Data Linking - Linking survey data with geospatial, social media, and sensor data (Version 1.0)

Beuthner, Christoph; Breuer, Johannes; Jünger, Stefan

Veröffentlichungsversion / Published Version
Arbeitspapier / working paper

Zur Verfügung gestellt in Kooperation mit / provided in cooperation with:
GESIS - Leibniz-Institut für Sozialwissenschaften

Empfohlene Zitierung / Suggested Citation:
Beuthner, C., Breuer, J., & Jünger, S. (2021). *Data Linking - Linking survey data with geospatial, social media, and sensor data (Version 1.0).* (GESIS Survey Guidelines). Mannheim: GESIS - Leibniz-Institut für Sozialwissenschaften. https://doi.org/10.15465/gesis-sg_en_039

gesis Leibniz-Institut für Sozialwissenschaften

Data Linking - Linking survey data with geospatial, social media, and sensor data

Christoph Beuthner, Johannes Breuer, & Stefan Jünger

March 2021, Version 1.0

Abstract

Survey data are still the most commonly used type of data in the quantitative social sciences. However, as not everything that is of interest to social scientists can be measured via surveys, and the self-report data they provide have certain limitations, such as recollection or social desirability bias, researchers have increasingly used other types of data that are not specifically created for research. These data are often called “found data” or “non-designed data” and encompass a variety of different data types. Naturally, these data have their own sets of limitations. One way of combining the unique strengths of survey data and these other data types and dealing with some of their respective limitations is to link them. This guideline first describes why linking survey data with other types of data can be useful for researchers. After that, it focuses on the linking of survey data with three types of data that are becoming increasingly popular in the social sciences: geospatial data, social media data, and sensor data. Following a discussion of the advantages and challenges associated with linking survey data with these types of data, the guideline concludes by comparing their similarities, presenting some general recommendations regarding linking surveys with other types of (found/non-designed) data, and providing an outlook on current developments in survey research with regard to data linking.

Citation

Beuthner, Christoph, Johannes Breuer, and Jünger, Stefan (2021). Data Linking - Linking survey data with geospatial, social media, and sensor data. Mannheim, GESIS - Leibniz Institute for the Social Sciences (GESIS Survey Guidelines).

DOI: 10.15465/gesis-sg_en_039

Why link survey data with other types of data?

Linking survey data with other types of data allows researchers to perform analyses and gain insights that would not be possible with survey data alone. In this paper, we define data linking as the process combining data from multiple sources for joint analyses. While surveys can be used to collect data on a huge variety of subjects related to human behavior, they also have limitations. For example, answers by respondents may be heavily influenced by social desirability or recall bias (e.g., if they are asked about events or behaviors that are quite rare or occurred long ago) or it may simply not be possible to reliably measure certain constructs of interests with surveys (e.g., about environmental or social characteristics of specific areas of residence or other geographic regions).

Data from other sources can be used to contextualize survey data and the combination of survey data with additional data types enables researchers to answer novel research questions or to test the robustness of findings that are based exclusively on self-reported data from surveys. Linking survey data with other data types enables the inclusion of new variables to explain statistical relationships and can lead to a more comprehensive understanding of social phenomena.

There are various types of data that survey data can be linked with. Three types of data that are being increasingly used in the social sciences are geospatial, social media, and sensor data. Unlike survey data, these data are typically not produced by the research process itself. Hence, such data (especially those from social media and sensors) are also often referred to as found or non-designed data. These data can be a byproduct of behavior or technology use or be collected independently by public institutions (examples of this latter category, e.g., include official statistics on land use or unemployment rates for specific regions). The availability of such data and the opportunity of linking them with survey data can also reduce response burden by shortening questionnaires. If there is external data available, there is no need to ask survey respondents about a given topic. In addition, these recorded data are less likely to be (directly) influenced by social desirability and are also unaffected by problems of recollection (as they measure activity when it happens instead of retrospectively).

How can survey data be linked with other types of data?

Within the type of data linking that we focus on in this survey guideline, there are different ways in which the data can be linked. Importantly, the specific type of data with which survey data are combined and the way they are collected determine the way in which they can be linked. In general, there are two key dimensions on which the ways of linking surveys and other types of data can differ: 1) The level on which they are linked and 2) the phase of research during which they are linked. With regard to the level of linking, the data can either exist and be linked on the individual level or the aggregate level. While the units of observation in surveys are individuals, the data to be linked does not have to be individual-level data. It is possible to link aggregate data to survey data based on geographic regions or time periods. For the second dimension, both ex ante and ex post linkage are possible. Survey data and the additional data can be collected for the explicit purpose of being linked. This is what we would call ex-ante linking, meaning that the linking is considered already in the research design phase, and the data are collected accordingly in a way and format that enables or facilitates the linking.

Conversely, the data can also be collected independently and linked at a later point in time. This is typically done in cases where the dataset(s) with which the survey data will be linked already exist(s) and the term we use for this is ex-post linking. As these dimensions of linking are somewhat abstract, Figure 1 provides examples of the different ways of linking with survey data for the three types of data that we

focus on in this survey guideline.

