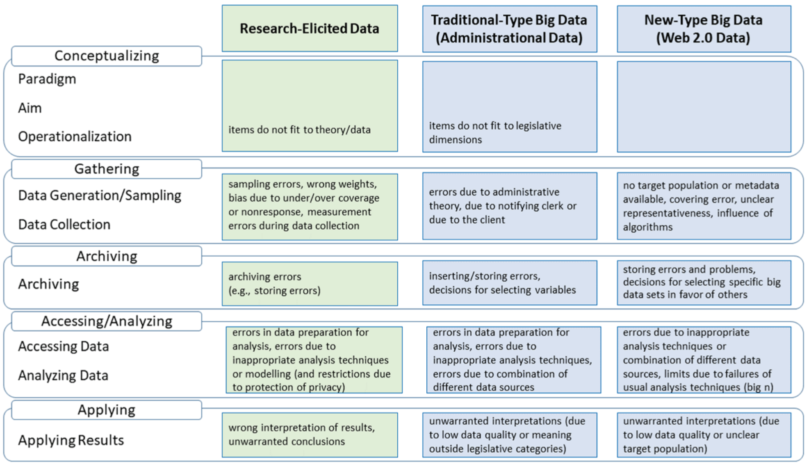


	Research-Elicited Data	**Traditional-Type Big Data (Administrational Data)**	**New-Type Big Data (Web 2.0 Data)**
Conceptualizing			
Paradigm	Survey Methodology: Total Survey Error	Legislation/Legal Principles	Companies' Principles
Aim	Answer Research Question	Conduct Data Process Demanded by Law	Collected due to Secondary Reasons
Operationalization	Reflecting Theoretical Basis	Legislative Directives	Differs Depending on Media
Gathering			
Data Generation/Sampling	Random Sample	Notification/Case Selection	Need to be Preprocessed to be Gathered
Data Collection	Survey Units of Interest (e.g., Persons)	Register Data	Rather "data selection"
Archiving			
Archiving	Archiving	Archiving	Sometimes Archived, Sometimes not
Accessing/Analyzing			
Accessing Data	Researchers	Administrations/Institutions	Via Data Collecting Organizations
Analyzing Data	Multivariate Analysis Methods	Analysis for Administrative Needs	Need to be Preprocessed to be Analyzed ("Big N Problem")
Applying			
Applying Results	Scientific (Sometimes also Practical) Implications/Conclusions	Practical Implications: Administrative, Political, Governmental Purposes	Only Specific Implications due to Lacking Target Population and Lacking Representativeness

Recently, big data sources have been picked up as alternative sources for scrutinizing social phenomena of crime. Social media data were used as predictors in a study by Gerber (2014) for predicting different types of crime in Chicago.

Williams, Burnap and Sloan (2017, 2) suggest that social media (such as Twitter) "[…] generate 'naturally occurring' socially relevant data that can be used to complement and augment conventional curated data to estimate the occurrence of offline phenomena." Referring to the broken window theory, they posit that Twitter posts about disorder are related to actual crime rates. They merge new media and administrative record data, at the London borough level as the unit of spatial analysis, and find evidence for an association between aggregated Twitter posts and police-recorded crime rates.

Figure 7: Errors by Data Type (Including New-Type Big Data)



	Research-Elicited Data	Traditional-Type Big Data (Administrational Data)	New-Type Big Data (Web 2.0 Data)
Conceptualizing Paradigm Aim Operationalization	items do not fit to theory/data	items do not fit to legislative dimensions	
Gathering Data Generation/Sampling Data Collection	sampling errors, wrong weights, bias due to under/over coverage or nonresponse, measurement errors during data collection	errors due to administrative theory, due to notifying clerk or due to the client	no target population or metadata available, covering error, unclear representativeness, influence of algorithms
Archiving Archiving	archiving errors (e.g., storing errors)	inserting/storing errors, decisions for selecting variables	storing errors and problems, decisions for selecting specific big data sets in favor of others
Accessing/Analyzing Accessing Data Analyzing Data	errors in data preparation for analysis, errors due to inappropriate analysis techniques or modelling (and restrictions due to protection of privacy)	errors in data preparation for analysis, errors due to inappropriate analysis techniques, errors due to combination of different data sources	errors due to inappropriate analysis techniques or combination of different data sources, limits due to failures of usual analysis techniques (big n)
Applying Applying Results	wrong interpretation of results, unwarranted conclusions	unwarranted interpretations (due to low data quality or meaning outside legislative categories)	unwarranted interpretations (due to low data quality or unclear target population)

