Conceptualizing a Spatial Data Infrastructure for the Social Sciences: An Example from Germany

STEFAN SCHWEERS, KATHARINA KINDER-KURLANDA, STEFAN MÜLLER, and PASCAL SIEGERS
GESIS–Leibniz Institute for the Social Sciences, Cologne, Germany

There is an increasing demand for spatial data within the social sciences as such data promise to contribute to a better understanding of how the concrete living environment influences individuals’ attitudes and behaviors. Spatial data can complement data from surveys; however, as of yet there is no spatial data infrastructure for the social sciences in Germany that facilitates merging of spatial data with survey data. This article explores avenues for creating such a spatial data infrastructure (SDI) for the social sciences in Germany. We examine the role that librarians, archivists, and curators can play in offering spatial data infrastructure services for social scientists, and show the scope and nature of necessary tasks in areas such as harmonization, archiving, dissemination, and user support. As the case of Germany is similar to that of other European countries, we expect our results to be helpful in the creation of SDIs in other countries as well. This article suggests technical and organizational measures for merging survey data and spatial data in compliance with German privacy legislation. Measures for storage and dissemination of enriched survey data, for example by providing restricted access to the data in a secured environment, also are explored.
INTRODUCTION

In recent years, an increasing demand for a spatial data infrastructure (SDI) emerged in the social sciences. Indeed, SDIs are not fundamentally new. Their implementation and integration within existing data infrastructures, such as archives or repositories, however, require technical and conceptual work. This article describes these considerations and the conceptualization of an SDI in a georeferencing project, which was initiated at the Data Archive for the Social Sciences (DAS) at GESIS, Cologne, in 2015.1 Because the archive infrastructure at the DAS is comparable to other archive institutions, the steps outlined in this paper can be followed by other research data infrastructure institutions.

The term “spatial data infrastructure” (SDI) stands for a “framework that consists of institutional arrangements, policies, and technologies that would create a conducive environment for the exchange of geographic information related resources in order to create a better information sharing community” (Tumba and Ahmad 2014, 85). At the global,2 continental3 and national levels,4 SDIs have been developed that allow users to visualize data and, in the case of some SDIs, to download spatial data.

A considerable potential of spatial data, on the other hand, is as of yet mostly untapped, namely the opportunities for research provided by the merging of spatial data with survey data. Such merging allows researchers to analyze individual human behavior with attention to its spatial dimension. However, there is currently no spatial data infrastructure in Germany that allows research projects to merge available geodata with social science survey data. Developing a service for harmonizing, merging, archiving, and disseminating spatial data and survey data is, though, the objective of the project that we describe in this article, using the example of environmental noise data. Additionally, the SDI is intended to give researchers the option to upload and share high-quality geodata.

In this project, a spatial component will be added to survey data from the archive's research data holdings by merging these survey data with spatial data provided by the national statistics office’s existing geoportals. The challenge is to embed the merged data within the archiving and data provisioning workflows as they are described by the Open Data Archival Information System (OAIS) reference model used by many data archives (Vardigan and Whiteman 2007). In this article, we therefore describe the workflow of georeferencing survey data, from collecting spatial data, data merging, and data archiving, all the way through to the dissemination of the merged datasets.
The SDI itself is not yet established. However, the contribution of this article is the presentation of an innovative way of providing access to data resulting from the merging of spatial data and survey data. The procedures we describe suggest solutions to all major hurdles for the use of spatial data encountered by social scientists using the example of environmental noise data. Moreover, our solutions are embedded into an existing social science data archive and therefore may help other data archives develop similar services for their national communities.

First, we show examples of the use of spatial data in the social sciences and explain how merging spatial data with survey data improves the analytical potential of survey data. The example of environmental noise introduced therein will be used as a demonstration of the concepts explained in the following sections. The next section addresses legal barriers for merging spatial data with survey data, explaining why the use of spatial data is still limited in the social sciences although studies on human attitudes and behavior would benefit from a consideration of including spatial information. Each of the following three sections then explains the functions a social science SDI needs to fulfill and how we suggest providing the following services:

- First, guidance needs to be provided to users as the creation and provision of geodata in Germany is strongly fragmented and it is difficult for researchers to gain an overview of available data.
- Second, researchers need support in merging spatial data with survey data because most social scientists are not familiar with geoinformation technologies.
- Finally, the archiving and dissemination of georeferenced survey data has to be performed in accordance with existing infrastructures and standards for the social sciences.

