Generation and use of unstructured data in the social, behavioural, and economic sciences: challenges and recommendations

Veröffentlichungsversion / Published Version
Arbeitspapier / working paper

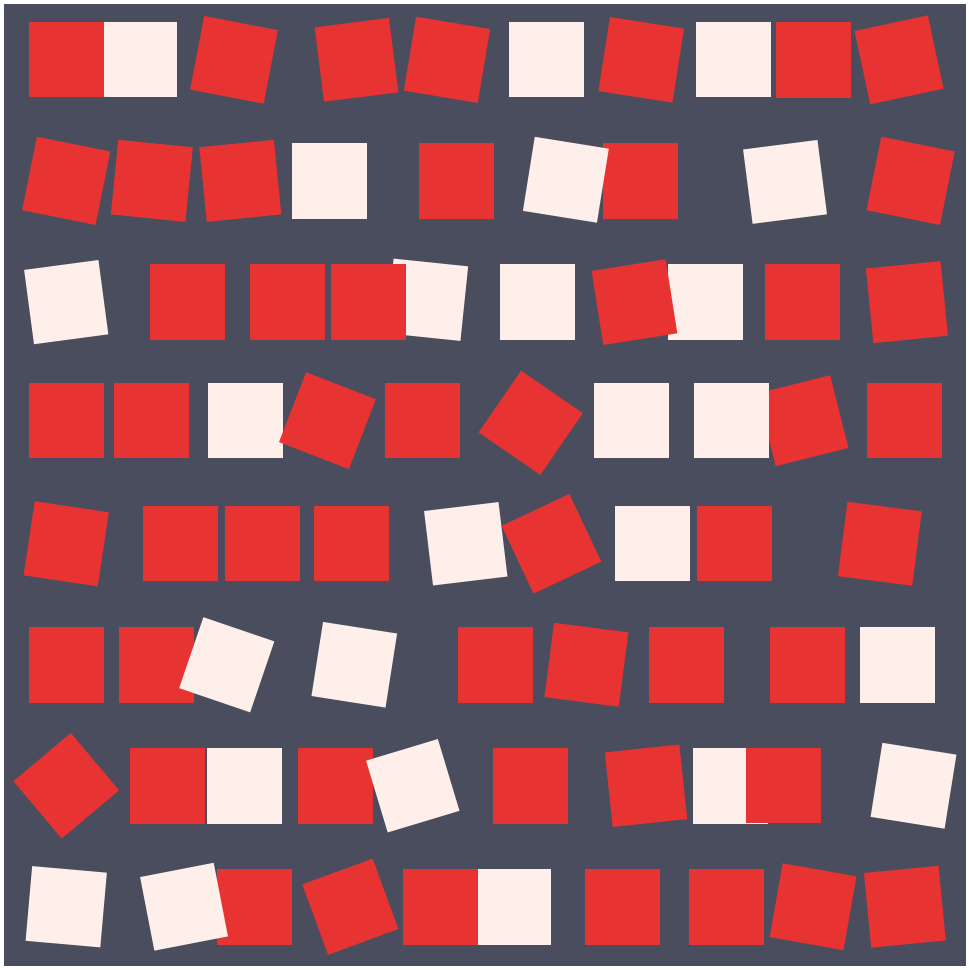
Empfohlene Zitierung / Suggested Citation:
Rat für Sozial- und Wirtschaftsdaten (RatSWD). (2024). *Generation and use of unstructured data in the social, behavioural, and economic sciences: challenges and recommendations.* (RatSWD Output Series, 2 (7)). Berlin. https://doi.org/10.17620/02671.92

2 | Output
7th appointment period (2020-2023)

RatSWD.
German Data Forum

Generation and use of **unstructured data** in the social, behavioural, and economic sciences

Challenges and recommendations



German Data Forum (RatSWD)

Generation and use of **unstructured data** in the social, behavioural, and economic sciences

Challenges and recommendations

List of Abbreviations

API . . . . . Application Programming Interface
DGPuK . . . . . Deutsche Gesellschaft für Publizistik und Kommunikationswissenschaft (German Communication Association)
EDMO . . . . . European Digital Media Observatory
GPS. . . . . Global Positioning System
HTML . . . . . Hypertext Markup Language
NFDI. . . . . Nationale Forschungsdateninfrastruktur (National Research Data Infrastructure Germany)
TEF . . . . . Total Error Framework
TSE framework . . . . . Total Survey Error Framework
RatSWD . . . . . Rat für Sozial- und Wirtschaftsdaten (German Data Forum)
URL. . . . . Uniform Resource Locator
XML . . . . . Extensible Markup Language

Content

List of illustrations

Abstract

The increasing digital transformation of society in recent decades has resulted in a number of new data sources for the social, behavioural, and economic sciences. Among many others, they include unstructured data, which are characterised by not being available in a fixed data format and are therefore not easy to process for data analysis (e.g., Facebook posts, Instagram images, YouTube videos, Twitter[1] messages). The use of unstructured data is linked to specific challenges, which arise precisely because the data are not typically collected as part of a controlled, scientific study but are often created in people's natural environments. Building on the results of an expert workshop, we describe the specific challenges of generating and using unstructured data and formulate recommendations for their use. Our recommendations are based on the total error framework and take into account data generation (definition of the units of analysis, coverage and sampling error, non-response, and missing data error), post-collection processing (specification error, validity, measurement error, and error in terms of content), and, lastly, data analysis (record linkage and processing errors, modelling errors, analytical errors). Finally, we discuss open questions and challenges to research using unstructured data. This output paper is aimed at students and researchers in the social, behavioural, and economic sciences on the one hand, and everyone working with unstructured data and drawing inferences from them for practical applications on the other.

1 During the translation process, the platform Twitter was renamed to X. However, for consistency with the original German text, we continue to use the term „Twitter" in this document.

1 Introduction

1.1 Definition of unstructured data and distinction from other terms

The increasing digital transformation of society in recent decades has resulted in a number of new data sources for the social, behavioural, and economic sciences. What these data often have in common is that they are not available in a fixed data format (e.g., a rectangular matrix with cases/observations in rows and variables in columns) and are therefore not easy to process further for data analysis (Eberendu, 2016). In the following, we group these data sources under the term ***unstructured data*** and distinguish them from traditional survey data. Figure 1 presents examples of unstructured data. Unstructured data are often characterised by high volume and require extensive processing to make them available to social, behavioural, and economic research. By contrast, structured data are available in a fixed data format (e.g., tables, datasets, databases) (Tanwar et al., 2015). Semi-structured data occupy an intermediate position between the two. Unlike structured data, they, too, are not available in a fixed format. However, they may contain structuring elements and may be exchanged easily (Eberendu, 2016; Tanwar et al., 2015). An example of semi-structured data is Extensive Markup Language data (XML data; Tanwar et al., 2015). XML data have a partial structure (e.g., hierarchical) and have some structural elements, including tags (Nyhuis, 2021). Documents using Hypertext Markup Language (HTML documents), for example, which many will know from the internet, are special types of XML documents (Bosse et al., 2021).

Figure 1: Selected examples of unstructured data

Social media	Facebook posts, Instagram images, YouTube videos, Twitter messages
General media	Text, images, video, voice recordings, music
Geodata	GPS data
Log data	Visits to websites, time spent on websites, email behaviour
World wide web	Websites, messages, blogs
General documents	Text, PDF files, scanned files
Financial data	Bank transactions, stock market data
Health data	Patient records, radiographs, scanner images

Source: Adapted from Eberendu (2016) and Taleb et al. (2018)

Unstructured data are similar to other types of data. Thus, unstructured data can be grouped under the term *big data*, whereby the term *big data* itself is not clearly defined. The term ***big data*** is typically understood to mean data that are characterised by high volume and high variety and are generated at a high velocity (Gandomi & Haider, 2015; Lazer & Radford, 2017; Tanwar et al., 2015). High variety is characterised by structural heterogeneity and by the fact that it can include structured data as well as semi-structured and unstructured data (Gandomi & Haider, 2015).

Unstructured data often encompass data that are generated using ***new information technology*** (RatSWD, 2020) (e.g., on the internet or from smartphones) and that represent people's digital lives (e.g., Facebook and Twitter data) (Lazer & Radford, 2017). They can capture aspects of the so-called *digitalised* life, which, according to Lazer and Radford (2017), represents the aspects of digital life that are not actively produced by a person (as opposed to aspects that a person actively produces, e.g., tweets) but were generated as a side product of digitalisation (e.g., capturing social closeness via Bluetooth). The distinction between digital and digitalised life corresponds to the distinction between intentional and non-intentional (Hox, 2017) or between traces of participation (active initiative) and transactional data (e.g., metadata on digital behaviour such as one's location) (Menchen-Trevino, 2013), respectively. Lastly, it is possible to record digital traces (record data) that are left behind when using digital devices. Lazer and Radford (2017), for example, referred to records of phone use in this regard. However, unstructured data are not to be

equated with data generated by new information technology because the latter also captures structured data, and unstructured data can also be collected through other means. Extensive text data could be mentioned here (Grimmer et al., 2022).

Unstructured data are often data that were obtained using ***non-reactive data collection*** methods. Data collection is considered reactive if the values it produces can be influenced by either the survey respondents or the investigators due to the nature of the data collection method (Fritsche & Linneweber, 2006). When unstructured data are generated, the people who are providing the data are often not aware that they are participating in a study or that their data are being used for such purposes (*unobtrusive measures*; Webb et al., 1966). This lack of awareness prevents data biases that may result from data collection or at least makes such biases highly unlikely to occur. However, unstructured data should not be equated with non-reactively collected data because structured non-reactive measurements also exist (e.g., structured observation of behaviour), and it is also possible for such data to be produced in contexts in which the people are aware that their data may be subjected to further use (e.g., the use of search engines). In the social sciences, the distinction between ***found data*** and ***designed data*** is also common (Biemer & Amaya, 2020). ***Found data*** are data that were not collected purposefully for a scientific study but were found instead (e.g., archival data), whereas survey data are an example of ***designed data*** because they were designed and collected for scientific purposes. Found data are often unstructured data, which must be prepared first before they can be analysed statistically.