While first empirical studies that apply new-type big data for scrutinizing social science topics now exist, the quality of these data is a topic that still remains open for discussion and assessment. As new-type big data are usually generated in conjunction with modern technologies, for example computer systems, the quality criteria need to meet the requirements of the research field of computer science. These requirements are different from those in the social sciences. Lohsin (2013, 41) considers "data quality dimensions" (such as the amount of accurate data values) of "traditional approaches" as hardly adoptable to big data since "[…] big datasets neither exhibit these characteristics, nor do they have similar types of business impacts. Big data analytics is generally centered on consuming massive amounts of a combination of structured and unstructured data from both machine-generated and human sources. Much of the analysis is done without considering the business impacts of errors or inconsistencies across the different sources […]." Data quality approaches are, therefore, scarce in computer science. Merino et al. (2016, 124) have delivered a data quality model that "[…] can be used to assess the level of Quality-in-Use of the data in Big Data." Since firms or administrations are the clients and contracting authorities of big data applications, the model considers the (international) industry standards (Merino et al. 2016, 124): "[…] it is paramount to align the investigation with the best practices in the industry in order to produce repeatable and usable research results. Taking advantage of the benefits of using international standards is one of those best practices." The "quality dimensions" suggested by Merino et al. (2016, 127) take Laney's "three V's" (volume, velocity, variety) into account. Whether their quality approach and similar ones are commensurable with “traditional approaches” remains an open question.

6. Conclusion

If data are considered to be information that reflects the materialization of a social phenomenon, it is striking that no data source (neither survey nor administrative data nor digital indicators) is capable of measuring the social phenomenon of corruption directly. The previously mentioned act of bribing (in order to get a contract for a firm) may really have taken place but its “measurement” is necessarily error-prone and should rather be treated as a probabilistic realization than a factual result. This holds true despite the fact that the data sources utilize rather different means for approaching (e.g., measuring) such incidents. *Surveys* investigating deviance usually aim to measure the propensity or willingness to commit a crime, which is moderately correlated with the committing of criminal acts (cf. Fishbein and Aizen 1980). Whether these acts have actually taken place remains questionable; this is also true for traditional-type big data such as *administrative data*. Any reported case can (potentially) be fake information based on various motives, such as the intention to denounce someone else by means of framing them for something they have not done. The same applies to new-type big data such as social *media data* on crime in which reporting biases also occur (Nasser and Tariq 2015; Williams, Burnap, and Sloan 2017, 15). If new-type big data are used for social science research, it might therefore be advisable to maintain an “error perspective” – both in regard to measurement issues and the analysis. This might come into conflict with data quality considerations from computer science which posits that “[d]ata quality is reached and preserved within a computer system” (Merino et al. 2016, 125). Further discussion of the requirements of data quality standards seems to be necessary here.

Bick and Müller’s “*data lore approach*” suggests that data quality – in equivalence to the perspective of the total survey error approach – depends on the distortions that come about when generating or processing traditional-type big data. Other scholars have defined data quality of traditional-type big data differently, usually referring to the structure and features of the administrations in their home country. The idea that errors and distortions spoil data quality is, however, predominant (Hand 2018, 562). The specific errors in administrative data, however, typically differ between the administrations of different countries. For example, Wallgren and Wallgren (2014, 292) assess the quality of register surveys based on the occurrence of relevance and integration errors. If an administrative definition of a statistical term, such as “target population,” deviates from the actual administrative content of interest (such as the population of the country), “relevance errors” can occur. Wallgren and

Wallgren (2014, 133) illustrate this error by pointing to the persons who are registered as Swedish citizens but do not permanently live in the country. If administrative data from different registers are integrated into a common register, any inconsistencies occurring during this process are considered to be "integration errors." Other authors have pointed to similar problems (describing these as errors), some (such as Berka et al. 2012) have come up with solutions for specific types of administrative data.

A promising expectation that is often stated is that different data sources can complementarily add to each other in the research process. This seems to be particularly important if the *complementary nature of data* refers to the same phenomenon, such as in the case of corruption. Connelly et al. (2016, 10) refine this point for traditional-type big data data:

> Administrative social science data offer the opportunity to study policy changes, social problems and societal issues using information which may not routinely be available in social surveys. The large size of many administrative social science data resources may offer the opportunity to study sub-groups, and could potentially lead to analytical approaches such as quasiexperimental methods being used more routinely. The re-purposing of these data could also result in long term savings for government departments, and social science data producers.