DEMAND FOR SPATIAL DATA IN THE SOCIAL SCIENCES

The use of spatial data is not new to the social sciences. Already at the beginning of the 20th century, French political scientist André Siegfried included geological characteristics of the country, such as soil composition, in his studies of electoral outcomes (Siegfried 1913). However, the social sciences have not systematically developed a spatial approach to human attitudes and behaviors since Siegfried’s study. Currently, spatial data in the social sciences are often primarily used for visualization purposes. Monnot (2013), for instance, maps the density of religious communities in Switzerland using geocoded data of all religious communities in Switzerland, whereas Schräpler (2009) computed the density of welfare recipients in the German federal state of North Rhine-Westphalia using address data from the German Federal Employment Agency.
An analytical use of spatial data from a social science perspective resulted from a body of research on environmental inequalities that started in the United States but also gained some momentum in European research (Saib et al. 2014). The most relevant data for this research are information about the exposure to noise and air pollution. Spatial data are used to study inequalities in the exposure to air pollution or traffic noise by correlating these data with aggregate data on the demographic composition of the population at the local level (Zwickl, Ash, and Boyce 2014) or at the level of Census Block Groups in the United States (Clark, Millet, and Marshall 2014). Correlations of aggregate data, however, bear the risk of an ecological fallacy. To avoid this, spatial data have to be linked with data about individuals, for example by merging spatial data from the surveys. This was done in a study by Diekmann and Meyer (2010). They use spatial data on air pollution and noise provided by the Swiss government using a 10-meter grid that they merge with geocoded data from a survey on the perception of pollution and noise, including socio-demographic variables. The results show that foreigners have a significantly higher exposure to ecological strain than Swiss citizens, whereas the effects of income and education are small or nonsignificant.

Several authors deplore that analyses of survey data largely ignore the spatial dimension of human behavior (e.g., Diekmann and Meyer 2010; Chhetri and Stimson 2014). Models that do not account for the situative context of behavior are incomplete and potentially biased because relevant predictors are not included in the statistical analysis. Moreover, merging spatial data with survey data is difficult because, for the mostly legal reasons we discuss in the next section, individual level data in general do not include geocodes. MacKerron and Mourato (2013) tried to overcome this problem using Global Positioning Systems (GPS) in smartphones to locate respondents of surveys in the geographical space. Merged with data on land cover, this enabled the authors to estimate the impact of individuals’ geographical location on momentary subjective wellbeing. Although this study used highly innovative data collection techniques, it lacked a consistent sampling procedure, which casts doubt on whether the findings can be generalized to the overall population.

Two examples from social sciences research illustrate how the use of spatial data might improve their research outcomes. First, social scientists recognized that the characteristics of the neighborhood play an important role for individual outcomes (Dietz 2002), such as deviant behavior (Sampson, Raudenbush, and Earls 1997; Sampson, Morenoff, and Gannon-Rowley 2002; Oberwittler and Wikström 2009) and educational achievement (Ainsworth 2002; Helbig 2010). Studying effects of neighborhoods is possible only if official statistics provide data at the appropriate level of detail and if respondents can be located within the neighborhoods. The latter is particularly
challenging due to privacy protection regulations in many countries. Using spatial data and Geographic Information System (GIS) techniques, inclusion of spatial data would improve the study because it allows for modeling context effects independently from administrative units (Crowder and South 2011). For instance, Schröpfer’s (2009) estimation of the density of social welfare recipients (see above) is used to measure the ‘burden’ of school neighborhoods. The measurement he suggests is finer-grained than data available from official statistics (such as census data). There are several examples of studies where spatial data at the neighborhood level provided by commercial suppliers were merged with survey data in Germany.

Voigtländer, Berger, and Razum (2010) showed that individual health is partly determined by the socio-economic composition of the neighborhood and Knies (2010) examined whether changes in the average income of the neighborhood affect life satisfaction. Sager (2012) showed that strong residential isolation of the principal migrant groups in Germany is observable at the neighborhood level and Dittmann and Goebel (2010) studied the impact of the neighborhood on individual life satisfaction in Germany. These studies all use data from the German Socio-Economic Panel (SOEP) that provides access to georeferenced data within a secure environment.

A second example concerns the study of ecological strains on individual health outcomes. Many studies revealed that exposure to noise in particular has a detrimental impact on individuals’ health outcomes (Passchier-Vermeer and Passchier 2000). Two different methods are frequently used in these studies. On the one hand, there are large-scale population surveys based on random sampling techniques that use self-reported ecological strain (Kohlhuber et al. 2006) because measuring exposure to noise and air pollution for each participant is too expensive if respondents are spread over the whole country. Subjective measures of exposure might vary depending on individuals’ characteristics (e.g., sensitivity to noise and perceptions of pollution) and therefore might be biased (Bodin et al. 2012). On the other hand, there are studies of small local samples, including measurements of exposure for each respondent (Öhrström 2004; Bocquier et al. 2013). The disadvantage here is that only small populations can be studied and data collection is most often limited to urban areas. One possibility for overcoming these shortcomings is to merge spatial data on exposure to environmental strains with survey data on health outcomes, following the example from the study by Diekmann and Meyer (2010) on environmental inequality in Switzerland mentioned above.

Data on environmental noise are available for Germany and most other European countries because the collection and publication of the data are compulsory for member states of the European Union (c.f. EG Guideline 2002/49/EG). The EG Guideline ensures a uniform and comprehensive data collection process. More questions on the effects of ecological strain on in-
Individuals could be answered if the data on air pollution and noise were to be merged with individual level data. For instance, more research is needed about coping mechanisms of individuals who suffer from ecological strain. These are substantial arguments for why georeferencing survey data with environmental noise data is promising. Paradoxically, the data are used neither for research on environmental inequality nor for studies on health outcomes in Germany. The reason for this is that there are significant technical burdens for using the data that are related to the federal governance structure in Germany because the data are collected by many different government actors and are not provided in a harmonized format (see section “Identifying and Harmonizing Available Spatial Data” for details). A spatial data infrastructure for the social sciences could help overcome these burdens. Data on noise exposure is particularly suited as an example because it is unequally distributed in space. Therefore, aggregation of noise exposure at different administrative levels (as it is done for census data) or spatial units is not appropriate for assessing effects from noise exposure on individual outcomes.