Unstructured data are typically data that were created in people's ***natural environments*** (naturalistic data, fieldwork) rather than in a laboratory. However, they are not identical to naturalistic data because they can also refer to structured data (e.g., capturing moods using *Ambulatory Assessment*).

1.2 Relevance of unstructured data

There are various reasons why unstructured data are of great importance to the social, behavioural, and economic sciences. First, unstructured data are generated in many life areas; they depict many important parts of human life that could not be depicted in the same way using structured data. A large portion of these data are not yet being analysed (Eberendu, 2016).

As laid out in the introductory section, unstructured data are often data that are collected or created non-reactively in people's natural environments. This lack of reactivity helps prevent biases that may occur when data are collected in a reactive or laboratory setting. By the same token, naturally occurring phenomena can be captured with high ecological validity. They are potentially able to depict the reality of natural environments, social groups, and organisations to a high degree.

As unstructured data often mirror concrete behaviour, they partly circumvent problems linked to self-reporting in classic survey studies (e.g., response styles, social desirability, self-deception and the deception of others) (Borkenau, 2006). Unstructured data can therefore be useful in complementing traditional survey studies by offering a means for validating survey data (Jürgens et al., 2020), enriching survey data with other data sources (e.g., behavioural data) to explain certain social phenomena, and thus increasing the explanatory power of survey studies (Reveilhac et al., 2022).

1.3 Goals and addressees of this output paper

Unstructured data open up new perspectives for social, behavioural, and economic research. However, the use of unstructured data is also linked to specific challenges, which arise precisely because the data are not typically collected as part of a controlled scientific study. We illuminate these particular challenges in the following and – as much as possible – formulate recommendations on how to tackle these challenges.

This output paper is aimed at students and researchers in the social, behavioural, and economic sciences on the one hand and everyone who works with unstructured data and draws inferences from them for practical applications on the other.

1.4 Short report on the survey and the workshop

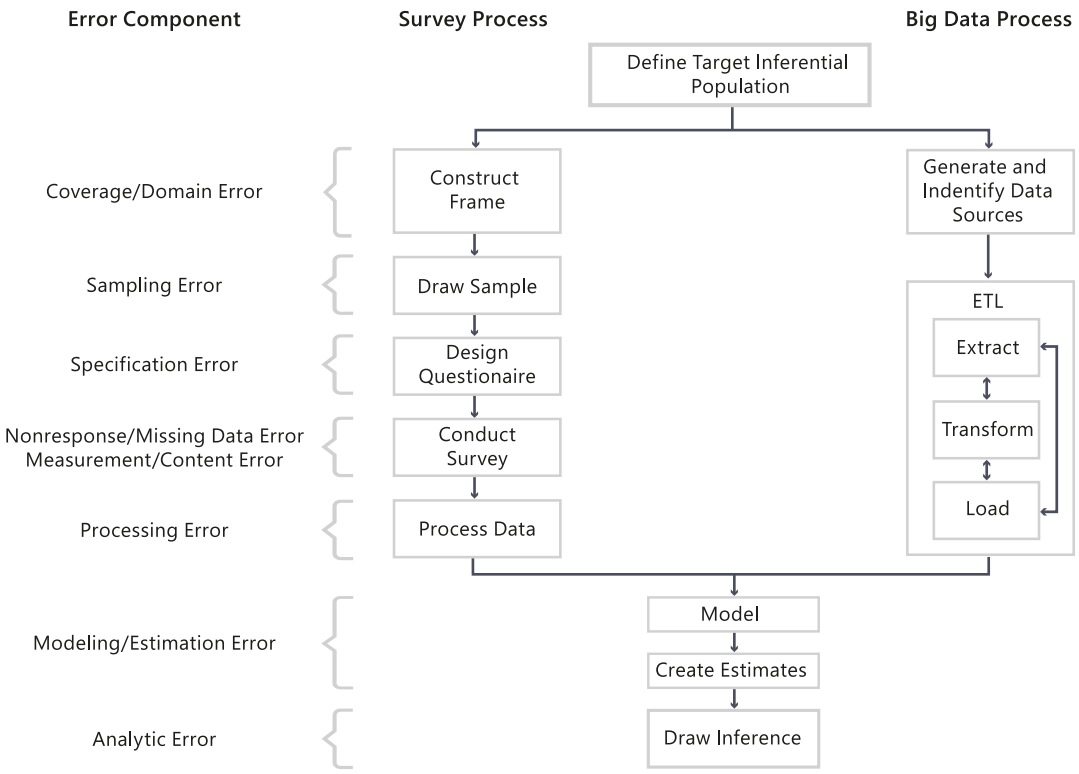
A questionnaire on the quality assurance of unstructured data was developed on the basis of the *Total Error Framework* (TEF) for *big data* (Amaya et al., 2020) as part of the creation of this output paper. It consisted of a total of 32 questions, and, in addition to more general technical information about the participants, it included questions about the generation, preparation, and analysis of unstructured data, as well as a final question in an open format. Nineteen scientists from different relevant areas of science (business administration, educational research, computational social science, communication studies, political science, psychology, sociology, and economics) answered the questionnaire. These scientists reported using a broad spectrum of various types of data for their research: Social media data (e.g., Twitter, Facebook), smartphone data, audio data, traditional media (e.g., newspapers), and text documents (e.g., parliamentary speeches, company agreements, accounting).

The main results of the survey were discussed with the scientists at a two-day online workshop on 13 and 14 October 2021. The workshop was divided into four thematic sessions: Data generation, post-collection processing, data analysis, and open questions about the availability of new types of data. The output paper at hand summarises the results of the survey as well as the discussion and embeds them into the TEF.

1.5 Brief introduction to total error frameworks for evaluating data quality

Many different versions of total error framework approaches that focus on different kinds of data are currently being created. For the purpose of discussing the quality assurance of unstructured data, we present ***Amaya et al.'s (2020) TEF*** as a concrete example. The TEF extends the error framework principle to the context of big data and builds on the *total survey error (TSE) framework* (Groves & Lyberg, 2010) that is commonplace in survey research. The *TSE framework* helps summarise all sources of error that have a possible influence on the results of an empirical, survey-based study. In total, the TEF considers eight error sources that aim to cover the entire process of data generation, post-collection processing, and data analysis. Figure 2 presents the TEF (Amaya et al., 2020). The boxes on the left side (survey process) highlight steps that are specific to survey research, whereas the ones on the right side (big data process) highlight the steps that are specific to handling *big data*. The middle part is common to both approaches and independent of the data source or data structure. The TEF offers a useful way to initially structure research for discussing methodological challenges when using unstructured data for scientific purposes. The report at hand is divided into the chapters Data generation, Post-collection processing, and Data analysis and is guided by the challenges addressed in Figure 2.

Figure 2: Total Error Framework (TEF)



Source: Amaya et al. (2020)

The **Data generation** chapter addresses definitions of units of analysis and the closely related *coverage error*, which arises because the target population and survey population are not identical. Moreover, classic sampling errors are addressed and distinguished from errors that are caused by non-response at the unit level (*unit non-response*) or at the item level (*item non-response*). The **Post-collection processing** chapter looks at specification error and thus examines possible deviations between the concept that was meant to be captured by the research question and the concept that was measured empirically. These deviations include limitations relating to the available data as well as distortions that can be the result of technical errors or the individual settings of devices. In the social sciences and psychometrics, these aspects are grouped under terms such as the validity and reliability of measurements (Eid & Schmidt, 2014). The **Data analysis** chapter looks at *processing errors* and thus addresses possible biases resulting from the input, transformation, and coding of variables. Processing errors can also result from linking different data sources for one unit of analysis (*record linkage*). Other sources of errors that occur in data analysis include the treatment of missing data in statistical analyses (*modelling error*) and biases that result from an incorrect application of statistical models and interpretation of results (analytical errors). According to the TEF, the special challenges that come with working with unstructured data can be found in the areas of **data generation** and **post-collection processing** (see Figure 2), whereas possible sources of error in data analysis are similar in nature to those in traditional survey research.

In conclusion, the report at hand should be seen as a mere first step towards a common understanding of quality standards for working with unstructured data. The primary aim is to carve out particular challenges that arise from working with them as compared with working with data from traditional survey research. To identify these challenges, the TEF offers useful guidance in a rapidly developing research field. The explicit formulation of standards should be done in view of concrete types of data. The basic idea of *error frameworks* is currently being further differentiated and applied to various types of data (e.g., Twitter data, see Hsieh & Murphy, 2017; data from online platforms, see Sen et al., 2021; *metred data*, see Bosch & Revilla, 2022).

2 Data generation

This chapter carves out the specifics of ***data generation*** and presents both special problem areas and challenges and recommendations for handling unstructured data during the generation process. First, we discuss the definition of the units of analysis, then *coverage error* and *sampling error*, and, finally, *non-response* and *missing data error* with regard to unstructured data.

2.1 Definition of units of analysis and data structure

Very different units of analysis can be chosen when working with unstructured data. Even when looking only at social media data, the unit of analysis could be defined as the individual, the account, the content of a post (e.g., the post itself or a comment), or the interaction (e.g., between followers in a network). Furthermore, entire texts in archives could be chosen as units of analysis, or individual sequences in visual or audio material. The definition of the unit of analysis has ***far-reaching consequences*** for the classification of various sources of error, especially when working with unstructured data. *Coverage error* and *sampling error*, which we discuss in the following section, are measured by defining the unit of analysis and the population (e.g., people in general vs. users/accounts).

Data often have a ***multi-level structure***. For example, tweets are clustered in accounts (Fischer et al., 2019), observation times in individuals, and individuals in spatial aggregates. Depending on the definition of the units of analysis at each respective level, error-free information about the structure of the clustering is not always available. It may be possible, for example, to correctly assign tweets to accounts but not necessarily to individuals. One person can use multiple accounts (e.g., a teenager may have different social media accounts for friends and parents/family), or multiple individuals can operate one Facebook account (e.g., in a business context). This partially opaque data structure, or improper assignment of units of analysis at Level 1 to the superordinate Level 2, will have adverse effects, at the latest, when choosing adequate modelling, which should take into account the violation of the assumption of independence of cases (by applying, e.g., multi-level or panel data analysis or robust standard errors) (see also Chapter 4.2).