These hopeful expectations require empirical evidence. Moreover, these expectations must be met when new-type big data are also utilized as a complementary source. Up to now, there have been a number of negative examples in which the results do not fit together, such as Lazer et al. (2014). All in all, *the focus of discussion should shift from debating which is the best data type (there is not one best data type) to the advantages and disadvantages of using specific data types for answering specific research questions* – and if and how they can and should be possibly mixed (e.g., Baur 2011; Baur and Hering 2017). In addition, we should start reflecting on what it means for data quality and the power balances between institutions and citizens revealing their data when the data producing institutions are no longer government agencies, but companies instead.

References

Abowd, John M., and Lars Vilhuber. 2005. The sensitivity of economic statistics to coding errors in personal identifiers. *Journal of Business & Economic Statistics* 23 (2): 133-152.

Baumann, Thomas. 2015. Staatsanwaltschaftliche Ermittlungstätigkeit in Deutschland. Umfang und Struktur der Verfahrenserledigung. *Statistisches Bundesamt, Wista* 3: 74-88.

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

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 Jörg Blasius, eds. 2019. Handbuch Methoden der empirischen Sozialforschung. Wiesbaden: Springer Fachmedien. doi: 10.1007/978-3-658-21308-4.

Baur, Nina, Peter Graeff, Lilli Braunisch, and Malte Schweia. 2020. The Quality of Big Data. Development, Problems, and Possibilities of Use of Process-Generated Data in the Digital Age. *Historical Social Research* 45 (3): 209-243. doi: 10.12759/hsr.45.2020.3.209-243.

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 (2): 387-414. doi: 10.1007/s11577-017-0468-8.

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, B. 246-284. Weinheim: Beltz Juventa.

Berka, Christopher, Humer, Stefan, Moser, Mathias, Lenk, Manuela, Rechta, Henrik and Eliane Schwerer. 2012. Combination of evidence from multiple administrative data sources: quality assessment of the Austrian register-based Census 2011. *Statistica Neerlandica* 66 (1): 18-33.

Bick, Wolfgang and Paul J. Müller. 1980. The nature of process-produced data – towards a social scientific source criticism. In *The Use of Historical and Process-Produced Data*, eds. Clubb, Jerome M. and Erwin K. Scheuch, 369-413. (HSF-Historisch-Sozialwissenschaftliche Forschungen, Vol. 6). Stuttgart: Klett-Cotta. Full Text available at <https://www.gesis.org/hsr/volltext-archiv/buchreihe-historisch-sozialwiss-forschungen-hsf/hsf-6>.

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

Biemer, Paul P. 2010. Total survey error: Design, implementation, and evaluation. *Public Opinion Quarterly* 74 (5): 817-848.

Bonewasser, Manfred 2000. Mitarbeiterzufriedenheit in der Polizei: Weg von der abstrakten Beschreibung hin zur konkreten Veränderung. In Empirische Polizeiforschung. Interdisziplinäre Perspektive in einem sich entwickelnden Forschungsfeld, eds. Liebl, Karlhans and Thomas Ohlemachern, 35-47. Holzheim: Centaurus Verlag & Media.

Brame, Robert, Turner, Michael G. and Raymond Paternoster. 2010. Missing data problems in criminological research. In *Handbook of Quantitative Criminology Piquero, Alex R. and David Weisburd*, 273-288. New York: Springer.

Brusten, Manfred 1984. Die Akten der Sozialbehörden als Informationsquelle für empirische Forschungen. Möglichkeiten und Grenzen der wissenschaftlichen Konstruktion von Wirklichkeit auf Grundlage prozessproduzierter Daten aus Institutionen der Sozialverwaltung und der sozialen Arbeit. In *Sozialforschung und Verwaltungsdaten*, eds. Bick, Wolfgang, Mann, Reinhard and Paul J. Müller, 238-258. Stuttgart: Klett-Cotta.

Carr, Indira and David B. Lewis. 2010. Combatting corruption through employment law and whistleblower protection. *Industrial Law Journal* 39 (1): 1-30.