	Data type	Ex ante	Ex post
Aggregate level	**Social media**	• Collecting tweets for the same region and period of time as the survey data using the Stream API	• Linking survey data with counts of posts (about a certain topic) or aggregate sentiment scores for posts from existing social media data collections for specific regions or time periods
	Geospatial data	• Simultaneous recording of data for surveyed area (e.g., weather, pollution or noise data collected via sensors)	• Linking aggregated survey data for specific geographic areas to available geospatial data (e.g., on access to certain amenities, pollution, noise, etc.)
	Sensor data	• Simultaneous recording of health data of a surveyed group (e.g., a sports team)	• Linking aggregated medical data for specific populations (e.g., blood oxygen levels in previous studies)

Individual level	**Social media data**	• Ask survey respondents for consent to collect their current/latest social media data (e.g., via an API or a browser plugin) for a specified period of time (during + maybe also after the survey field time)	• Ask individuals in surveys for informed consent to collect their historical digital trace data…
			  • via social media APIs
			  • via data donation (e.g., personal Google, Twitter or Facebook archives)

	Geospatial data (note: these are usually not generated/available on the individual level)	• Record location data from respondents (self-report, e.g., via experience sampling or tracked GPS data from devices)	• Linking survey data to existing geospatial data (e.g., on access to certain amenities, pollution, noise, etc.) on the level of the location/address of individual participants
	Sensor data	• Equiping respondents with fitness trackers for the time of the study	• Accessing fitness tracker data stored on respondents devices (e.g., via data donation)

Figure 1: Types and examples of data linking approaches for social media, geo, and sensor data

Linking surveys with geospatial data

One of the most common approaches to enrich survey data with auxiliary information (i.e., using data linking techniques) is to use information from geospatial data sources. In this effort, survey respondents' location information is used to link survey data, which in this case have to be georeferenced survey data, and geospatial data using Geographic Information Systems (GIS). As both data sources are projected in space, this linking can either be done by a one-by-one match, for example, by extracting information from the geospatial data sources at the survey respondents' housing address. Alternatively, it is possible to choose from proximity measures, such as distances of points in space, or extract descriptive statistics from aggregated circular areas around a specific point, so-called buffer areas. No matter which approach of data linking is chosen, these flexible methods provide different outcomes.

Figure 2 exhibits some of the available GIS approaches to link georeferenced survey data and geospatial data. This figure consists of three separate maps, all displaying road traffic noise measurements in decibels on main roads as the geospatial data source. The black dot in the middle is a geo-coordinate of a fictional survey respondent. In subfigure (a), the road traffic measurement's decibel value is added

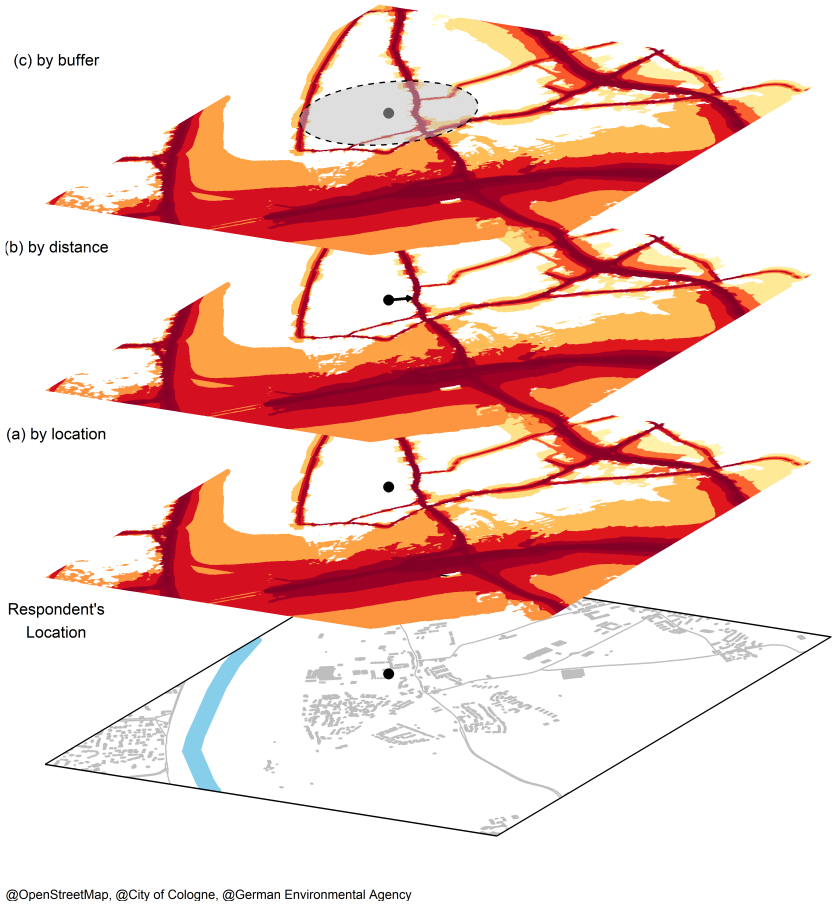

Figure 2: Spatial linking methods

to the survey respondent's geo-coordinate solely by the one by one location. Subfigures (b) and (c) exemplify the actual flexibility of the joined projection in space. (b) shows the capturing of a distance to the next road traffic noise source of $\leq$ 65 dB(A), which can be an approximation of noise for respondents in surveys who do not live on the main road. (c) shows an even more advanced approach as it draws a circular buffer around respondents' geo-coordinates. It calculates some descriptive statistics, e.g., the mean dB(A) level within 500 meters, and adds them to the respondent's location. These examples show that even with the same data sources, different approaches exist to extract information from geospatial data sources.