2.2 Coverage error and sampling error

The *TSE framework* clearly distinguishes between *coverage error* and *sampling error*. ***Coverage error*** refers to differences between the statistical population to which the results of the research are to be generalised, and the sampling plan, the list, or the procedure for how the sample elements should be drawn. ***Sampling error*** refers to differences between the sampling plan and the sample drawn on the basis of the plan. When such differences are random, they are considered sampling error, or they can be systematic. Both the expert survey and the workshop that were conducted as part of this paper suggested that, in many cases, this distinction between the sampling plan and the actual sampling could not be made for unstructured data. This inability to make a distinction can have different causes: In many cases, the sampling plan and the actual sampling are identical because – unlike in surveys – there is no economic necessity to limit the number of units of analysis by using a sample. At the same time, there are cases in which data providers deploy samples directly through an *application programming interface* (API), leaving researchers without access to the sampling plan and putting the plan beyond separate investigation. For this reason, we jointly investigated selections made at the level of the sampling plan and the actual sampling in the following.

A distinction should be made between studies that can extract data from platforms ***without the consent*** of the participants (e.g., the people who created accounts) (Type A) and studies that can collect data only ***with the consent*** of the participants and, possibly, with extensive participation (e.g., installing a smartphone app) or the use of their own hardware (e.g., wearables) (Type B).

For Type A (studies ***without*** the consent of participants), we identified the following ***problems and challenges***:

1. ***Adjusting the population to match data availability ('availability research'):*** The problem that the users of certain services (e.g., Facebook, Twitter) do not represent the entire population (e.g., of a country) can be solved, in principle, only by declaring that the users for whom data are available are the statistical population. When doing so, however, no inferences should be drawn about larger populations without addressing the selection problem.
2. ***Opaque algorithms:*** Data providers use proprietary and opaque algorithms to select samples, and such algorithms are made available to researchers through APIs. These algorithms can affect both selection and sorting. For example, the free-of-charge 1% sample of Twitter data differed from the purchased 10% sample (Morstatter et al., 2013). Researchers who want to work with the provider's data often do not have a choice but to use data selected by opaque algorithms.
3. ***Paywalls:*** The fact that many text documents are not openly available and therefore cannot be used by researchers in many cases plays a significant role, especially in the analysis of text documents. It is safe to assume that the features of these documents differ from the openly available ones.
4. ***Personalised Content:*** The content displayed (e.g., when visiting a website) may depend on the characteristics and previous behaviour of the users and thus cannot be determined objectively. This problem has a strong effect on the replicability of research.
5. ***Deletion of content:*** Content or accounts on platforms are frequently deleted if, for example, they violate a platform's terms and conditions, or other users request it. The population of all 'posts' made on a platform may already be unavailable after a very short time. The longer the gap between when the post was made and the extraction of the sample, the more the sample is distorted in favour of rule-compliant contributions. Such changes can be a massive problem if rule violations are the subject of research. In other cases, technological change can also be a cause for the unavailability of content (e.g., outdated data formats). This problem, too, has a large effect on the replicability of research.
6. ***Overcoverage through bots (automated accounts):*** A considerable part of accounts and posts on online platforms can be traced back to automated accounts, so-called bots, and can therefore not be included in the population of most research endeavours. If bots are not recognised as such, results can be significantly biased.
7. ***Selection error in samples:*** If samples are taken from the total stock of a platform for research purposes, for example, by selecting posts containing certain keywords, selection error at the level of posts as well as at the level of user groups can result. This kind of selection error can occur, for example, when certain age groups have a preference for a certain terminology or when ethnic groups use different languages (Sen et al., 2021).
8. ***Data protection:*** Restrictions based on data protection and research ethics can also be a reason why selective access to part of a data stock is not possible. Additionally, data protection policies often prevent researchers from sharing the data used in a publication, thus limiting their reproducibility.
9. ***Duplication:*** Duplicates of content are commonplace. However, software is usually able to easily identify only exact duplicates.
10. ***Uncertainty regarding affiliation with a population:*** Researchers might lack the information needed to judge whether a case belongs to a population (e.g., creation date and location of a document).

For Type B (studies ***with*** the consent of participants), we identified the following ***problems and challenges***:

1. ***Missing sampling plans:*** Due to a lack of sampling plans (e.g., for users of a certain technology), arbitrary or deliberate selection or snowball sampling is often used. In such samples, biases are quantifiable only for items that (a) were surveyed during the study and (b) are available in an unbiased form for the population from other sources.
2. ***Difficulty in distinguishing sampling error from non-response error:*** With this type of sample, sampling error cannot be neatly distinguished from non-response error because there is no sample available beyond the individuals who take part voluntarily. It is often only a very small portion of the population that consents to this type of participation.

3. ***Matching individuals to devices:*** Devices can be used by multiple individuals (*clustering*), or individuals may own several devices (*duplication*). If this lack of one-to-one matching is not surveyed and taken into account in the analyses, biases can result.
4. ***Commercial access panels:*** Commercial access panels are now sometimes used for data collection and recruiting, and such panels are also not based on samples with known or estimable errors.

2.3 Non-response/missing data error

Whereas it is relatively easy to make a neat distinction between ***unit non-response*** (a respondent fails to take part in a survey at all) and ***item non-response*** (a respondent does not answer a certain question) in survey research, this distinction is much harder to make for unstructured data. There are many causes of missing data in unstructured data as the following examples illustrate:

1. ***Later deletions*** of user-generated content can be responsible for missing data as well as hidden content or blocked users.
2. ***Privacy concerns*** can lead to systematic ***unit non-response***, especially in surveys that require the respondent's consent but also more generally for social media use.
3. ***Technical problems*** can be the cause of missing records as well as *dropouts*.
4. With textual data, ***changing the character encoding*** can cause problems.
5. ***Systemic failures*** that are correlated with variables that were supposed to be measured (e.g., deviant behaviour) are a particularly serious problem. Respondents may decide to switch off tracking for specific behaviours, thus temporarily interrupting data collection.
6. The ***cause of missing data*** is then often unclear. When data are generated with smartphones and wearables, it might not be possible to distinguish whether the device was forgotten and is possibly recording the sounds of a different environment, whether the respondent is asleep, and the device is not recording because of that, or whether there is a technical problem.

Moreover, ***non-response*** in unstructured data is often ***confounded with undercoverage***. For example, when analysing Twitter data over a set period of time, categorisation could differ depending on the definition of the unit of analysis (tweets vs. the individual): When the unit of analysis is the individual, undercoverage could be the result of a missing Twitter account or non-response in the sense of tweets that do not occur during that period. If the unit of analysis is a tweet, all missing tweets can be considered undercoverage (Amaya et al., 2020).

2.4 Recommendations

Due to the dynamic development and variety of types of data, developing standards to handle most of the abovementioned problems in data generation is a difficult task. This professionalisation runs in parallel or with a time lag across disciplines. In our workshop, we deduced the following ***recommendations***:

1. Procedures/software for ***identifying bots*** should be deployed when using platform data. There is an urgent need for further research on the quality of existing procedures (Rauchfleisch & Kaiser, 2020) and on improving detection.
2. The ***limitations*** of studies with unstructured data, especially ***regarding certain selectivities***, should be extensively discussed in publications.
3. If very large samples or even the entire population are available for analyses, we recommend using ***effect size measures*** (and their confidence intervals) in addition to or instead of the reported statistical significance of the results.
4. We recommend ***power analyses*** to determine sample size, especially when the costs per unit of analysis are high.
5. ***Checking for duplicates*** should take place when analysing documents. ***Fuzzy String Matching*** can be applied to detect documents that are not identical (Leskovec et al., 2020).
6. A promising approach for facilitating the ***replicability*** of results on the basis of privacy-sensitive data despite the requirements of data protection is the ***differential privacy approach*** (Dwork et al., 2006;

Dwork & Roth, 2014). This approach has already been applied, for example, by Evans & King (2022) to a Facebook URL dataset (Evans & King, 2022; see also the possibility of *Remote Execution Solutions* by van Atteveldt et al., 2021).

7. There are studies on mobile devices that require that participants using smartphones or wearables actively participate. If it is not possible to use random samples for data availability, data protection, or economic reasons, we recommend that such studies take measures to increase the ***generalisability of research results***. The following measures could be considered:
   a. Recruiting from and ***linking with existing studies*** (e.g., longitudinal studies) (Kreuter et al., 2020) to facilitate a comprehensive analysis of selectivity. If the respective questions are available, a distinction can also be made between ***non-coverage*** (who does not have the appropriate devices) and ***non-response*** (who does not participate despite having the device) (Keusch et al., 2020; Keusch et al., 2022). Moreover, the questionnaire can help clarify issues of jointly used devices or multiple devices per participating person.
   b. For hidden populations, the application of, for example, ***respondent-driven sampling*** (Heckathorn, 1997) facilitates estimates of the probability of inclusion.
   c. ***Effective weighting*** can be achieved by ***collecting*** participant information that is as meaningful as possible and is ***correlated*** with target variables (not merely demographic information) that are also known for the population (e.g., by using official statistics).
8. Technical causes for missing values can be limited, for example, by using ***high-frequency monitoring***. It is possible to validate missing values by ***comparing different sensor data***.
9. ***Coverage bias and non-response bias*** should, in principle and not just for unstructured data, be ***evaluated together*** (Eckman & Kreuter, 2017).
10. The ***causes of sources of error*** should be ***named and categorised***, the categorisation should be made transparent, and the resulting errors should be quantified and compared with other data sources (Amaya et al., 2020).

3 Post-collection processing

In classic survey research, instruments are explicitly developed to capture constructs. These are often items or questionnaires (see ***Design Questionnaire*** in Figure 2) that are formulated or selected, driven by theory, in relation to the construct that is to be measured. However, research with unstructured data draws on data that already exist but were not generated explicitly for research purposes. This creates special challenges for assessing the quality of measurements, which we address in the following chapter.