Many more applications for merging geocoded survey data with spatial data might also be relevant. Diekmann and Meyer (2010) in particular mention the computation of distances to public transport facilities, schools, childcare, and shopping facilities and the quality of living environments (Lersch 2013).

Other approaches adding value to survey data with spatial data focus on visualizing survey data within a spatial framework and applying techniques of spatial data analysis (Chhetri and Stimson 2014). The UK Data Archive (UKDA) developed a tool for linking the regional information included in the surveys archived at the UKDA (for example, county or province of residence) to digital boundary files from the UK to facilitate the visualization of survey data. However, access to low-level geographic data is restricted due to privacy legislation. Survey data might also improve the use of spatial data. Survey data merged with spatial data on noise exposure might show how individuals perceive ecological strain.

To summarize, merging survey data with spatial data enables social scientists to model the spatial dimension of human behavior. Improving the linking of survey data and geographical information promises to result in a substantial improvement of explanatory models in the social sciences. Therefore, it is safe to assume that social scientists will more frequently make use of publicly available data and start to produce relevant spatial data themselves. At the same time, however, there is no infrastructure available that supports researchers in (1) harmonizing and documenting spatial data and (2) merging spatial data with data from other sources, particularly survey data. This lack of infrastructure increases the costs of investing in the creation and analysis of spatial data. The construction of a spatial data infrastructure for the social sciences is required to facilitate access to, and the use of,
spatial data. In addition, such an infrastructure can support researchers in addressing the data protection challenges of merging spatial data with survey data.

LEGAL BARRIERS TO THE USE OF SPATIAL DATA IN SOCIAL SCIENCE RESEARCH

All countries in Europe have legal frameworks protecting individuals from the abuse of private data. Data protection legislation restricts the use of spatial data for research in those cases where data on the individual level (survey data) is merged with spatial data. When conducting survey research in Germany, direct identifiers (such as respondents’ names and addresses) must not be stored in the same place as the data collected in the survey (including socio-demographic characteristics of the respondents and data about attitudes and behavior). The separation of direct identifiers and survey responses ensures that participants remain anonymous and that contact information can be stored for follow-up surveys (Metschke and Wellbrock 2002). The contact details of respondents that have been collected for sampling but who have decided not to participate in the survey or have not responded must be deleted. In general, survey research institutes are urged to delete addresses as soon as possible unless respondents have given informed consent to the storage of their contact details for follow-up questions. Geographical coordinates of individuals’ addresses are considered to be direct identifiers and therefore cannot be offered for scientific use in the same place as data collected in the survey.

Data disseminated for the purpose of secondary analyses usually is anonymized in order to protect individuals’ privacy. Datasets offered to researchers must be changed (by deleting or coarsening information) to such a degree that a de-anonymization of the data would require a “disproportionate effort.” De-anonymization is particularly a risk for respondents who show a rare combination of socio-demographic characteristics (Wirth 2013). For example, it would be fairly easy to identify a person working as a lawyer, has six children, and is living in a small town if the name of the municipality was known; anonymization in this case cannot be achieved by removing the name of the municipality. In general, the more information that is added to a dataset about an individual, the more likely it is that the individual can be identified. The growing availability of various information about individuals on the Internet and in social media makes reidentification even more likely. In practice, anonymization of survey data often is achieved by removing more detailed geographical and biographical information. For example, only province names could be provided in the data set example above (and not municipalities) and the number of children in those cases where there are
more than three could be given as “four or more.” Access to more comprehensive and detailed datasets may still be possible for researchers but always requires specific contractual arrangements and analysis within a secured environment such as a safe room employing various organizational, contractual, and technical controls to prevent de-anonymization. Safe rooms and other secure data access services are thus particularly interesting for providing spatial data because they allow access even to data containing fine-grained geographical information that is required for many of the new research opportunities mentioned above.

For example, access to the spatial data merged to the German SOEP mentioned in the examples in the previous section is only possible in the safe room at the German Institute for Economic Research in Berlin. Moreover, researchers can only use spatial data for their analyses that have been merged to the data by the primary investigators of the SOEP, which further limits the potential use of the data.\(^8\)

In our example of environmental noise data, the imperative of data protection made data research and data preparation an effort that had to contend with a considerable time-pressure element. Because direct identifiers needed to be deleted as soon as possible, georeferencing of survey data had to be performed immediately after sampling and data collection. Adjustments would not be possible after deletion of contact details and geocodes. The practice of deleting addresses and storing data in different places to ensure data protection thus led to time-pressure and having to resolve certain issues very early on. Data quality issues such as formats, errors, etc. (see next section) had to be solved and decisions about, for example, coarsening of respondents’ coordinates in the form of grids had to be made right away.

A spatial data infrastructure has to provide solutions for using spatial data combined with survey data in compliance with the legislation in Germany and other European countries. In the following section, the various steps of setting up a service that addresses the data protection challenges will be explained in detail.