Accordingly, researchers use these methods in various settings. The method of drawing buffers is particularly useful if the aim is to grasp the geographic variation of some effects, for example, the influence of immigrant rates on social trust (Sluiter, Tolsma, & Scheepers, 2015). Other applications include the calculation of distances to pollution sources to estimate social inequalities of environmental hazards (Crowder & Downey, 2010). Generally, the method of spatial linking with GIS can provide new insights in research either by providing opportunities to answer innovative research questions or to corroborate or even reject previous findings (Jünger, 2019).

Apart from some technical challenges of using GIS methods, there is also one major challenge concerning georeferenced survey data: data protection. Using survey respondents' location information on the

smallest scale possible –the address level– poses some challenges in that regard. One challenge is the legal situation because, according to data protection legislation (in Germany and the EU), we cannot store personal information (i.e., addresses and geo-coordinates) together with survey information. The other challenge affects the disclosure risk of the data. Additional information on respondents' living environment increases the risk of re-identification. A typical example is a person with a rare job sharing some unique sociodemographic characteristics, such as a female lawyer with seven children from whom we know the neighborhood location (Schweers, Kinder-Kurlanda, Müller, & Siegers, 2016). The good news is that both challenges can be navigated by imposing organizational arrangements within the institution which holds the survey and address data.

This organizational arrangement comprises separating the spatial linking of personal information with geospatial data attributes from linking these derived attributes with the actual survey attributes. The workflow we developed at GESIS to adress this is as follows: We first apply spatial linking methods on the survey respondents' georeferenced addresses and extract some geospatial data attributes, such as road traffic noise dB(A) values, distances, or related measures. In the next step, we delete the geo-coordinates in this dataset, or, as a measure of data protection, we coarsen them to some higher aggregated spatial units. What follows is the most crucial part: Through a correspondence table of individual identifiers in the survey data and different individual identifiers in the address data we change the identifiers in our intermediate dataset to the survey data ones. These new identifiers can then provide the basis for linking the derived attributes from the geospatial data to the actual survey data. This final linked dataset comprises the original survey data *and* geospatial data attributes.

Still, these data may be sensitive, and it might not be possible to distribute them openly. For example, suppose these data contain information on immigrant rates from the German Census on 1 km^2 grid cells with detail of 2 decimal points. In that case, potential privacy attackers could use these census data at least to substantially narrow down the potential set of 1 km^2 grid cells. Additional sociodemographic information could then be used to identify individual survey respondents. In the conclusion of this survey guideline, we discuss potential solutions to this issue as they are quite similar for the two other types of data we will discuss in the next sections.

Recommendations for linking surveys with geospatial data

- Decide which type of spatial linking you aim to use, taking into account the complexity of their implementation.
- Consider being flexible and choosing variations of the selected method since you have to do these steps before the actual analysis (but don't forget theory!).
- Assess the sensitivity of linked datasets even after deleting or coarsening direct spatial information.

Linking surveys with social media data

Data from social media platforms can be used to answer a wide range of social science research questions. If researchers want to include (dimensions of) social media use as a predictor or outcome variable in their research, using data from the platforms of interest is more reliable than using self-reports of usage behavior. Several studies have shown that self-reports tend to be inaccurate due to social desirability or recall bias (Araujo, Wonneberger, Neijens, & de Vreese, 2017; Prior, 2009; Scharkow, 2016). Of course, social media data can also be used to measure other variables, such as the formation and expression of opinions.

A variety of research questions can be investigated with social media data due to the large variety of data types in this broad category. Social media data can come from a wide range of platforms (e.g., *Twitter*,

Facebook, or *reddit*) and encompass various data types. These can be textual (posts, comments, etc.), audiovisual (images, audio, video), network (friends/followers/contacts), or other forms of data and metadata (e.g., information about the time or location of a post). Importantly, the type of data that can be used and the format they are in also depend on the collection method. There are various options for collecting social media data. Researchers can purchase them from data re-sellers or market research companies, cooperate with social media companies to access their data, or collect the data themselves via the Application Programming Interfaces (API) of platforms or web scraping (Breuer, Bishop, & Kinder-Kurlanda, 2020). In addition, researchers can also (re-)use data from existing collections, such as the *GESIS Social Media Monitoring*, *TweetsKB* or archived social media collections (Kinder-Kurlanda, Weller, Zenk-Möltgen, Pfeffer, & Morstatter, 2017). If the data are collected through APIs, they usually come in a structured format (often in the form of `.json` files). However, if the data are, for example, gathered via web scraping, represent networks (of users or content), or include (audio-)visual material, they have to be processed into a format that social scientists usually use for their analysis: rectangular tabular data. This format is also required for linking these data with survey data.