3.1 Specification error and validity

Specification error occurs when the construct that the research question refers to is not aligned with the construct represented by the data (Amaya et al., 2020). This part of the TEF thus refers to construct validity. ***Construct validity*** is given when the inferences drawn about the underlying construct on the basis of the data are adequate and appropriate (Eid & Schmidt, 2014; Messick, 1989, 1995). Ensuring construct validity in social, behavioural, and economic research usually begins as early as when tests or questionnaires are being constructed. Typically, the construct is first defined in theoretical terms. Then, on the basis of theoretical deliberations, items that are designed to cover the construct in a valid way are formulated or selected (Eid & Schmidt, 2014). As part of extensive investigations of validity, researchers then empirically test whether the data, which are collected using the constructed measurement instrument, follow theoretical expectations. Due to its primary recourse to already existing data, empirical research using unstructured data often differs from this prototypical approach in social, behavioural, and economic research:

1. ***Construct-related data selection and construct adaptation:*** Assuming that a researcher wants to study an a priori construct, they are faced with the problem of selecting and processing unstructured data in such a way that the data provide the best possible match to the a priori construct. Depending on the data situation, this matching can be more or less successful, and it may require the original construct of interest to be adjusted (Sen et al., 2021).
2. ***Determining the construct through the data:*** Research using unstructured data can also be designed in such a way that it does not begin with an a priori construct. Instead, it may investigate available data in an exploratory manner, seeking out interesting concepts and constructs that could be researched on the basis of the available data (Sen et al., 2021). The research questions and constructs of interest are thus inferred from the data and determined by the available data.
3. ***Defining new constructs:*** Working with unstructured data could also introduce entirely new constructs into the scientific community because new collection methods may also result in the definition and establishment of new constructs (e.g., analogous to the distinction between explicit and implicit attitudes in social psychology by considering reaction times).

This use of unstructured data implicates ***specific problems and challenges*** to securing construct validity and the manner in which validation strategies are chosen:

1. ***Lack of validity studies:*** Ideally, an investigation of the construct validity of a questionnaire or other measurement instrument is undertaken as part of an independent research programme in which, driven by theory, hypotheses are tested in relation to the behaviour of the measurement instrument (Eid & Schmidt, 2014). For unstructured data, there are no such comprehensive validity studies. The validity of specific inferences may therefore be compromised.
2. ***Lack of gold standard in convergent validity:*** Concerning convergent validity, it is possible to examine whether measures of constructs obtained with unstructured data are theoretically related to other measures of constructs that were obtained using other methods of data collection or other types of data. Survey data, interview data, and ethnographic data, for example, may be used to validate conclusions that were made on the basis of unstructured data (Reveilhac et al., 2022; Tufekci, 2014). However, there will not always be a gold standard measurement of a construct available to which the

unstructured data can be related. Further, by using unstructured data, researchers are aiming to avoid the kinds of problems that occur with other methods of data collection (e.g., self-reporting) and arrive at more valid conclusions. Lastly, there is also the question of whether the constructs captured using different methods really represent the same construct or different constructs.

3. ***Securing content validity:*** When taking texts, images, videos, or other data that have been found and using them to make inferences about constructs (e.g., the attitudes of the individuals under investigation), one question that arises is how representative the data are of the construct of interest (e.g., the attitudes). Typically, this representativeness can be verified only by using additional data and sources of information.
4. ***Analysing discriminant and incremental validity:*** When the concepts and constructs under investigation are adjusted to the available data or when defining new data-driven constructs, a question that arises is how they are related to established constructs. It is necessary to ask whether they are sufficiently distinct from established constructs (discriminant validity) or whether they contribute to better predictions or explanations of phenomena by going beyond established constructs (incremental validity). The analysis of these facets of validity also often requires additional data to be included.

3.2 Measurement errors and errors in content

The precision of a measurement instrument in the social sciences is measured by its reliability. Reliability is defined as the proportion of the variance of true scores to the total variance of observed scores, where the total variance is made up of the variance of true scores and the unsystematic variance. Reliability refers to the extent to which observed differences between individuals can be attributed to true (measurement-error-free) differences (Eid & Schmidt, 2014; Schnell et al., 2011). There are various methods that can be used to determine reliability. Such methods focus on either the consistency of measurements within a measurement instrument (e.g., split-half method, internal consistency, parallel testing method) or the consistency of measurements that use the same measurement instrument over time (e.g., test-retest method). High reliability indicates that the results of repeated measures of the same feature are highly correlated (i.e., the measurements are highly correlated) and the measurements are overlaid by unsystematic variance to only a limited extent.

Poor reliability of measurement instruments affects further analyses because poor reliability can lead to biased estimates of correlations between features. In principle, there are three strategies for dealing with measurement error in the social sciences. Firstly, if reliability is poor, researchers can try to optimise measurement instruments (e.g., by excluding items or developing new items). Secondly, measurement error can be taken into account in further analyses when interpreting results (e.g., as part of sensitivity analyses). Thirdly, some methods explicitly seek to correct the effect of erroneous measurements when estimating correlations (e.g., structural equation modelling).

When working with unstructured data, it is often necessary to deal with potentially erroneous data, which can result in various ***problems and challenges:***

1. ***Found data:*** Unstructured data are often available in the form of found data. As such, the conditions under which they were created remain unclear to a certain extent. For example, it is not always possible to determine exactly whether different digital traces were left by the same person or which concrete technical settings were made by a platform. This lack of clarity may obstruct the replication of measurements as well as the specification of measurement models, which, in the traditional approach, are needed to estimate reliability.
2. ***Individual collection devices:*** Studies using unstructured data often employ measurement devices (e.g., sensors or trackers) or draw on available data that depend on an individual's settings (e.g., of a smartphone or a platform). The question that therefore arises is which strategies for due diligence are available to identify and correct possible biases in data collection and post-collection processing.
3. ***Automatically created data:*** Data can be generated automatically (e.g., by bot accounts on social media or automated content in web tracking). By including such data in addition to naturally collected data, reliability analyses can become distorted.

3.3 Recommendations

Overall, data quality issues play an important role in research with unstructured data, and a considerable amount of time must be invested into preparing and checking such data. However, even though advanced statistical methods are being used to handle such data, little to no general standards have yet been established for assessing the quality of measurements. This lack of standards is partly due to the fact that research teams are working with very different apps and devices and that platforms and apps are constantly changing. There is also often a lack of systematic validation studies. We identified the following starting points for future research on the validity and reliability of unstructured data:

1. ***Resorting to established validation strategies:*** Studies for testing the validity of unstructured data can resort to established procedures from social, economic, and behavioural research (see, e.g., Eid & Schmidt, 2014; Krippendorff, 2008; Lamnek & Krell, 2016; Schnell et al., 2011) and can apply these procedures to unstructured data:
    a. ***Convergent validity*** can be examined by drawing from additional data sources, such as survey data, interview data, or ethnographic data (Reveilhac et al., 2022; Tufekci, 2014). If construct measures are obtained with algorithms, external data or specially generated data can be used to test for whether the algorithm results in similar construct measures.
    b. As part of ***criterion validation***, a researcher can analyse whether a construct that is measured with unstructured data can predict a criterion in the expected way. In addition, a researcher can analyse whether measures of constructs obtained with unstructured data contribute to the prediction of a criterion (e.g., specific behaviour or experience) beyond other previously available measures of the same construct and thus represent added value (***incremental validity***).
    c. Regarding ***content validity***, researchers can analyse the extent to which the selected sample of data is representative of the construct under investigation. Using the topic modelling procedure (Heyer et al., 2018) based on database metadata, for example, researchers can investigate whether the found distribution of words corresponds to the distribution in other databases. It may also be useful, for instance, to examine the extents to which the life situations that the text, audio, or images were collected in are representative of the lives of the individuals under investigation. Such comparative data can be obtained through comprehensive panel studies that deal explicitly with human behaviour in the digital world (Tufekci, 2014). A compliance questionnaire can be used to survey the times at which or the situations in which the participants did not wear the recording device. Such information can help to highlight possible biases in the recording.
    d. When evaluating the content of unstructured data, methods of ***semantic validation*** can be used (Krippendorff, 2012) to examine whether the meaning of the analysed text was correctly represented.
    e. If unstructured data are analysed using qualitative methods, it can be helpful to take into account ***validation strategies*** of qualitative research (Lamnek & Krell, 2016).
2. ***Lack of information on validity:*** If it is not possible to present convincing evidence of construct validity, the validity of the inferences drawn from the data to the underlying construct is not ensured. In these cases,
    a. it is advisable not to draw these conclusions but to make interpretations only in relation to the available data (i.e., 'close to the data');
    b. it may be useful to revise the research question or switch to a different data source (and transparently communicate what was done);
    c. it may be necessary to refrain from using or publishing the data altogether.
3. ***Data-driven constr uct definition:*** If the definition of constructs is driven by the data, it is necessary to ensure high transparency regarding the relationships between the construct and the data points. Example data can be used to illustrate the ways in which the data are related to the construct of interest. Alternatively, a more theory-driven approach is conceivable in which concrete expectations about the relationships between the concept and the available, or collected, data are formulated

(auxiliary theories; Schnell et al., 2011). Such an approach would enable theory-driven testing of validity and reliability and result in better integration into the theory of classic measurement models.

4. ***Reflecting on the effect of measurement error:*** The extent and possible effect of measurement error should be communicated and taken into account when interpreting the results. Data points that are obviously erroneous should be excluded. There are several strategies for checking data quality (measurement quality, measurement error influences), which can differ depending on the type of unstructured data.
    a. ***Smartphone data:*** The ***correctness of the recorded data*** can be checked with the help of standardised action sequences:
        i. For example, a certain ***protocol*** can be specified (open the app, make a call, etc.) and then compared with the recorded data. Moreover, certain smartphone settings could be ***standardised***.
        ii. ***Identification and careful examination of values that lie outside the expected range:*** For example, these values can be variations in heart rate or very rapid changes in location (several kilometres in a few seconds) in the Global Positioning System (GPS).
        iii. It is possible to partially ***record measurement inaccuracies*** and to take them into account for further analyses. When measuring coordinates (longitude and latitude), for example, the precision of the measurement can also be recorded.
    b. ***Textual data:*** Different strategies can be recommended for textual and content analysis:
        i. ***Pre-processing of text:*** Scripts for preparing text and the deployed algorithms can be made available in replication scripts, and plausibility checks can be carried out.
        ii. ***Correction of measurement error in content analyses:*** In content analyses of social media data (e.g., Facebook data), the extent to which coders' observations agree can be used to determine the precision of the content coding (Bachl & Scharkow, 2017). Crowdsourcing approaches can also be used, for example, to code random text samples in order to be able to examine convergence across coders.
        iii. ***Robustness of text analysis algorithms:*** The parameters of the employed text algorithms may vary, and researchers can investigate how sensitive the main results are to this variation. For example, it might be possible to vary the threshold for semantic or structural similarity in text analysis.
5. More ***basic methodological research*** should be conducted to develop procedures to test for validity and reliability and to develop criteria for the publication of data quality checks. One challenge will be to develop procedures that can be generalised to different apps and platforms.