### IDENTIFYING AND HARMONIZING AVAILABLE SPATIAL DATA

A spatial data infrastructure must first facilitate access to spatial data. This can be done by a specialized spatial data librarian who acts as an interface between data producers and researchers. In Germany and most other countries of the European Union, spatial data are provided either by government agencies (in particular statistical offices) or by private business companies. Both types of spatial data are used for research purposes.

Commercially supplied spatial data concerns mostly data used for targeted marketing; in Germany the companies microm Consumer Marketing,\(^9\) infas 360,\(^10\) and GfK\(^11\) are competitors in this market. Typical kinds of spatial
data provided by these commercial suppliers are, for example, characteristics of neighborhoods, such as housing quality and household income. The problem with using spatial data from commercial suppliers is that in most cases no information is published about the processes of gathering and generating the data. Business suppliers collect spatial data from various sources and merge them into spatial indices (e.g., purchasing power). Some of the sources used for generating spatial data are publicly available survey data. In many cases, the data are extrapolated to achieve complete geographic coverage of Germany. The algorithms for extrapolation are part of the business secrets of commercial suppliers and are not known to researchers using the produced datasets.

As noted above, a second source for spatial data is government agencies. Spatial data from this source have the advantage that the process of data generation is transparent so that researchers can judge the quality of the data, including its validity and reliability. Moreover, most public spatial data are open access and meet the high quality criteria for data from official statistics.

The use of spatial data collected by public agencies is embedded within the INSPIRE Directive of the European Union. This directive aims to stimulate open access to spatial data and the interoperability of spatial data by developing a harmonized metadata scheme for data documentation. Reacting to the open access policy concerning spatial data of the European Union, public administrations of all governance levels cooperate to establish a single online catalogue for available public spatial data in Germany. The repository counts 120,000 entries covering many different topics, such as school population, election results, and vital statistics. The portal enables users to find 1,800 public data providers. There are, however, major inconveniences in the use of these data for social scientists. First, although the catalogue provides a scheme for metadata within the INSPIRE directive, for most data sets the information is incomplete. Most noteworthy, details on how to access the spatial data are lacking in the metadata. Second, the catalogue does not give information on which data are available for the whole of Germany, even in those cases where data are produced by subnational government agencies. Thus, the spatial data infrastructure in Germany is still in its very beginning stages.

A major disadvantage of using public spatial data is that the federal structure of governance in Germany leads to a fragmented production of spatial data. The responsibility for data collection can vary between institutions on the national, state, or local level. A comprehensive supply of spatial data exists only for those policy domains where local governments are required to deliver data to the national federal office. In particular, spatial data on welfare infrastructures (such as child care and health facilities) are collected only at the local level because municipalities are responsible for providing these services in Germany.
Moreover, data provided by the federal statistical offices at local (municipalities) and regional (counties) levels (see endnote 6) lack spatial references; thus, an analysis of the data using geographical software requires adding georeferences to the data using spatial basis data (e.g., Esri shapefiles, see below). Even at the state level there is a high degree of fragmentation. There are seventeen public statistical offices in Germany, one for each of the sixteen states and then the federal office. Most spatial data are collected by the various nonfederal statistical offices. In addition, several national agencies collect their own spatial data.

Many data exist at the substate level. For example, some states publish maps of existing health infrastructures, whereas in other states these data are published at the municipality level. Consequently, in Germany one can see that there is a huge amount of spatial data that are publicly available for free, however, these data are often fragmented and therefore incomplete. In addition, spatial data offered by the various government institutions are not available in the same formats (i.e., they are not harmonized) and are not documented within a consistent metadata scheme.

The provision of data on environmental noise in Germany perfectly illustrates the fragmentation described above. Despite the fact that there is an existing EU guideline for collecting and publishing data on environmental noise, access to this data remains difficult. Federal states or single municipalities publish the results as maps in the form of simple PDFs, geospatial and geoPDFs, or as interactive web services. A primary issue is that there is no single access point to the data where researchers can download them in a coherent format. The federal states and municipalities collect the data and deliver them to the Federal Agency of Environment, which stores them in a data repository run by the European Environment Agency (EEA) and its European Environment Information and Observation Network (EIONET). This repository, called the Central Data Repository (CDR),\textsuperscript{15} collects, among other data, all environmental noise data deposited by the EEA member countries, including Germany, under a Creative Commons License (CC-BY).

However, neither the German nor the EEA harmonize and merge the data delivered by the states and municipalities.\textsuperscript{16} This means that the data cannot be easily used. First, there are 525 Esri files that can be downloaded through the repository that contain polygon vector layer data.\textsuperscript{17} Having said this, 32 files are in fact line vector layers, which contain a huge amount of dangling and open lines, especially in the areas of administrative borders. As a consequence, coordinates between two of these lines are not attributed with environmental noise values in the form of decibels or decibel ranges. To conduct the georeferencing process, however, the area between the two lines must be completed with a corresponding value. Otherwise, it is simply not possible to assign noise values to respondents that live in such areas. The second issue we faced with the data obtained from the CDR are missing georeference system definitions. This means that the projection of single
shapefiles related to the whole set of shapefiles for which we targeted an EPSG:25832 definition was not possible in the first place. This issue affected 78 of 525 shapefiles. The third and last challenge concerned missing data for whole federal states or municipalities. These data are either nonexistent in the repository or they exist in a format, such as simple PDF maps, that are not georeferenced. We observed this problem in one federal state and in eighteen municipalities.