To date, in (computational) social science research with social media data, collecting data via APIs has been the most commonly used approach. Gathering social media data through platform APIs is a data collection method explicitly allowed by the platform providers (provided that their Terms of Service are respected), whereas this is usually not the case for web scraping. However, despite this fact and the advantage that APIs provide structured data, collecting social media data through APIs entails limitations and risks. APIs typically have rate limits that regulate what data can be accessed, how much can be collected, and how often data requests can be sent. APIs also have Terms of Service (ToS) that specify how the data can be used (and often also if or how they can be shared or published). In addition, social media platforms may change and completely shut down APIs as they wish and at any time. As some major platforms, most notably *Facebook*, have already begun to reduce access to or capabilities of their APIs drastically, researchers have suggested that (computational) social science research is facing an "APIcalypse" (Bruns, 2019) or may entering a "post-API age" (Freelon, 2018).

Given these developments, researchers have started to discuss and explore alternative models of access to social media data. One proposed solution is to collaborate with platform users instead of the platform providers (Halavais, 2019). Most platforms offer users the option to export their personal data, and researchers can invite them to make all or parts of those data available to them (Thorson, Cotter, Medeiros, & Pak, 2019). Alternatively, researchers can also ask study participants to use tools created by them or other researchers, such as browser plugins (Haim & Nienierza, 2019), to collect social media data. Apart from being independent of APIs, such approaches also increase the transparency for the individuals whose data are being collected.

When working with social media data, what is important to consider is that while self-report data can be biased due to social desirability or problems with recollection, social media data often lack relevant individual-level information about users. While there are tools (see, e.g., Z. Wang et al., 2019) for inferring user attributes from social media profiles, these can be wrong or uncertain, and the direct information that social media data provide about the users is usually quite limited (in addition, some ToS of social media APIs, e.g., explicitly prohibit inferring certain individual characteristics). Moreover, relevant outcome variables (e.g., voting intention) are often missing from social media data. Linking surveys and social media data is an approach that allows combining the unique strengths of these two data types while addressing some of their respective limitations (Stier, Breuer, Siegers, & Thorson, 2020).

As shown in Figure 1, social media data can be linked with survey data in various ways. Of course, within these linking categories, researchers can make additional choices that affect what the combined data look like. For example, in the case of individual-level ex-ante linking, researchers can start with/from

the survey or the social media data in the case of ex-ante linking on the individual level. They either first collect social media data (e.g., via an API) and invite individuals whose data are included in this collection to participate in a survey. Alternatively, they ask survey respondents whether they are willing to share or agree to the tracking/collection of (parts of) their social media data.

Notably, both options are associated with specific sampling biases (Jürgens, Stark, & Magin, 2020; Sen, Flöck, Weller, Weiss, & Wagner, 2019). One factor that can introduce biases is that the willingness to have their survey and social media data linked differs among respondents. Al Baghal, Sloan, Jessop, Williams, & Burnap (2020), for example, found that survey mode matters as consent rates were higher in face-to-face surveys in their study. Other factors, such as the (perceived) sensitivity of the social media data in question, privacy concerns, or the size of incentives, are also likely to play a role in this.

The more common option in social science research is to start with the survey and then link social media data from the respondents. This approach has the clear advantage that researchers can directly get informed consent to collect and link the respondents' data. When seeking informed consent, researchers need to inform participants about what data they collect, for what purpose(s) the data are used, how the data are stored, and who will access them. Of course, informed consent needs to adhere to international and national legal regulations (in Europe, e.g., the General Data Protection Regulation) and satisfy ethical standards (as defined by Institutional Review Boards or professional societies). A practical challenge here is to properly inform respondents without overwhelming them with information and technical details. Although informed consent must always be adapted for the specific study, Sloan, Jessop, Al Baghal, & Williams (2020) have developed a flexible template for a study in which they linked data from surveys and Twitter.

Concerning the practicalities of linking survey and social media data, researchers need a unique identifier (or a combination of identifiers that allows for an unambiguous matching of cases) that they can use to match units of observations in the datasets. For individual-level linking, respondents' user names or IDs are a natural choice. If users are asked to provide their user/screen names in the survey, however, it is important to keep in mind that they might misremember or misspell those or provide a user name that is not their own (intentionally or unintentionally). Another thing consideration is that many platforms also allow users to change their user names, which may cause problems if, for example, there is a time gap between asking for consent and collecting the social media data (e.g., via an API). A potential solution for the first issue (incorrect user names) is to have users follow/friend, message, or otherwise contact an account created by the researchers. Regarding the second issue (changing user names), many platforms use unique ID keys to identify users/accounts. Those typically remain the same, even if user names change, and can often be accessed through APIs. Of course, these user IDs would have to be collected right away or as soon as possible after respondents have provided their user names.