4 Data analysis

In contrast to data generation and post-collection processing, the analysis of unstructured data in principle consists of the same steps – and thus the same sources of error – as traditional survey research (see Figure 2). A large part of the data analysis recommendations from the methodological literature on research using unstructured data can therefore be directly applied to working with unstructured data (Cohen et al., 2003; Shadish et al., 2002). The chapter at hand focusses on specific challenges in data analysis that the experts who took part in our survey or during the subsequent discussion in the workshop viewed as characteristic of unstructured data.

4.1 Record linkage and processing error

Unstructured data are often combined with other sources of data (***record linkage***) so that additional analyses can be applied to the data. For example, web tracking and social media data are often linked to traditional survey data. Other examples of record linkage include linking TV video content with content analysis data (linked through channels, dates, times), social media data with website content (linked through URLs), or behavioural data with media data (linked through tracking data – URLs) (see Stier et al., 2020). The same standards apply to this linkage as to structured data formats (Christen, 2012; Tokle & Bender, 2021). The aim must be to minimise possible ***processing error*** (e.g., no clear assignment of individuals) that can arise from linking different data sources.

Furthermore, a special challenge to data analysis is that unstructured data are predominantly collected for ***purposes other than research***, and therefore – to render such data useable in research – they must be transformed into research data at great expense. To do so, an understanding of what these data represent and how they were measured must be gained before each instance of analysis of unstructured data. Sometimes the exact meaning or the context of data collection/generation of unstructured data is not clear for certain variables when they are viewed 'from a distance'. For example, there is often no demographic information available, and thus, key reference information might be missing (Stier et al., 2020). Moreover, ***different devices*** from different manufacturers (e.g., for recordings) can result in disparities in the data because recordings may differ by the type of device. Individual processing steps (e.g., pre-processing of textual data) must therefore be documented and made transparent. When working with textual data, it is possible to encounter ***encoding problems*** or differences in data due to differences in or a different order of clean-up or transformation steps. Rather than viewing them as programming errors, it underscores the importance of documenting the programming steps in detail. Overall, it was noted that, although individual best practice recommendations for programming exist, pre-processing and the corresponding processing steps also depend on the concrete research question.

Another special feature of unstructured data is that ***system-generated data*** are created in addition to user-generated data. It is therefore possible that system-generated data will be mistakenly processed as user-generated data.

4.2 Modelling error

Another possible source of error lies in the dependency between the results of the data analysis and the choice of statistical modelling (see Figure 2). ***Modelling error*** includes both the possibility that the statistical models were applied incorrectly and the fact that results/conclusions can vary significantly depending on the choice of statistical modelling.

Incorrect application of statistical models can include failing to choose the appropriate analysis model (e.g., choosing the wrong method to calculate the standard error) (West et al., 2016) but also mistakes that might occur accidentally in the specification of the analysis model. A prerequisite for identifying and correcting modelling errors is the reproducibility of the results of an analysis (*computational reproducibility*) (Stodden et al., 2018). Computational reproducibility means that the results can be reproduced by

applying the analysis code (e.g., a syntax file for the statistical software that was used) to the data used for the analysis (Christensen, 2018; Hardwicke et al., 2021).

Additionally, there is the issue of the robustness of the ***results of the analysis***. Results are generally viewed as less robust if they are sensitive to the choice of alternative analysis models (Simonsohn et al., 2020; Young & Holsteen, 2017). For example, key analysis results may depend on which covariates are included in the model and how the effects of the covariates are specified (e.g., linear vs. non-linear effects). Robustness issues in analysis results can be best addressed if both the data and the analysis code are accessible to the scientific community.

4.3 Analytical error

Finally, a key source of error lies in the inferences drawn on the basis of the data analysis (***analytical error***). Here, it is important to emphasise that, also when working with unstructured data, the robustness of inferences depends on the research design (Salganik, 2018). Here, too, the royal road towards robust causal inferences is paved with randomised, longitudinal experiments or field studies (Shadish et al., 2002). If randomisation is not possible, the effects of possible confounders can be controlled by taking covariates into account or by cleverly choosing study designs (e.g., natural experiments or quasi-experimental designs) (Angrist & Pischke, 2009). Alternatively, it is also possible to refrain from answering causal analytical questions and to focus on describing interrelationships.

A special aspect of analysing unstructured data is the ***test fairness*** of the statistical algorithms used for analysis. Test fairness refers to systematic discrimination against certain individuals on the basis of their ethnic, sociocultural, or gender-specific group background. This discrimination can be the result of, for example, the use of biased training data (*sample bias*) for classification algorithms (Rodolfa et al., 2021). In computer science, the verification of fairness in flexible machine-learning procedures is gaining momentum (e.g., algorithmic bias/fairness in *artificial intelligence*). There was consensus among the participants of the expert workshop we conducted that the verification of model fairness should also play a large role in social science applications, especially if the aim is to apply the procedures in practice.

4.4 Recommendations

Making unstructured data useable for analyses is linked to a great deal of effort that is not always visible to third parties. Generally speaking, the ***pre-processing*** of data plays a substantial role in any type of subsequent statistical analysis. The following recommendations were deduced from the results of our workshop:

1. ***Documentation of processing steps:*** When transforming unstructured data into research data, every pre-processing step should be documented and examined for possible biasing effects. Such documentation and examination can sometimes be difficult or even impossible because *pre-processing* on a platform or an app does not provide researchers with insights into the processes, as these are proprietary procedures. Moreover, the version of the software can change and may therefore produce deviating results. Such problems can be addressed by using solutions for version or dependency management (e.g., 'packrat' in R).
2. ***Programming:*** Erroneous data can also be the result of simple programming errors that are not subsequently found due to a lack of plausibility checks. It would therefore be sensible to adopt the error-reducing strategies from professional programmers. Examples include pair programming, peer review of code, and, of course, replication, transparency, and meta-analysis. However, efficient post-collection processing often requires, for example, in-depth knowledge of object-oriented programming. Most social scientists do not have such knowledge, and it is difficult to recruit software engineers with comprehensive training to work in research.
3. ***System vs. user-generated data:*** To distinguish system-generated from user-generated data, the following approaches may be useful:
    a. ***Triangulation:*** Subsequent qualitative interviews can help determine the extent to which the study's participants were aware of their behaviour in certain situations (e.g., during eye-tracking: Is a person at all aware that their eyes were resting on an object and, if so, why they were resting there?). Such interviews can supply additional information to help validate the data.

b. ***Integration of methods:*** In text analysis procedures, there is the possibility of expert validation. Subject-specific unstructured data (e.g., on the political situation of a certain region) could be underpinned by professional expertise. This type of method integration would be especially desirable from the point of view of qualitative researchers.

c. ***Combination of methods:*** When methods are combined, the data are enriched with additional knowledge from classic text analyses. In communication studies, for example, discourse analyses are carried out on controversial topics (e.g., internet regulation). In doing so, large amounts of data can be registered by using computer-aided procedures. It is nevertheless important to extract small samples afterwards to try to understand the broader patterns.

4. ***Transparency of analyses:*** For the sake of both the reproducibility and the robustness of the analyses, the analysis code and data should be made available to the scientific community. However, this availability will not always be possible for all datasets for data protection reasons. When data cannot be made available, it is important to think about alternatives (e.g., synthetic data, or limited data access; van Atteveldt et al., 2021).

5 Outlook: Open questions and challenges to research with unstructured data

The scientific analysis of new, unstructured types of data results in additional issues that go beyond the specifics of the research process. Which new challenges arise in terms of data access? How transparent is the selection process? How must governance structures be designed, and which resources do researchers require for research using unstructured data?

5.1 Data access

Unstructured data encompass a broad spectrum of types of data. They range from newspaper articles to images and videos to sensor data. Moreover, they are often generated by very different, often non-academic, global companies that work for profit. Such companies are rarely interested in publicly disclosing the existence of these data to independent researchers. Additionally, both the structure of and access to data change continuously at a technical and legal level. This change creates major technical, legal, and procedural challenges for researchers (Breuer et al., 2020). The sheer quantity and heterogeneity in terms of time and content impede the development of clear recommendations (or textbooks) that remain valid in the long term.

Whereas this situation is problematic, it is also very similar across disciplines in social, behavioural, and economic research. Coupled with rapid changes in the structure of and access to data, the result is that ***opportunities for interdisciplinary networking*** and an institutionalised, up-to-date exchange on concrete research projects are the best-possible responses to the rapidly evolving field of the analysis of unstructured data. Individual disciplines, such as journalism and communication studies, where an exchange is already taking place, especially among researchers working with social media, are pioneers in this regard. Moreover, communication researchers are involved in various initiatives within their academic associations that are developing recommendations on these matters. The aim should be to initiate and consolidate an ongoing exchange of this kind for all the academic associations represented in the German Data Forum (RatSWD) in order to help researchers in other disciplines work with unstructured data, too. Moreover, ***interdisciplinary conferences focussing on methods and data*** would be useful for an exchange on issues, such as data sharing and the re-use of unstructured data.

Another challenge is to create a secure legal framework for research using data from private platforms. Universities should provide ***support staff for data management*** and provide advice on ***issues of copyright and data protection***.