It is important to note that the German Federal Agency for Environment only serves as an agent in depositing the data to the CDR. All data preparation, despite the naming conventions of single file names, is initially done by the authorities of the federal states or municipalities. Different measures to address the issues described above are conceivable.

We came up with several approaches to answer the challenges outlined above. Relating to the issue of dangling and open lines vector data, we observed mixed results either by using automatic snapping algorithms from common GIS software, such as ArcGIS (Esri 2015) or QGIS (QGIS Development Team 2015), or by doing it manually within the GIS. None of them were satisfying or scalable. At the same time, a fix for the missing georeference system definitions issue was determined after checking every affected Esri shapefile for an appropriate reference system. This was done by an automatic routine written in the statistical programming language R that can be fed with external data tables, such as a collection of common reference systems and the coordinates of affected shapefile regions (R Core Team 2015).

Eventually, we researched and contacted all states and municipalities whose data still were nonconforming after our attempts, or were missing in the first place, in order to acquire or re-acquire the data, respectively. As it turned out, these actions were successful, and as of August 2015, we have received data from twelve municipalities, while five requests are still pending and one municipality is still in the process of data collection.

Figure 1 shows the data types available after consulting the CDR and the data we received on request from the responsible authorities. It contains a visualization of the line and polygon vector data from the CDR and raster data we received from some authorities. Raster data may be more exact and also more convenient for georeferencing purposes. On the other hand, because we preferably targeted a harmonious data structure for the environmental noise data, and because we had only raster data from few municipalities, we finally had to reclassify them into the classes of the polygon attributes (classes from 50 to 54 decibels, 55 to 59 decibels, and so forth). The example of environmental noise data shows that using data requires considerable effort due to the need to harmonize the data. One important function of a social science SDI would therefore be to support researchers in this task because they are generally not experts on different formats of spatial data.
Another function of a spatial data infrastructure for the social sciences would be to facilitate the reuse of available spatial data. Facilitating reuse primarily concerns data collected by government agencies and data created for the purpose of scientific research. For these data, the quality of the data can be assessed because the process of data generation is transparent. Moreover, publicly available spatial data can be used without additional costs and can be disseminated for secondary use, which is usually not the case for data from commercial suppliers. Specialized data librarians working with a social science SDI would act as mediators between government agencies and the spatial data needs within the social science research community by curating and disseminating the data with the tools researchers use (see last section). This is particularly true when government agencies that initially collected the data for purposes other than specific disciplines’ scientific usage do not have the resources to harmonize the data. Environmental noise data are only one example among many.
The first element of a social science SDI, therefore, consists of a database that records which spatial data are available for social science research and how access to the data is possible. Documentation of the data contains, at a minimum, information about the following items:

1. Content of available spatial data;
2. Geographical coverage;
3. Temporal reference of spatial data;
4. Producer/owner of the data;
5. Format of the spatial reference included in the data.

The last point is important because it addresses the need to harmonize spatial references before merging spatial data from different sources, as described above. Moreover, the database should contain information about shapefiles for harmonization and about spatial data for merging with survey data.

To summarize, the federal governance structure in Germany leads to a situation in which the availability of spatial data depends on which institutions at the different levels (regional and state) are responsible for collecting and providing data in a specific domain or about certain topics. Although there are national data agencies responsible for environmental data that collect spatial data for the entire national territory, data on health and welfare infrastructures are available only at local and regional levels. Even if data are available for all or almost all territories, the data have to be harmonized and merged for use by (social science) researchers. Harmonizing available spatial data is therefore a major mission for an SDI.

MERGING SPATIAL DATA WITH SURVEY DATA

At the beginning of this article we showed that merging spatial data with data from social science surveys provides many opportunities for improving explanatory models. Therefore, another core element for a spatial data infrastructure for social science applications would be a tool for merging spatial data with survey data. The survey data, including the coordinates of the respondents, spatial basis data in the form of shapefiles, and spatial data, have to be merged. In this section, we sketch how a tool for merging spatial and survey data should be conceptualized within a spatial data infrastructure for the social sciences. The objective is to provide a service that is user-friendly and in compliance with German data protection legislation. The tool (1) creates transparency and (2) reduces costs for georeferencing in research projects that produce or re-use survey data.

Transparency is created by recording each step in the process of georeferencing and merging the steps with the help of proper documentation
approaches. This is accomplished by making the scripts and routines used available for future reuse, as well as attributing the steps taken and the final “product” with appropriate metadata. The marginal costs for future georeferencing projects are then reduced by the work already done in past projects.

Georeferencing of survey data requires information about the location of participants in a study. This is the case for face-to-face surveys, which is the standard mode of collection for many high quality social science surveys. Postal addresses can then be used to locate individuals within space. This task is performed with the help of a GIS. Because GIS technologies work with coordinates, addresses have to be converted into coordinates. There are several commercial and noncommercial providers offering conversion services for geographical references in Germany. Examples of commercial providers are the Internet companies Google and Yahoo. For reasons of data protection, conversion for scientific purposes cannot be commissioned to commercial providers who, as stated in their terms of service, often claim access to addresses or even store the addresses being converted into coordinates (cf. Google Terms of Service). Due to German data protection requirements, a commercial service provider must not access or store the original survey respondents’ contact details.