When storing and working with the data, researchers should store and process the survey and social media data in a way that minimizes disclosure risk. As social media data tend to be highly disclosive, the survey data and the social media data should be kept separate. This means that they should at least be stored in separate files, while storing them in different places (i.e., on different drives or computers) can further increase data privacy (as can additional measures, such as password protection and encryption of files). Researchers should only combine parts of the data required for specific analyses. Sloan, Jessop, Al Baghal, & Williams (2020) propose such a principled workflow that ensures that no combined dataset contains the full social media and survey data. More specifically, this model proposes that researchers have/use two versions each of the survey data and the social media data: One with the unique identifier that can be used to link them (usually a user name) and one without the identifier. To be able to link the data sets, researchers should create and use unique ID keys that are pseudonymized (unlike the user names, which often are or include real names). When the survey and social media data are combined, this

combined dataset should include these IDs (not the user names) as unique identifiers. Figure 3 is based on Figure 1 in the paper by Sloan, Jessop, Al Baghal, & Williams (2020) and presents a generalized and slightly simplified version of the workflow they proposed for Twitter data for social media in general. The different colors in this figure represent different data sources: The light blue fields represent information from the social media data, the grey-colored boxes are (parts of) the survey data, and the dark blue boxes represent the common identifier that is used for linking the two data types. One thing the figure shows, is that there is no combined dataset that includes the common identifier (participant ID) together with the full survey and social media data. Only parts of the the data from the two sources and/or derived variables are combined into the same dataset (which is used for the analyses). This reduced combined dataset used for the analyses is also what researchers can potentially share as "replication data" (King, 1995), given, of course, that the legal requirements and ethical considerations allow sharing the data included therein.

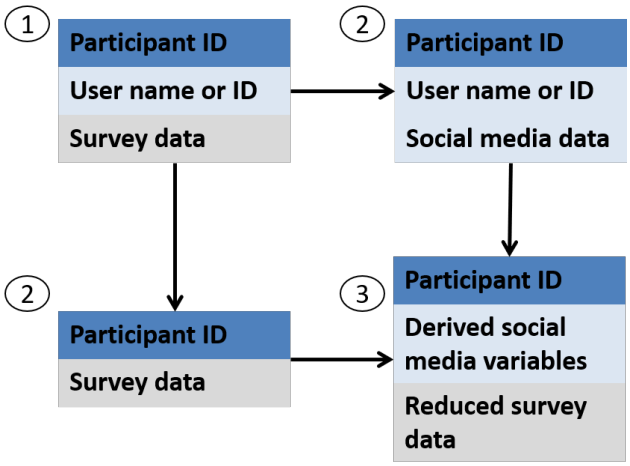

Figure 3: Proposed workflow for linking survey and social media data based on Sloan et al. (2020)

Recommendations for linking surveys with social media data

- Make an informed choice regarding the way you collect or access the social media data (e.g., via APIs or through data donation) as this determines how they can be linked with the survey data
- If you collect the data yourself, take into account that the order in which you collect the data influences the composition of the sample included in the linked dataset
- Be aware of and take into account the sampling bias associated with the social media data that you use
- Employ measures to eliminate or reduce the risk of mismatches or unmatchable cases between the data types (e.g., because of errors in self-reported user names)

Linking surveys with sensor data

Today a wide range of devices is available at relatively low costs to measure environmental factors, such as air pressure or temperature with built-in sensors. Many of the data that these sensors generate are also interesting for social scientists. Examples of such data include RFID (radio frequency identification) Chips (Elmer, Chaitanya, Purwar, & Stadtfeld, 2019) used to generate social networks or the measurement of environmental pollution in urban areas (Brunekreef & Holgate, 2002). In this section, we focuses on two of the most widely used and accessible devices for the collection of sensor data: smartphones and

fitness trackers. Using smartphone sensors and fitness trackers to collect data on behavior enables many novel research designs in the social sciences. Some of the many types of measures that these devices can provide that can be of particular interest for social scientists are physical and medical ones, such as acceleration or heart rates, or environmental ones, such as noise levels.

Users of devices like fitness trackers typically use them to monitor themselves or assess their physical performance. Apps that these people use to collect these data usually allow for additional insights that are also interesting for social science research. To link such data with survey data, researchers need respondents to complete an identification process, and find or develop methods for accessing their data. Notably, the data obtained may come in larger quantities and different formats than survey data. Hence, collecting data generated by smartphones and fitness trackers and linking them with survey data is a multistep process (see Figure 4).

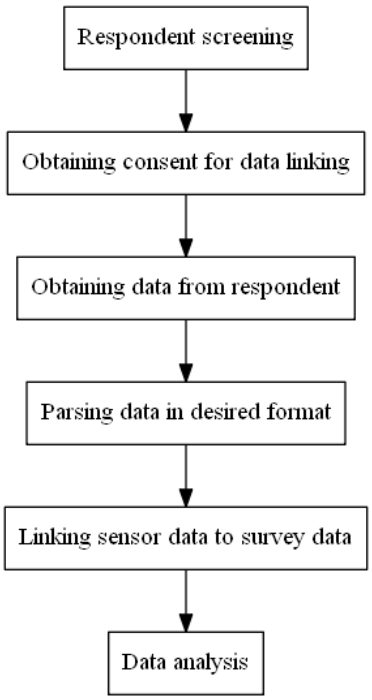

Figure 4: Sensor data linking process

Data generated by smartphone sensors and fitness trackers are stored in apps that come either with the operating system, such as *Apple Health* on *iOS*, or are provided by the manufacturer of the fitness tracker. In many cases, data can be linked between different apps in many cases. Getting access to data stored in a fitness app has the advantage that researchers only need to access one data source, which also allows more straightforward data analysis as the data are typically preprocessed and delivered in a similar format for every respondent. Nevertheless, researchers should investigate beforehand which devices and apps they want to include in their study as they differ drastically in measurement accuracy (Battenberg, Donohoe, Robertson, & Schmalzried, 2017) as well as with regard to how the data are stored and how they can be exported.