Charron, Nicholas, Dahlström, Carl, Fazekas, Mihaly and Victor Lapuente. 2017. Careers, connections, and corruption risks: Investigating the impact of bureaucratic meritocracy on public procurement processes. *The Journal of Politics* 79 (1): 89-104.

Connelly, Roxanne, Playford, Chris, Gayle, Vernon and Chris Dibben. 2016. The role of administrative data in the big data revolution in social science research. *Social Science Research* 59: 1-12.

Dickel, Petra and Peter Graeff. 2016. Applying Factorial Surveys for Analyzing Complex, Morally Challenging and Sensitive Topics in Entrepreneurship Research – The Case of Entrepreneurial Ethics. In *Complexity in entrepreneurship, innovation and technology research - Applications of emergent and neglected methods*, 199-217. Berlin: Springer.

Dickel, Petra and Peter Graeff. 2018. Entrepreneurs’ propensity for corruption: a vignette-based factorial survey. *Journal of Business Research* 89: 77-86.

Doorn, Peter and Heiko Tjalsma. 2007. Introducing: archiving research data. *Archival Science* 7 (1): 1-20.

Eterno, John A., Verma, Arvind and Eli B. Silverman. 2014. Police manipulations of crime reporting: insiders’ revelations. *Justice Quarterly* 33 (5): 811-835.

Elsner, Beatrix 2008. *Entlastung der Staatsanwaltschaften durch mehr Kompetenzen für die Polizei? Eine deutsch-niederländisch vergleichende Analyse in rechtlicher und rechtstatsächlicher Hinsicht*. Göttingen: Universitätsverlag Göttingen.

Enzmann, Dirk 2017. Reporting behavior and police recording practices. In *Victimisation Surveys in Germany,* eds. Leitgöb-Guzy, Nathalie, Birkel, Christoph and Robert Mischkowitz, 66-67.Wiesbaden: Bundeskriminalamt.

Gerber, Matthew S. 2014. Predicting crime using twitter and kernel density estimation. *Decision Support Systems* 61: 115-125.

Göhler, Johanna 2018. Strafverfahren ohne Hauptverhandlung im Rechtsvergleich – Ein Bericht über die 36. Tagung der Gesellschaft für Rechtsvergleichung aus strafrechtlicher Perspektive. *Zeitschrift für die gesamte Strafrechtswissenschaft* 130 (2): 513–525.

Graeff Peter, Sattler, Sebastian, Mehlkop, Guido and Carsten Sauer. 2014. Incentives and inhibitors of abusing academic positions: Analysing students´ decisions about bribing academic staff. *European Sociological Review* 30 (2): 230-241.

Groves, R.M. 2004. *Survey Errors and Survey Costs*. Hoboken: John Wiley & Sons.

Groves, Robert M. and Lars E. Lyberg. 2010. Total survey error. Past, present, and future. *Public Opinion Quarterly* 74 (5): 849-879.

Hand, David J. 2018. Statistical challenges of administrative and transaction data. *Journal of the Royal Statistical Society, Series A* 181(3): 555-605.

Heinz, Wolfgang 2003. Soziale und kulturelle Grundlagen der Kriminologie – der Beitrag der Kriminalstatistik. In *Kriminologie zwischen Grundlagenwissenschaft und Praxis,* eds. Dittmannn, Volker and Jörg-Martin Jehle, 149-185. Mönchengladbach: Forum Verlag Godesbergen.

Hermann, ieter. 2009. Soziologie des Strafverfahrens. In Handbuch der Forensischen Psychiatrie, eds. Kröber, Hans-Ludwig, Dölling, Dieter, Leygraf, Norbert and Henning Sass, 645-662. Beerfelden: Steinkopff.

Hermann, Dieter 2016. Die Konstruktion von Realität in Justizakten. *Zeitschrift für Soziologie* (16) 1: 1-12.

Hox, Joop 1997. From theoretical concept to survey question. In *Survey Measurement and Process Quality*, eds. Lyberg, Lars et al., 47-69. New York: John Wiley.

Kersting, Stefan and Julia Erdmann. 2014. Analyse von Hellfeld-Daten – Darstellung von Problemen, Besonderheiten und Fallstricken anhand ausgewählter Praxisbeispiele. In *Empirische Forschung über Kriminalität: Methodologische und methodische Grundlagen*, eds. Eifler, Stefanie and Daniela Pollich, 9-30. Wiesbaden: Springer Fachmedien.