For this reason, research projects in Germany can rely on geocoding services provided by the German Federal Office for Cartography and Geodesy (BKG). This office has accurate and reliable addresses and coordinates available for every building in Germany, and allows the conversion without relying on commercial service providers. Coordinates are added in a UTM32 projection to a comma-separated values (CSV)-file listing the addresses and a goodness of conversion is estimated for each address. Table 1 shows the structure of the comma-separated values (CSV)-file required for the conversion.

Problems do occur if addresses are incomplete or faulty. In this case, the conversion into coordinates fails. Rules for handling incorrect addresses have to be developed based on the purpose of each research project. In the future, it may be possible to use GPS technology to automatically generate coordinates for respondents in face-to-face interviews and thus reduce the overall error rate of converting addresses into coordinates. The X- and Y-coordinates are added to the CSV-file. The respondent’s address now is no longer of interest, but the coordinates are added to the survey data.

Figure 2 illustrates a fictional example of respondents’ coordinates mapped to environmental noise data as polygon layers. Area information

| Respondent ID | Street    | House Number | Postal Code | Town            |
|---------------|-----------|--------------|-------------|-----------------|
| 1             | Pattern   | 1            | 12345       | Pattern City    |

TABLE 1 Structure of CSV file required for conversion of addresses
from the spatial data is added to the individual characteristics of each participant in a survey (Meyer and Bruderer Enzler 2013). In this case, participants A and C are both assigned a decibel range value of 70, whereas B is assigned a value of 75.

As part of the implementation of an SDI at the DAS, environmental noise data was merged with survey data from the German General Social Survey 2014 (ALLBUS) (GESIS–Leibniz Institute for the Social Sciences 2015), which constitutes a social monitoring of trends in attitudes, behavior, and societal change in the Federal Republic of Germany. Furthermore, it contains, among others, questions on subjective wellbeing and noise burden that are interesting for comparisons with actual measured noise exposure. In addition, a set of socioeconomic spatial “standard” variables was added, as shown in Table 2, using the example of unemployment rate. Table 2 also outlines the structure of the resulting merged data file.

Referring to the data model for spatial data suggested by Haining (cited by Stimson 2014), Table 2 shows the cases in the rows and the variables in the columns. The table includes the data of three sources. First, the data set comprises the data referring to the individual participants in a survey. The coordinates stem from converting addresses into coordinates, and they constitute the spatial reference for the survey data. Variables V1 to Vn were collected through the survey. Second, the spatial base data contain harmo-
nized shapefiles with boundaries. These allow assigning individuals to specific administrative units or grids. Third, the contextual spatial data includes information about respondents’ living environments. The spatial base data act as a pivot between respondents’ spatial references and the data about the environment. Moreover, the data model contains a temporal reference for each of the three categories of data. The temporal reference is needed to ensure that spatial base data and contextual data refer to the same point in time as the data from the survey. The temporal reference is particularly needed if time-series of data are constructed. The example from Table 2 shows two measurement time points in the survey data, 2012 and 2014.

The dissemination of survey data enriched with geocodes needs to comply with existing data protection legislation and to fit into existing infrastructures for data dissemination (for example, data catalogues). As outlined earlier, two main data protection aspects need to be considered for the dissemination of spatial survey data. First, direct identifiers (i.e., exact geolocations) and other information about the participants (i.e., survey responses) need to be kept separate to prevent direct reidentification of individual respondents. Second, access to datasets containing no direct identifiers but enough information about participants to permit reidentification of individuals must be secured to prevent such reidentification.

A functional spatial data infrastructure for the social sciences needs to address both aspects. To address the first issue, direct identifiers in the data need to be removed in preparation for archiving and dissemination. Such direct identification information is at least addresses, coordinates, territorial limits, and ID areas. In the flow sheet in Figure 3, this step is shown under “removing identifiers.” These identifiers are not necessary for data analysis, and their removal complies with legislation requiring that these data are not stored together with the survey data. One possible way to address the second issue of reidentification is employing existing infrastructures for secure access (see next section).

| TABLE 2 Structure of merged data |
|---------------------------------|
| **Spatial survey data** | **Spatial base data** | **Contextual spatial data** |
| Respondent ID | Coordinate V1–Vn | Area ID | Polygon | Time | Unemployment rate | Noise pollution (dB) | Time |
| 1 | (X, Y) | 2012 | 1 | Shape 1 | 2012 | 6.4 | 50–54 | 2012 |
| 2 | 2012 | 2 | Shape 2 | 2012 | 9.3 | 55–59 | 2012 |
| 3 | 2012 | 1 | Shape 1 | 2012 | 6.4 | 50–54 | 2012 |
| 4 | 2012 | 4 | Shape 4 | 2012 | 14.0 | 60–64 | 2012 |
| 1 | (X, Y) | 2014 | 1 | Shape 1 | 2014 | 6.2 | 50–54 | 2014 |
| 2 | 2014 | 2 | Shape 2 | 2014 | 9.4 | 55–59 | 2014 |
| 3 | 2014 | 1 | Shape 1 | 2014 | 6.2 | 50–54 | 2014 |
| 4 | 2014 | 4 | Shape 4 | 2014 | 13.8 | 60–64 | 2014 |
Removing any spatial reference from the data would, however, not allow a visualization of survey results (e.g., aggregates). Therefore, to enable visualization spatial information can be coarsened so that de-anonymization becomes impossible, which can be done by creating grids. Grids must be chosen that are large enough to preclude the de-anonymization of respondents in the data. The appropriate size of the grid depends on the density of the population and the size of the sample from the survey study. Therefore, different grids need to be tested (for example, 1 kilometer by 1 kilometer) and documented in a script as well as in the metadata.20