It makes sense to link data from smartphone sensors and fitness trackers with survey data on an individual level as smartphones are a highly personalized device and typically not shared between persons. As displayed in Figure 4, in the first step, respondents need to be screened to identify those who are using a device or app of interest. Then the identified respondents have to be asked for their consent to collect

and use the data. The next step depends mainly on the devices and apps which are used.

Generally, respondents have to make their data accessible to researchers. This access can happen through data donation (see the previous section on secial media data), meaning that respondents download their data and upload them to a server provided by the researcher(s) (Bietz, Patrick, & Bloss, 2019). In this case, it must be ensured that the connection to and the data stored on the server are encrypted and only accessible by the research team. Another option is that respondents allow access to their data via an API request (J. Wang, Coleman, Kanter, Ummer, & Siminerio, 2018). This option depends on the manufacturers of devices and fitness apps. Not every company allows third parties to access their users' data. For example, *Apple Health* allows users to export their data themselves, but does not offer an API accessible from a third device.

Similar to geospatial and social media data, data obtained from smartphone sensors and fitness trackers typically differ substantially from survey data in several regards. The data may be delivered as `.json` or `.xml` files which have to be parsed before the data can be analyzed, e.g., using the `jsonlite` package for `R` or the `json` library for `Python`. In addition, it is not uncommon that the dataset delivered by a single respondent can be over a gigabyte in size and contain several million observations for a single variable. Hence, researchers need to learn how to handle this kind of data if they want to link it with survey data. Finally, some variables, such as the step count of a person, might not be normally distributed. Such data often show long-tailed or Poisson distributions which means that appropriate analysis methods have to be chosen for them.

While using data from smartphones, fitness trackers (or other sensors) and linking them with survey data can provide many novel insights and research opportunities for social scientists, researchers should thoroughly plan all of the steps described above when collecting, processing, and analyzing these data. Due to the variety of smartphone and fitness tracker providers it can be challenging to collect enough data for research purposes using only one type of device or app. On the other hand, when collecting data from various devices, comparability between different datasets might be limited, and the complexity of the data collection and later analyses can increase. Finally, as with the other types of data discussed in this guideline, researchers have to consider that the data collected can be very sensitive, especially if they include geo-locations and further personal information, such as detailed medical information that might identify a respondent, such as a name, e-mail or home address.

Recommendations for linking surveys with sensor data

- Develop and test a a simple and reliable process to collect the sensor data (that minimizes respondent burden)
- Inform yourself about the apps and devices available, and make an informed choice on which to include and which not. Your decision will have a mayor impact on data quality and the analysis
- Use an effective screening procedure to identify respondents who belong to your target population
- Be aware of the biases arising from limiting your data collection to a specific group i.e. Apple users or smartwatch owners

Conclusion

One thing the data types discussed in this guideline - geospatial, social media, and sensor data - have in common is that they provide a host of opportunities for social scientists when combined with survey data. All of them come in data formats, e.g., shapefiles for geospatial data or `.json` files for sensor and social media data, that are typically less familiar to social scientists. They often do not come as structured tabular data and need extensive preprocessing, as the data are often not produced specifically for

research purposes. To access the data, researchers often need to make use of Application Programming Interfaces (APIs). Further, all three data sources tend to fall into the category of big data (at least from a social science perspective), meaning that their size tends to exceed that of typical survey datasets by orders of magnitude. Hence, even with modern computer technology, these datasets can be too large for use on a local computer which may force researchers move their analyses to more powerful servers, computer clusters or cloud computing services. In sum, working with these data requires social scientists to extend their methodological toolbox and skill sets.

When it comes to the linking process itself, all three data sources need a link between auxiliary data and survey data. This link is typically a geocoordinate in geospatial data, identifying a respondent as belonging to a specific geometry projected on the earth's surface. For the other two types of data, these are typically user names or user IDs for the platforms, devices or services in question. Hence, while there are numerous differences between the three data types we focused on, there are essential structural similarities in tn the general linking workflows. In order to link survey data to additional data using deterministic data linking, we need an identifier, which is typically provided by the respondents. This identifier is then used to link the survey and auxiliary data using a correspondence table. Notably, this simple-sounding process is subject to several obstacles. For one, researchers have to consider how they can reduce the burden associated with the data linking process. Making this process difficult for respondents' is likely to lead to non-compliance and refusal. This issue may be less relevant for data types like geospatial data where the information stems existing aggregate data but can be critical when using a linking process based on data donation (Bietz, Patrick, & Bloss, 2019).