Finally, it is necessary to change the ***incentive system for researchers*** in order to reduce the high costs that are often associated with the sharing and documenting of unstructured data. Whereas such sharing and documenting are already commonplace in disciplines such as psychology, this tradition has not yet become established in other disciplines, for example, in business administration. The topic still does not get enough attention in academic associations or during the training and qualification phases. However, the requirements of research sponsors and publications are now creating new incentives in this regard. Journals (e.g., in political science) are also increasingly transforming themselves and are creating incentives to showcase new datasets and to briefly describe their potential to be analysed, for example, through the research note format. Such a format, which increases the citability and reputation of datasets, would also be desirable for datasets containing unstructured data. However, again, these new developments underpin the need to create points of contact that help researchers find out whether these data can legally be published at all, especially unstructured data.

5.2 Transparency

Because unstructured data are often generated and made available by non-academic digital companies, typical criteria for transparency are difficult to meet. Researchers gain access to data through unknown application interfaces without knowing what the populations look like or which selection criteria were used to gather the data made available to them. Documentation is often lacking, and metadata are rarely provided or are provided with insufficient accuracy. Another problem is the fast-moving nature of this field. Not only do platforms change their selection algorithms and the structure of the data itself, but they also do not document or insufficiently document these changes, thus rendering these data very volatile and calling into question the replicability of studies.

The nature of unstructured data itself also means that certain quality criteria, which are tried-and-tested for the analysis of self-generated data, cannot be met by researchers or are met only at high costs. Existing templates for questionnaires for the ***pre-registration*** of studies with unstructured data are frequently an obstacle because they are not always transferable. Pre-registration is especially difficult for exploratory or data-driven research. However, the many pre-processing steps and various options for operationalisation when conducting research with unstructured data are, in fact, an important argument for pre-registration. It enables researchers to mull things over in advance, to transparently document decisions (and revise them if necessary), and, in the best case, to bring in constructive feedback from colleagues. If pre-registration is not possible or has not been done, the work steps should be documented and published retrospectively in order to increase transparency as much as possible. Transparency is especially important when working with unstructured data.

Regarding the *FAIR principles*, according to which data should be *findable*, *accessible*, *interoperable*, and *re-usable*, re-usability of data is a key issue. Research with unstructured data is often very specific, and it is often not allowed to share data in full. This restriction further increases the requirements for the ***documentation of data processing and the deployment of metadata*** to enable other researchers to replicate results with the data they requested from platforms. Both project funding schemes and the funding of science institutions should take these requirements for the documentation of data processing and the deployment of metadata into account.

5.3 Governance

The widespread use of unstructured data also poses new questions regarding the establishment of an appropriate institutional framework for scientific work. Adequate approaches to governance must keep different goals and requirements in mind: How are the legitimate interests of the users who are leaving the traces of data protected? How does one take into account the legitimate interests of organisations, companies, and state actors that create, store, and analyse unstructured data? How does one make sure researchers have access to socially relevant troves of unstructured data? And, finally: How does one find a common understanding of the rules of good scientific practice and ensure that they are widely adhered to?

To do so, appropriate governance models must tackle two challenges. First, they must reconcile norms that have very different objectives. Take data protection requirements, for example. Anonymisation and pseudonymisation procedures for data make scientific use costlier, create obstacles for certain analytical access paths, and make the shared use of data more difficult. Another challenge is establishing an appropriate balance between the (intellectual) property rights of corporations, whose value creation networks generate these unstructured data, and the legitimate interest of science and research, enshrined in the constitution, in accessing data troves that are both economically valuable and socially relevant.

By the same token, governance structure must integrate ***a large number of diverse actors***. First of all, these actors include the ***researchers*** themselves, who are working with unstructured data with very different access paths and research questions in a field that is seeing dynamic growth. Here, we observe a heterogeneous scientific practice that is finding different responses to the regulation goals sketched out above.

Furthermore, ***academic organisations*** must be considered. Such organisations include universities and research institutions that create frameworks for handling unstructured data in terms of rules and resources for their members. Moreover, academic associations are relevant institutions for finding a common under-

standing of good scientific practice as well as actors in research sponsoring, who shape scientific practice through, for example, requirements for data management. Academic journals and other relevant publications are also significant because they shape the field through their regulation of manuscripts and the availability of data that published articles are based on.

Regarding structures and resources, ***actors in (science) policy*** play a particularly relevant role. Moreover, a factor of central importance is that unstructured data are generated to a relevant extent by ***commercial companies***, specifically, by ***global platforms***. Their sovereignty over user data plays a key role in their business models, thus impeding the accessibility of these data troves for research purposes, especially seeing as access is possible only via APIs that could be reconfigured unilaterally at any given time.

Well-developed activities in science and science policy, for example, in the area of the National Research Data Infrastructure (NFDI), include working on making unstructured data systematically available to other researchers. As yet, these activities are contrasted by a comparatively low density of regulations at the level of academic associations. In addition to data availability issues, there are questions about the methods needed to generate valid and reliable findings in their respective fields. Recommendation papers on working with unstructured data are available at very different levels of detail. One example is the working group of the German Communication Association (DGPuK), which developed recommendations for handling research data in communication studies (Peter et al., 2020). Generally speaking, the recommendation is to intensify the discourse within scientific disciplines and to find formats that can be used to develop a ***common understanding of good scientific practice*** that goes beyond the specifics of certain disciplines. Here, it would also make sense to include the expertise of research sponsors, infrastructures, and academic publications.

In terms of deliberations on appropriate governance structures, the need for ***regulation regarding the cooperation with commercial providers at a political level*** is of central importance. On the one hand, it is problematic that companies can curate their data themselves and, on the other hand, that it is up to them to decide whether to cooperate with science and research. This situation creates strong dependencies where the economic and strategic interests of platforms exert a significant influence on the research questions that are being pursued and the data that are being analysed. Whereas there are first beginnings in the current debate on regulating digital offerings that are capable of increasing researchers' access to data, the extent to which these beginnings will be realised and whether they can lower the access barriers for researchers remain to be seen. There is still a need for policymakers to increase the pressure on platform providers. However, it is difficult to place political demands on commercial companies as long as public institutions, too, are not taking on a pioneering role in providing researchers with easy access to data, for example, official statistics data. Here, coherent action towards the most open access to data possible would be desirable. Some first initiatives are moving in this direction, for example, the European Digital Media Observatory (EDMO[2]).

5.4 Resources

Due to the volume and temporal dynamism of unstructured data, it takes significantly more resources to analyse such data than it does to handle more selectively gathered datasets. One reason (among many) is that there is little to no standardisation, making case-by-case examinations necessary to a larger extent.

From a resource perspective, this problem is not just a technical one because the technologies that are necessary to archive and analyse larger amounts of data are basically available. Existing infrastructures, too, can be used for this purpose and can be scaled up with comparatively little financial expenditure. Rather, the challenge consists of configuring these technologies in technical platforms (e.g., Dataverse[3]) in such a way that they create added value for users from science and research (Hemphill, 2019). Such configuration is as much about ***setting up the platform*** as it is about the maintenance and continuous adjustment to new technological requirements in terms of, for example, availability, efficiency, data protection, and the prevention of cyber attacks. Moreover, the focus is on content requirements such as those relating to findability, data quality, formatting, and metadata (Breuer et al., 2021).

2 https://edmo.eu/
3 https://dataverse.harvard.edu/

The aforementioned challenges can be dealt with, at least in part, by ***investing in the technical infrastructure***. More important, however, are the human resources. Handling data with information technology requires new skill profiles, which are rarely reflected in established job descriptions in science organisations. If it is possible to define new job profiles and make the necessary resources available for qualification and additional training, recruiting personnel who are qualified to work in science still requires considerable effort. When competing with commercial actors with considerable financial resources, it might be more effective to highlight the value for society that working in research can produce in order to attract talented applicants to work in the field of scientific data management, at least for limited periods of time.

In addition to issues related to the ***recruitment training of qualified employees***, additional human resources are needed to reliably provide and document unstructured data. In this context, the absolute size of budget funds is also relevant. However, the logic applied to funding is almost equally important. Here, the demands for a permanent infrastructure, which is necessary to make unstructured data useable, collide with a project-oriented approach to research funding, which, while able to create incentives for systematic data management, is not fully capable of safeguarding the long-term availability of unstructured data. In the long run, it is important to think about the decentral support functions that should be made available at the level of universities and institutes. Furthermore, it seems sensible to work towards greater cooperation between research infrastructure providers that already have extensive experience in handling unstructured data.

Another way to move towards qualification and knowledge-building is to increase the ***cooperation between the social sciences and computer science***. Ideally, the two disciplines will complement each other, and knowledge transfer will take place in both directions. This ideal situation could be very productive because one complements the other in terms of content and method: Whereas the social sciences focus on describing and explaining, computer science is aimed at making data-driven predictions (*prediction tasks*). In principle, research design (including choice of methods) and quality standards are no different between studies in the social sciences and those in computer science. However, the publication cultures in the two disciplines are very different. Before embarking on a joint project, it might therefore be advisable to agree in advance on whether a study will be a social science or a computer science study. There is a risk of conflict in the context of such interdisciplinary cooperation if one of the two sides ends up in a secondary service role. The emergence of *computational social science* can be seen as a reaction to this problem: People would rather acquire the relevant skills than take on a service role.

6. References

Amaya, A., Biemer, P. & Kinyon, D. (2020). Total error in a big data world: Adapting the TSE framework to big data. *Journal of Survey Statistics and Methodology, 8*(1), 89–119. https://doi.org/10.1093/jssam/smz056

Angrist, J. D. & Pischke, J.-S. (2009). *Mostly harmless econometrics: An empiricist's companion.* Princeton University Press.

Bachl, M. & Scharkow, M. (2017). Correcting measurement error in content analysis. *Communication Methods and Measures*, *11*(2), 87–104. https://doi.org/10.1080/19312458.2017.1305103

Biemer, P. & Amaya, A. (2020). Total error frameworks for found data. In C. A. Hill, P. P. Biemer, T. D. Buskirk, L. Japec, A. Kirchner, S. Kolenikov & L. E. Lyberg (Eds.), *Big data meets survey science* (pp. 131–161). Wiley. https://doi.org/10.1002/9781118976357.ch4

Borkenau, P. (2006). Selbstbericht. In F. Petermann & M. Eid (Eds.), *Handbuch der Psychologie: Vol. 4. Handbuch der psychologischen Diagnostik* (pp. 135–142). Hogrefe.