Coarsening of coordinates introduces imprecisions. This is especially relevant if respondents reside close to political boundaries that may change over time. For example, if differences between political areas are of interest, inferences based on respondent’s survey data may be biased because of a potentially wrong localization of those participating in the survey. An SDI must pay attention to issues such as these, and therefore, one must develop certain measures in order to estimate the bias that can arise. This can be done, for example, by assigning corresponding metadata.

Nevertheless, grids are especially important when running statistical analyses. They can be used to control for statistical dependence of respondents, e.g., if they live in the same neighborhood. It is conceivable that certain combinations of environmental noise, such as that caused by a large main road and a crossing tram, may render respondents’ noise sensations
to be more similar—independent from the actual decibel levels of the main road or the tram that are in the data. The same holds true for other covariates, including health measures, living conditions, or simple demographics, such as level of education obtained.

Removing direct identifiers and applying grids still may not be sufficient to achieve anonymization of the data. For example, it must be ruled out that a possibility exists to reidentify the coordinates of the respondents’ addresses by combining information from different contextual spatial variables included in a merged dataset. Specific combinations of values of different spatial variables may enable secondary users to deduce the identity of a respondent if the spatial data are publicly available (which is the case for most spatial data provided by government agencies). For example, a merged dataset (i.e., one including survey variables and spatial variables) may include data on several variables of ecological strain (such as air pollution, or noise from traffic, railways, and/or aviation). If a pattern of values of the spatial variables occurs only once in the data, a reidentification of the coordinates is possible even if the addresses and the coordinates are removed from the dataset because the original maps used for merging the spatial data with survey data can be combined to identify specific points in space.

To disseminate georeferenced data in accordance with German privacy legislation, the mode of access must preclude the de-anonymization of the data. It should be noted that privacy issues with respect to spatial data are discussed not only in Germany, but also other countries (Blatt 2012). Currently, there are, however, no clear rules for how to provide georeferenced survey data for primary or secondary use. Most experts recommend not publishing georeferences with survey data, but these recommendations refer to information on municipality and state levels. Whether spatial data can be published for scientific use once the spatial references are removed from the data has not yet been tested. Testing this outcome constitutes an important desideratum for facilitating the use of spatial data in the social sciences.

A spatial data infrastructure for the social sciences, therefore, would have to establish verification routines for checking whether spatial data merged to survey data can be disseminated to scientific users without access restrictions. Such verification requires transferring and adapting concepts developed for the anonymization of sociodemographic variables to the domain of spatial data. If the tests yield that the data do not include information that allows de-anonymization, the data set can be published for scientific use. These data contain the survey data, information at the grid level, and the contextual spatial data.

Another important advantage of spatial data is the possibility of harmonizing areas when constructing time series of survey data. The boundaries of administrative areas over time are sometimes modified. For example, the shape of municipalities, counties, or provinces can change due to territorial
reforms based on any number of factors. Estimating time series at local or regional levels becomes problematic because the areas covered by the administrative units are not necessarily static through time. Monitoring changes over time, however, is of high interest in the social sciences.

With GIS and boundary data, it becomes possible to solve this problem by assigning smaller level units (for example, individuals) from previous time points to boundaries of the current administrative units. With this harmonization over time, a consistent time series can be created and provided for research.

We have described how a fixed set of spatial data can be merged with survey data. The resulting merged dataset enables the exploration of various new research questions. For specific research questions, however, it may not be sufficient to have a predetermined, fixed set of spatial data. We therefore decided to develop a flexible and extensible georeferencing service, which gives researchers the option to submit their own spatial data to the SDI to be merged with survey data. To merge the data, information about the location of the respondent is necessary, such as addresses or at least zip codes. A merge performed by the researcher is therefore not possible because of the restrictions posed by German privacy legislation. Merging can be performed by data librarians serving as an intermediary, who thus play an essential role in ensuring data protection while facilitating new research opportunities. A further advantage of a merging service at an SDI is that resulting datasets can again be archived and disseminated easily within the framework of an existing infrastructure, and can be curated by data librarians.

However, archiving infrastructures need to be prepared for the specific requirements of spatial data. This concerns the fixed sets of spatial data, and also the data produced through merging of data submitted by individual researchers. In particular, the question of sustainable use of the data generated by geodata services is important and is discussed in the following section.