Finally, all three data types discussed here are likely to contain sensitive information. This requires that researchers working with these data pay special attention to questions of data protection and privacy. The general data protection measures presented by Sloan, Jessop, Al Baghal, & Williams (2020) for Twitter data can also be applied to all of the types of data discussed in this guideline. In addition to storing the survey data and the additional media data separately and only combining what is needed for analysis, Sloan, Jessop, Al Baghal, & Williams (2020) suggest three other strategies for increasing data security: 1) data reduction, 2) data deletion, and 3) controlled access. For many analyses, the full raw data are not required. In such cases, reduced or aggregated data can be used. In practice, this means that only a subset of variables is combined and used. An additional option in this regard is to use aggregated or derived variables. Once derived or aggregated variables of interest are created, a very effective measure for securing data privacy is to delete the raw data. Of course, this limits the reproducibility and reduces the potential re-use value of the data. The latter is especially vital if researchers want to share their data. In that case, controlled access needs to be considered for distributing data beyond the researchers who collected the data and their direct collaborators (e.g., within a project). Many data archives, such as the *GESIS* data archive, offer different types and degrees of access control. Depending on the type of data, it may, for example, be an option to use different levels of access control for the full (raw) data and a reduced or aggregated replication dataset (created, e.g., for a specific publication). For sharing very sensitive data, such as geocoded survey data, researchers can make use of solutions for secure data access, such as the Secure Data Center at *GESIS*. While the specific legal and ethical aspects that need to be considered when researchers want to share their data always depend on the type of data (and how they were collected), some publications provide useful general guidance. For example, Schweers, Kinder-Kurlanda, Müller, & Siegers (2016) discuss solutions for geospatial data and several other publications regarding social media data sharing (e.g., Bishop & Gray, 2017; Weller & Kinder-Kurlanda, 2016; Williams, Burnap, & Sloan, 2017). While the specifics are likely to differ, these resources can also be consulted by researchers who want to share survey data linked with geospatial, social media, or sensor data.

General recommendations for linking survey data with other types of data

Although the specific steps that need to be taken and things that should be considered depend on the research interest and the types of data that are used, there are a couple of general aspects that researchers should take into account and address when they want to link surveys with other types of data:

1. Basic requirements and decisions
   - Decide whether you want to link data ex-ante (collect data for the purpose of linking) or ex-post (link existing datasets)
   - Decide which data you need for your research and how to access them: The type and format of the data determines how they can be linked
   - Pick or define a suitable common identifier for linking the datasets
2. Legal and ethical considerations
   - Check the legal compliance of your data collection and use (use institutional help/legal counselling if necessary and possible)
   - Consider the ethical aspects of your data collection and processing: Consult ethics review boards or existing guidelines (e.g., from relevant professional societies)
   - If possible, get informed consent from survey respondents for collecting and/or linking additional data
   - In the informed consent, offer detailed information regarding data storage and processing (without overwhelming participants with too much technical information)
3. Data processing, storage, and sharing
   - Keep the datasets (survey + additional data) separate as much as possible and only combine reduced or processed data as required for your analyses
   - Minimize the data: Only collect the data that you need for your research purposes, remove direct identifiers, and reduce indirect identifiers before analyzing, and especially before sharing or publishing the data
   - Consult with archives regarding the sharing of your linked data

Overall, despite the challenges in collecting and working with geospatial, social media, and sensor data, linking them with survey data is already demonstrating great potential for social science research. Given the required effort as well as the issues related to privacy and data protection, data linking is not yet widely used in the large survey programs that *GESIS* is involved in. Exceptions in the realm of georeferenced survey data are, for example, the *German Socio-economic Panel* (SOEP), the *German General Social Survey* (GGSS/ALLBUS), or the *German Longitudinal Election Study (GLES)*, which actively engage their users in linking of geospatial data. Collecting and linking sensor and social media data is even more challenging for the existing survey programs due to the associated legal issues (e.g., related to the ToS of companies that own the platforms/services), the sensitivity of the data, and the burden their collection can place on respondents when a data donation approach is employed. Nevertheless, the *GESIS Panel* is currently also exploring the options of collecting and linking social media and sensor data, and other data collection services for combined survey, sensor, and social media data are currently being discussed and planned at *GESIS* as well as elsewhere. For the time being, however, linking survey data with other data is something that researchers or research projects who need this data to answer specific research questions need to engage in themselves. We hope that this guideline can aid these researchers and projects in successfully doing so.

References

Al Baghal, T., Sloan, L., Jessop, C., Williams, M. L., & Burnap, P. (2020). Linking Twitter and survey data: The impact of survey mode and demographics on consent rates across three UK studies. *Social Science Computer Review*, *38*(5), 517–532. https://doi.org/10.1177/0894439319828011

Araujo, T., Wonneberger, A., Neijens, P., & de Vreese, C. (2017). How much time do you spend online? Understanding and improving the accuracy of self-reported measures of internet use. *Communication Methods and Measures*, *11*(3), 173–190. https://doi.org/10.1080/19312458.2017.1317337

Battenberg, A. K., Donohoe, S., Robertson, N., & Schmalzried, T. P. (2017). The accuracy of personal activity monitoring devices. *Seminars in Arthroplasty*, *28*, 71–75. Elsevier.

Bietz, M., Patrick, K., & Bloss, C. (2019). Data donation as a model for citizen science health research. *Citizen Science: Theory and Practice*, *4*(1).

Bishop, L., & Gray, D. (2017). Ethical Challenges of Publishing and Sharing Social Media Research Data. In K. Woodfield (Ed.), *Advances in Research Ethics and Integrity* (Vol. 2, pp. 159–187). https://doi.org/10.1108/S2398-601820180000002007

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

Brunekreef, B., & Holgate, S. T. (2002). Air pollution and health. *The Lancet*, *360*(9341), 1233–1242.