Bosch, O. & Revilla, M. (2022). When survey science met online tracking: Presenting an error framework for metered data. *Journal of the Royal Statistical Society: Series A (Statistics in Society),* 1–29. https://doi.org/10.1111/rssa.12956

Bosse, S., Dahlhaus, L. & Engel, U. (2021). Web data mining: Collecting textual data from web pages using R. In U. Engel, A. Quan-Haase, S. Liu & L. Lyberg (Eds.), *Handbook of computational social science: Data science, statistical modelling, and machine learning methods* (pp. 46–70). Routledge.

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

Breuer, J., Borschewski, K., Bishop, L., Vávra, M., Štebe, J., Strapcova, K. & Hegedűs, P. (2021). *Archiving social media data: A guide for archivists and researchers.* https://doi.org/10.5281/ZENODO.5041072

Christen, P. (2012). *Data matching: Concepts and techniques for record linkage, entity resolution, and duplicate detection. Data-Centric Systems and Applications.* Springer. https://doi.org/10.1007/978-3-642-31164-2

Christensen, G. (2018). *Manual of best practices in transparent social science research.* UC Berkeley Initiative for Transparency in the Social Sciences. https://github.com/garretchristensen/bestpracticesmanual* (Accessed August 2, 2022)

Cohen, J., Cohen, P., West, S. G. & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Erlbaum.

Dwork, C., McSherry, F., Nissim, K. & Smith, A. (2006). Calibrating noise to sensitivity in private data analysis. In S. Halevi & T. Rabin (Eds.), *Theory of cryptography. TCC 2006. Lecture notes in computer science* (Vol. 3876, pp. 265–284). Springer. https://doi.org/10.1007/11681878_14

Dwork, C. & Roth, A. (2014). The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science, 9*(3–4), 211–407. https://doi.org/10.1561/0400000042

Eberendu, A. (2016). Unstructured data: An overview of the data of big data. *International Journal of Computer Trends and Technology (IJCTT), 38*(1), 46–50. https://doi.org/10.14445/22312803/IJCTT-V38P109

Eckman, S. & Kreuter, F. (2017). The undercoverage-nonresponse tradeoff. In P. P. Biemer, E. de Leeuw, S. Eckman, B. Edwards, F. Kreuter, L. E. Lyberg, N. C. Tucker & B. T. West (Eds.), *Total survey error in practice* (pp. 95–113). John Wiley & Sons, Inc. https://doi.org/10.1002/9781119041702.ch5

Eid, M. & Schmidt, K. (2014). *Testtheorie und Testkonstruktion*. Hogrefe.

Evans, G. & King, G. (2022). Statistically valid inferences from differentially private data releases, with application to the facebook URLs dataset. *Political Analysis*, 1–21. https://doi.org/10.1017/pan.2022.1

Fischer, C., Fishman, B. & Schoenebeck, S. Y. (2019). New contexts for professional learning: Analyzing high school science teachers' engagement on Twitter. *AERA Open, 5*(4). https://doi.org/10.1177/2332858419894252

Fritsche, I. & Linneweber, V. (2006). Nonreactive methods in psychological research. In M. Eid & E. Diener (Eds.), *Handbook of multimethod measurement in psychology* (pp. 189–203). American Psychological Association. https://doi.org/10.1037/11383-014

Gandomi, A. & Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management, 35*(2), 137–144. https://doi.org/10.1016/j.ijinfomgt.2014.10.007

Grimmer, J., Roberts, M. & Stewart, B. (2022). *Text as data: A new framework for machine learning and the social sciences*. Princeton University Press.

Groves, R. M. & Lyberg, L. (2010). Total survey error: past, present, and future: Past, present, and future. *Public Opinion Quarterly, 74*(5), 849–879. https://doi.org/10.1093/poq/nfq065

Hardwicke, T., Bohn, M., MacDonald, K., Hembacher, E., Nuijten, M., Peloquin, B., deMayo, B., Long, B., Yoon, E. & Frank, M. (2021). Analytic reproducibility in articles receiving open data badges at the Journal Psychological Science: An observational study. *Royal Society open science, 8*(1), 201494. https://doi.org/10.1098/rsos.201494

Heckathorn, D. D. (1997). Respondent-driven sampling: A new approach to the study of hidden populations. *Social Problems, 44*(2), 174–199. https://doi.org/10.2307/3096941

Hemphill, L. (2019). *Updates on ICPSR's Social Media Archive (SOMAR)*. https://doi.org/10.5281/ZENODO.3612676

Heyer, G., Wiedemann, G. & Niekler, A. (2018). Topic-Modelle und ihr Potenzial für die philologische Forschung. In H. Lobin, R. Schneider & A. Witt (Eds.), *Digitale Infrastrukturen für die germanistische Forschung* (pp. 351–368). De Gruyter. https://doi.org/10.1515/9783110538663-016

Hox, J. (2017). Computational social science methodology, anyone? *Methodology, 13*(1), 3–12. https://doi.org/10.1027/1614-2241/a000127

Hsieh, Y. P. & Murphy, J. (2017). Total Twitter error. In P. P. Biemer, E. de Leeuw, S. Eckman, B. Edwards, F. Kreuter, L. E. Lyberg, N. C. Tucker & B. T. West (Eds.), *Total survey error in practice* (pp. 23–46). John Wiley & Sons, Inc. https://doi.org/10.1002/9781119041702.ch2

Jürgens, P., Stark, B. & Magin, M. (2020). Two half-truths make a whole? On bias in self-reports and tracking data. *Social Science Computer Review, 38*(5), 600–615. https://doi.org/10.1177/0894439319831643

Keusch, F., Bähr, S., Haas, G.-C., Kreuter, F. & Trappmann, M. (2020). Coverage error in data collection combining mobile surveys with passive measurement using apps: Data from a German national survey. *Sociological Methods & Research, 0*(0), 1–38. https://doi.org/10.1177/0049124120914924

Keusch, F., Bähr, S., Haas, G.-C., Kreuter, F., Trappmann, M. & Eckman, S. (2022). Non-participation in smartphone data collection using research apps. *Journal of the Royal Statistical Society: Series A (Statistics in Society).* Advance online publication. https://doi.org/10.1111/rssa.12827

Kinder-Kurlanda, K. & Weller, K. (2014). „I always feel it must be great to be a hacker!". In F. Menczer, J. Hendler, W. Dutton, M. Strohmaier, E. T. Meyer & C. Cattuto (Eds.), *Proceedings of the 2014 ACM conference on web science – WebSci, 14,* Bloomington, Indiana, USA. 6/23/2014 – 6/26/2014 (pp. 91–98). ACM Press. https://doi.org/10.1145/2615569.2615685

Kreuter, F., Haas, G.-C., Keusch, F., Bähr, S. & Trappmann, M. (2020). Collecting survey and smartphone sensor data with an app: Opportunities and challenges around privacy and informed consent. *Social Science Computer Review, 38*(5), 533–549. https://doi.org/10.1177/0894439318816389

Krippendorff, K. (2008). Validity. In W. Donsbach (Ed.), *The international encyclopedia of communication*. Wiley-Blackwell.

Krippendorff, K. (2012). *Content analysis: An introduction to its methodology* (3rd ed.). SAGE Publications.

Lamnek, S. & Krell, C. (2016). *Qualitative Sozialforschung* (6th ed.). Beltz.

Lazer, D. & Radford, J. (2017). Data ex machina: Introduction to big data. *Annual Review of Sociology, 43*(1), 19–39. https://doi.org/10.1146/annurev-soc-060116-053457

Leskovec, J., Rajaraman, A. & Ullman, J. D. (2020). *Mining of massive datasets* (3rd ed.). Cambridge University Press. https://doi.org/10.1017/9781108684163

Menchen-Trevino, E. (2013). Collecting vertical trace data: Big possibilities and big challenges for multi-method research. *Policy & Internet, 5*(3), 328–339. https://doi.org/10.1002/1944-2866.POI336

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). Macmillan.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*(9), 741–749. https://doi.org/10.1037/0003-066X.50.9.741

Morstatter, F., Pfeffer, J., Liu, H. & Carley, K. (2013). Is the sample good enough? Comparing data from Twitter's streaming API with Twitter's firehose. In *Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media 7*(1) (pp. 400–408). https://ojs.aaai.org/index.php/ICWSM/article/view/14401

Nyhuis, D. (2021). Application programming interfaces and web data for social research. In U. Engel, A. Quan-Haase, S. Liu & L. Lyberg (Eds.), *Handbook of computational social science: Data science, statistical modelling, and machine learning methods* (pp. 33–45). Routledge.

Peter, C., Breuer, J., Masur, P. K., Scharkow, M. & Schwarzenegger, C. (2020). Empfehlungen zum Umgang mit Forschungsdaten in der Kommunikationswissenschaft: AG Forschungsdaten im Auftrag des Vorstands der DGPuK. *Studies in Communication and Media, 9*(4), 599–626. https://doi.org/10.5771/2192-4007-2020-4-599

RatSWD (Rat für Sozial- und Wirtschaftsdaten). (2020). *Datenerhebung mit neuer Informationstechnologie: Empfehlungen zu Datenqualität und -management, Forschungsethik und Datenschutz* (Output Series, 6. Berufungsperiode No. 6). Berlin. https://doi.org/10.17620/02671.47

Rauchfleisch, A. & Kaiser, J. (2020). The false positive problem of automatic bot detection in social science research. *PloS one, 15*(10), e0241045. https://doi.org/10.1371/journal.pone.0241045

Reveilhac, M., Steinmetz, S. & Morselli, D. (2022). A systematic literature review of how and whether social media data can complement traditional survey data to study public opinion. *Multimedia tools and applications, 81*(7), 10107–10142. https://doi.org/10.1007/s11042-022-12101-0

Rodolfa, K. T., Saleiro, P. & Ghani, R. (2021). Bias and fairness. In I. Foster, R. Ghani, R. S. Jarmin, F. Kreuter & J. Lane (Eds.), *Big data and social science: Data science methods and tools for research and practice* (2nd ed.). CRC Press.