Koster, Nathalie-Sharon N., Kuijpers, Karlijn F., Kunst, Maarten J.J. and Joanne van der Leun. 2016. Crime victims´ perceptions of police behavior, legitimacy, and cooperation: a review of the literature. *Victims & Offenders*: 1-44.

Laney, Doug 2001. 3D Data Management: Controlling Data Volume, Velocity and Variety. META Group Research Note 6.

Lazer, David, Kennedy, Ryan, King, Gary and Alessandro Vespignani. 2014. The parable of google flu: traps in big data analysis. *Science* 343: 1203-1205.

Lohsin, David 2013. *Big Data Analytics. From Strategic Planning to Enterprise Integration with Tools, Techniques, NoSQL, and Graph*. Waltham: Elsevier.

Mayntz, Renate 1985. *Soziologie der öffentlichen Verwaltung*. Heidelberg: C.F. Müller.

Merino, Jorge, Caballero, Ismael, Rivas, Bibiano, Serrano, Manuel and Mario Piattini. 2016. A Data Quality in Use model for Big Data. *Future Generation Computer Syste*ms 63: 123-130.

Nasser, Thabet and Rahim S. Tariq. 2015. Big data challenges. *Journal of Computer Engineering & Information Technology* 4 (3): 1-10.

Neumann, Robert and Peter Graeff. 2015. Quantitative Approaches to Comparative Analyses. *European Political Science* 14 (4): 385-393.

Oberski, Daniel L., Kirchner, Antje, Eckman, Stephanie and Frauke Kreuter. 2017. Evaluating the quality of survey and administrative data with generalized multitrait-multimethod models. *Journal of the American Statistical Association* 112 (520): 1477-1489.

Pepper, John, Petrie, Carol and Sean Sullivan. 2010. Measurement error in criminal justice data. In *Handbook of Quantitative Criminology*, eds. Piquero, Alex R. and David Weisburd, 353-374. New York: Springer.

Perry, Nick 1998. Indecent exposures: theorizing whistleblowing. *Organization Studies* 19 (2): 235-257.

Rasner, Anika, Himmelreicher, Ralf K., Grabka, Markus, and Joachim Frick. 2007. Best of Both Worlds: Preparatory Steps in Matching Survey Data with Administrative Pension Records; The Case of the German Socio-Economic Panel and the Scientific Use File Completed Insurance Biographies 2004. *DIW Data Documentation* 24. Berlin: Deutsches Institut für Wirtschaftsforschung (DIW).

Rüping, Hinrich 2009. Das Verhältnis von Staatsanwaltschaft und Polizei. Zum Problem der Einheit der Strafverfolgung. *Zeitschrift für die gesamte Strafwissenschaften* 95 (4): 894-917.

Schmidt, Georg 2018. Die richterliche Unabhängigkeit – eine Bestandsaufnahme. *Die Verwaltung* 51: 227-263.

Slocum, Lee A. 2018. The Effect of Prior Police Contact on Victimization Reporting: Results from the Police–Public Contact and National Crime Victimization Surveys. *Journal of Quantitative Criminology* 34 (2): 535-589.

Skogan, Wesley G. 1974. The validity of official crime statistics: An empirical investigation. *Social Science Quarterly* 55 (1): 25-38.

Tinati, Ramine, Halford, Susan, Carr, Les and Catherine Pope. 2014. Big data: methodological challenges and approaches for sociological analysis. *Sociology* 48: 663-681.

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

Van Dijk, Teun A. 2009. *Society and Discourse: How Social Contexts Infuence Text and Talk*. United Kingdom: Cambridge University Press.

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

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.

Williams, Matthew L., Burnap, Pete, and Luke Sloan. 2017. Crime sensing with big data: the affordances and limitations of using open-source communications to estimate crime patterns. *British Journal of Criminology* 57: 320-340.

Woollard, M. 2014. Administrative data: problems and benefits. A perspective from the United Kingdom. In *Facing the Future: European Research Infrastructures for the Humanities and Social Sciences*, eds. Dusa, Adrian, Nelle, Dietrich, Stock, Günter and Gert G. Wagner, Berlin: SCIVERO.

Historical Social Research
Historische Sozialforschung

All articles published in this Forum:

For further information on our journal, including tables of contents, article abstracts, and our extensive online archive, please visit https://www.gesis.org/en/hsr.