Bruns, A. (2019). After the 'APIcalypse': Social media platforms and their fight against critical scholarly research. *Information, Communication & Society*, *22*(11), 1544–1566. https://doi.org/10.1080/1369118x.2019.1637447

Crowder, K., & Downey, L. (2010). Inter-Neighborhood Migration, Race, and Environmental Hazards: Modeling Micro-Level Processes of Environmental Inequality. *American Journal of Sociology*, *115*(4), 1110–1149.

Elmer, T., Chaitanya, K., Purwar, P., & Stadtfeld, C. (2019). The validity of RFID badges measuring face-to-face interactions. *Behavior Research Methods*, *51*(5), 2120–2138.

Freelon, D. (2018). Computational Research in the Post-API Age. *Political Communication*, *35*(4), 665–668. https://doi.org/10.1080/10584609.2018.1477506

Haim, M., & Nienierza, A. (2019). Computational observation: Challenges and opportunities of automated observation within algorithmically curated media environments using a browser plug-in. *Computational Communication Research*, *1*(1), 79–102. https://doi.org/10.5117/CCR2019.1.004.HAIM

Halavais, A. (2019). Overcoming terms of service: A proposal for ethical distributed research. *Information, Communication & Society*, *22*(11), 1567–1581. https://doi.org/10.1080/1369118X.2019.1627386

Jünger, S. (2019). *Using Georeferenced Data in Social Science Survey Research. The Method of Spatial Linking and Its Application with the German General Social Survey and the GESIS Panel*. Retrieved from 10.21241/ssoar.63688

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

Kinder-Kurlanda, K., Weller, K., Zenk-Möltgen, W., Pfeffer, J., & Morstatter, F. (2017). Archiving information from geotagged tweets to promote reproducibility and comparability in social media research. *Big Data & Society*, *4*(2), 205395171773633. https://doi.org/10.1177/2053951717736336

King, G. (1995). Replication, Replication. *PS: Political Science and Politics*, *28*(3), 444. https://doi.org/10.2307/420301

Prior, M. (2009). The Immensely Inflated News Audience: Assessing Bias in Self-Reported News Exposure. *Public Opinion Quarterly*, *73*(1), 130–143. https://doi.org/10.1093/poq/nfp002

Scharkow, M. (2016). The accuracy of self-reported internet usea validation study using client log data. *Communication Methods and Measures*, *10*(1), 13–27. https://doi.org/10.1080/19312458.2015.1118446

Schweers, S., Kinder-Kurlanda, K., Müller, S., & Siegers, P. (2016). Conceptualizing a Spatial Data Infrastructure for the Social Sciences: An Example from Germany. *Journal of Map & Geography Libraries*, *12*(1), 100–126. https://doi.org/10.1080/15420353.2015.1100152

Sen, I., Flöck, F., Weller, K., Weiss, B., & Wagner, C. (2019). *A Total Error Framework for Digital Traces of Humans*. Retrieved from http://arxiv.org/abs/1907.08228

Sloan, L., Jessop, C., Al Baghal, T., & Williams, M. (2020). Linking Survey and Twitter Data: Informed Consent, Disclosure, Security, and Archiving. *Journal of Empirical Research on Human Research Ethics*, *15*(1-2), 63–76. https://doi.org/10.1177/1556264619853447

Sluiter, R., Tolsma, J., & Scheepers, P. (2015). At Which Geographic Scale Does Ethnic Diversity Affect Intra-Neighborhood Social Capital? *Social Science Research*, *54*, 80–95. https://doi.org/10.1016/j.ssresearch.2015.06.015

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

Thorson, K., Cotter, K., Medeiros, M., & Pak, C. (2019). Algorithmic inference, political interest, and exposure to news and politics on Facebook. *Information, Communication & Society*, 1–18. https://doi.org/10.1080/1369118x.2019.1642934

Wang, J., Coleman, D. C., Kanter, J., Ummer, B., & Siminerio, L. (2018). Connecting smartphone and wearable fitness tracker data with a nationally used electronic health record system for diabetes education to facilitate behavioral goal monitoring in diabetes care: Protocol for a pragmatic multi-site randomized trial. *JMIR Research Protocols*, *7*(4), e10009.

Wang, Z., Hale, S., Adelani, D. I., Grabowicz, P., Hartman, T., Flöck, F., & Jurgens, D. (2019). *Demographic Inference and Representative Population Estimates from Multilingual Social Media Data*. Presented at the WWW '19: The World Wide Web Conference. https://doi.org/10.1145/3308558.3313684

Weller, K., & Kinder-Kurlanda, K. E. (2016). A manifesto for data sharing in social media research. *Proceedings of the 8th ACM Conference on Web Science - WebSci '16*, 166–172. https://doi.org/10.1145/2908131.2908172

Williams, M. L., Burnap, P., & Sloan, L. (2017). Towards an Ethical Framework for Publishing Twitter Data in Social Research: Taking into Account Users' Views, Online Context and Algorithmic Estimation. *Sociology*, *51*(6), 1149–1168. https://doi.org/10.1177/0038038517708140