Salganik, M. J. (2018). *Bit by bit: Social research in the digital age*. Princeton University Press. http://bvbr.bib-bvb.de:8991/F?func=service&doc_library=BVB01&local_base=BVB01&doc_number=029920334&sequence=000001&line_number=0001&func_code=DB_RECORDS&service_type=MEDIA https://doi.org/Social

Schnell, R., Hill, P. B. & Esser, E. (2011). *Methoden der empirischen Sozialforschung* (9th ed.). Oldenbourg.

Sen, I., Flöck, F., Weller, K., Weiß, B. & Wagner, C. (2021). A total error framework for digital traces of human behavior on online platforms. *Public Opinion Quarterly, 85*, 399–422. https://doi.org/10.1093/poq/nfab018

Shadish, W. R., Cook, T. D. & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Houghton Mifflin Company.

Simonsohn, U., Simmons, J. P. & Nelson, L. D. (2020). Specification curve analysis. *Nature Human Behavior, 4*(11), 1208–1214. https://doi.org/10.1038/s41562-020-0912-z

Stier, S., Breuer, J., Siegers, P. & Thorson, K. (2020). Integrating survey data and digital trace data: Key issues in developing an emerging field. *Social Science Computer Review, 38*(5), 503–516. https://doi.org/10.1177/0894439319843669

Stodden, V., Seiler, J. & Ma, Z. (2018). An empirical analysis of journal policy effectiveness for computational reproducibility. *Proceedings of the National Academy of Sciences of the United States of America, 115*(11), 2584–2589. https://doi.org/10.1073/pnas.1708290115

Taleb, I., Serhani, M. A. & Dssouli, R. (2018). Big data quality assessment model for unstructured data. In *2018 International Conference on Innovations in Information Technology (IIT)*, Al Ain, United Arab Emirates. 11/18/2018 – 7 11/19/2018 (pp. 69–74). IEEE. https://doi.org/10.1109/INNOVATIONS.2018.8605945

Tanwar, M., Duggal, R. & Khatri, S. K. (2015). Unravelling unstructured data: A wealth of information in big data. In *4th International Conference on Reliability, Infocom Technologies and Optimization (ICRITO) (Trends and Future Directions)*, Noida, India (pp. 1–6). IEEE. https://doi.org/10.1109/ICRITO.2015.7359270

Tokle, J. & Bender, S. (2021). Record linkage. In I. Foster, R. Ghani, R. S. Jarmin, F. Kreuter & J. Lane (Eds.), *Big data and social science: Data science methods and tools for research and practice* (2nd ed., pp. 43–65). CRC Press.

Tufekci, Z. (2014). *Big questions for social media big data: Representativeness, validity and other methodological pitfalls: ICWSM '14: Proceedings of the 8th International AAAI Conference on Weblogs and Social Media*. https://arxiv.org/pdf/1403.7400

van Atteveldt, W., Althaus, S. & Wessler, H. (2021). The trouble with sharing your privates: Pursuing ethical open science and collaborative research across national jurisdictions using sensitive data. *Political Communication, 38*(1–2), 192–198. https://doi.org/10.1080/10584609.2020.1744780

Webb, E. J., Campbell, D. T., Schwartz, R. D. & Sechrest, L. (1966). *Unobtrusive measures: Nonreactive research in the social sciences*. Rand McNally.

West, B. T., Sakshaug, J. W. & Aurelien, G. A. S. (2016). How big of a problem is analytic error in secondary analyses of survey data? *PloS one, 11*(6), e0158120. https://doi.org/10.1371/journal.pone.0158120

Young, C. & Holsteen, K. (2017). Model uncertainty and robustness. *Sociological Methods & Research, 46*(1), 3–40. https://doi.org/10.1177/0049124115610347

Contributors

Members of the working group ‘Challenges in the scientific collection and use of unstructured data’

Prof. Stefan Bender
Deutsche Bundesbank, RatSWD

Prof. Dr. Michael Eid *(co-chair)*
Free University of Berlin, RatSWD

Prof. Dr. Christiane Gross
Julius Maximilian University of Würzburg, RatSWD

Prof. Dr. Stefan Liebig
Socio-Economic Panel (SOEP) at the German Institute for Economic Research (DIW Berlin), Free University of Berlin, RatSWD

Prof. Dr. Oliver Lüdtke *(co-chair)*
Leibniz Institute for Science and Mathematics Education (IPN), Christian-Albrecht University of Kiel, RatSWD

Prof. Dr. Laura Seelkopf
Ludwig Maximilian University of Munich, RatSWD

Prof. Dr. Lars Rinsdorf
Stuttgart Media University

Prof. Dr. Mark Trappmann
Institute for Employment Research (IAB) at the German Federal Employment Agency (BA), University of Bamberg, RatSWD

Consultation

Dr. Johannes Breuer
GESIS – Leibniz Institute for the Social Sciences

Prof. Dr. Christian Fischer
Eberhard Karl University of Tübingen

Dr. Stephanie Geise
University of Münster

Dr. Theresa Gessler
University of Zurich

Dr. Fenne große Deters
University of Potsdam

Dr. Pascal Jürgens
Johannes Gutenberg University Mainz

Prof. Dr. Florian Keusch
University of Mannheim

Prof. Dr. Wenzel Matiaske
Helmut Schmidt University

Prof. Matthias Mehl, PhD.
University of Arizona

Prof. Dr. Jürgen Pfeffer
Technical University of Munich

Julia Rakers
University of Duisburg-Essen

Christian Strippel
Free University of Berlin

Dr. Katrin Weller
GESIS – Leibniz Institute for the Social Sciences

RatSWD Office

Marie Eilers

Imprint

Publisher:
German Data Forum (RatSWD)
Office
Am Friedrichshain 22
10407 Berlin
office@ratswd.de
https://www.ratswd.de

Editor:
Marie Eilers

Layout:
Claudia Kreutz

Translation | proof-reading:
Simon Wolff, simon@translate.berlin | Dr. Jane Zagorski, janezagorski@gmail.com

Berlin, February 2024

RatSWD Output:
The RatSWD Output Series documents the work of the German Data Forum (RatSWD) in its 7th appointment period (2020–2023). It serves to publish its statements and recommendations and to make them available to a broad readership.

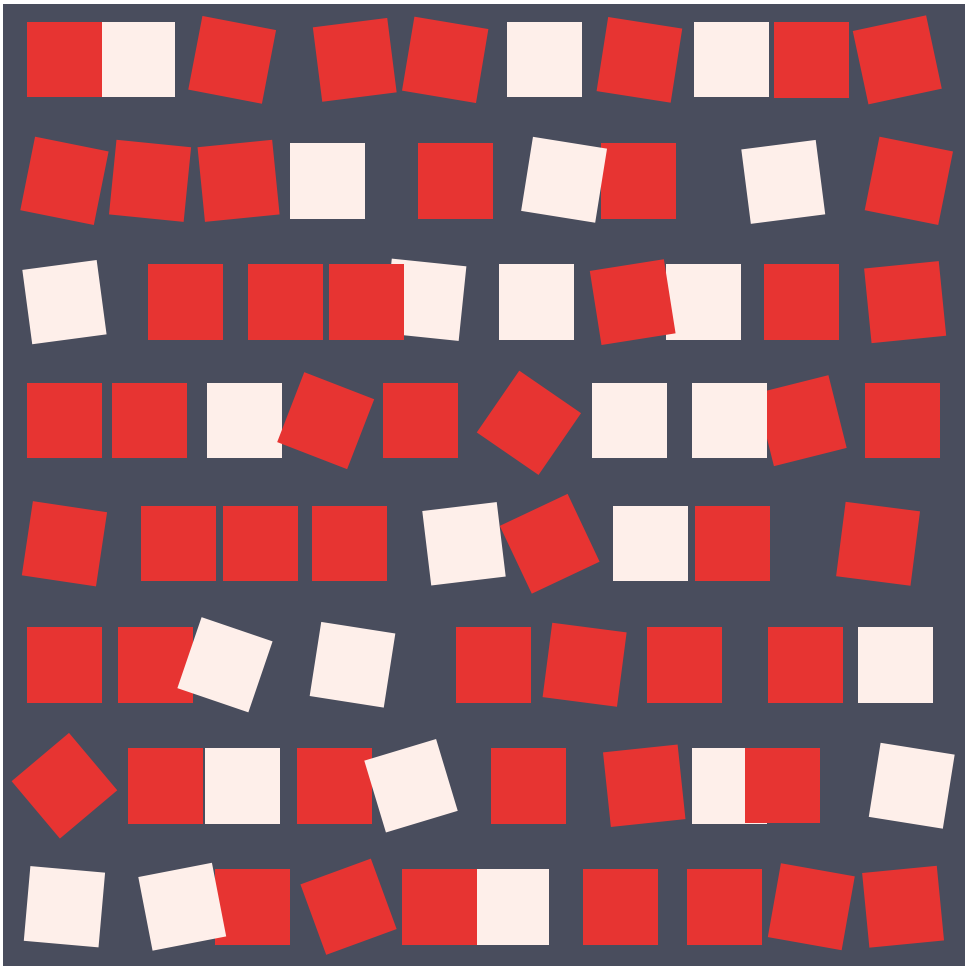
The RatSWD Office is funded as part of KonsortSWD within the NFDI framework by the German Research Foundation (DFG) – project number: 442494171.

DOI: 10.17620/02671.92

Suggested citation:
RatSWD [German Data Forum] (2024). Generation and use of unstructured data in the social, behavioural, and economic sciences: challenges and recommendations. (RatSWD Output Series, 7. Appointment period No. 2). Berlin. https://doi.org/10.17620/02671.92

The German Data Forum (RatSWD) has been advising the Federal Government and the governments of the German states on questions of research data infrastructure for the empirical social, behavioural and economic sciences since 2004. In the RatSWD, ten representatives of the social, behavioural and economic disciplines, legitimised by election, work together with ten representatives of data production.

The RatSWD is part of the Consortium for the Social, Behavioural, Educational and Economic Sciences (KonsortSWD) in the National Research Data Infrastructure (NFDI). It sees itself as an institutionalised forum for dialogue between science and data producers and develops recommendations and statements. In doing so, it is committed to an infrastructure that provides science with broad, flexible and secure access to data. These data are provided by governmental, science-based and private-sector actors. Currently, the RatSWD has accredited 41 research data centres and promotes their cooperation.



www.ratswd.de