ARCHIVING AND DISSEMINATING DATA

A social science SDI should not only facilitate the use of spatial data but also for sharing data for secondary use. This requires archiving of the data, and a prerequisite for archiving is documentation. Information about all aspects of the spatial data is required for the reuse and dissemination of the data, going beyond the documentation related to the surveys. This step concerns two different levels of documentation: (1) the catalogue metadata for making studies findable and (2) variable metadata used in the data analysis. Two standards have been developed for documenting spatial data: ISO 19915 and INSPIRE (Rat für Sozial- und Wirtschaftsdaten–RatSWD 2012, 28). The ISO 19115 standard is used for gathering metadata for spatial data.21 In the INSPIRE guideline compulsory elements are defined that can be mapped to the
ISO 19115 standard. Using established standards ensures the connectivity of documentation about georeferenced survey data to common international standards and enables passing on metadata to national and international search portals (Karschnick et al. 2003).

The challenge of archiving georeferenced data is to develop an integrated metadata scheme that includes both survey metadata and spatial metadata. There are metadata standards for surveys (DDI, SDMX, Dublin-Core, etc.) and for spatial data (e.g., ISO 19115 and INSPIRE). For an integrated metadata scheme the required metadata elements must be defined. The DDI metadata standard for social science survey data includes in its version 3.X (DDI Lifecycle) spatial metadata elements of ISO 19115. The support of spatial data within DDI facilitates the inclusion of spatial metadata; thus, it becomes possible to add spatial metadata to existing metadata schemes (Vardigan, Heus, and Thomas 2008).

Generally, researchers are not familiar with standards for the creation of metadata. This is particularly true if the standards for different types of data have been developed in different disciplines and independently from each other. For instance, documenting data on noise exposure according to the ISO 19115 standard will be difficult for social scientists to accomplish, particularly if the spatial data were collected by scholars trained in survey research. These metadata records provide deeper information about the method of data merging (on the level of individual studies) and the context variables (on the variable level). Within an SDI, data librarians should be responsible for supporting the documentation process of spatial data because they have been trained to create and organize metadata and therefore will implement needed professional standards for metadata management.

Two different types of data have to be archived. First are the merged data, which include both survey data and spatial data without spatial identifiers. These can be documented, archived, and disseminated through existing data catalogues for social science research, such as the data catalogue at the Data Archive for the Social Sciences in Germany. Data catalogues facilitate comprehensive access to all data holdings and provide information on the conditions of access. Documentation provides additional information about the methods used for data merging and a basic description of the result of the spatial data merged with the survey data. In addition, versioning control is provided as part of the data, which entails assigning a study number and a digital object identifier (DOI). Subsequently, data backup is performed and data storage is provided in the archiving system.

The second type of data is the spatial data used for merging with survey data. These data include the spatial references and can be reused for merging with other survey data. Archiving spatial data in a data catalogue for survey data is not appropriate because users will search differently for spatial data than for survey data. For instance, searches for spatial data require detailed
information on geographic coverage and the format of the boundary data. Such information is not fully included in the metadata scheme of standard social science data catalogues. Researchers searching for spatial data, moreover, need access to the raw data to assess their suitability for merging and their scientific potential. The spatial data must be documented in detail for appropriate reuse (i.e., the data generating process), particularly in those cases where data processing and harmonizing involved a modification of the original data. Documentation of the spatial data can be done only by a data expert, in this case, by the librarian who collected and harmonized the data.

An SDI should also include a repository for uploading raw spatial data that have been used for merging with survey data. An extended documentation of the data will support searches for, and retrievals of, the spatial data by secondary users. Here again, librarians are experts at designing repositories and will contribute to establishing an appropriate user interface for uploading and downloading spatial data.

After the data have been archived, they can be prepared for dissemination. Conditions for data access depend on the level of disclosiveness of the data. Assessing disclosiveness is performed in accordance with the rules applied in the data archive to provide both privacy-conscious and user-friendly access to research data and includes the examination of individual variables and combinations of variables to assess their de-anonymization potential (c.f. section “Legal Barriers to the Use of Spatial Data in Social Science Research”). The DAS at GESIS in Germany provides data access in three different ways, depending on the level of disclosiveness of the data. First, anonymized datasets are offered for download to registered researchers. Survey data merged with environmental noise data from our example that has undergone anonymization will be shared with other researchers in this way. Second, some data may have a negligible degree of disclosiveness while not being anonymous. In such cases, the following access method can be considered: researchers signing a contract for off-site access can be allowed to access selected datasets. Off-site access delivers data to users for a specified time period on the condition that users honor precise security requirements.

In most cases, however, it may be necessary to restrict access to the spatial survey data as anonymization would be difficult to achieve or detrimental to the quality of the data (Kounadi and Leitner 2014). Therefore, the third and most restrictive way of accessing spatial data is through the Secure Data Center proper at the Data Archive for Social Science in Germany, which provides controlled and secured access to disclosive data based on special access requirements and restrictions. Data can be analyzed by appointment and upon signing a contract for on-site use in a safe room within the archive. The SDC has implemented various technical (e.g., working on a virtual machine presented to the researcher via a thin client) and organizational (e.g., access control and prohibition of mobile phone or laptop use in the safe room) means to ensure a high level of security and to allow for responsible
use of disclosive research data. Survey data merged with spatial data that has not undergone anonymization will be provided in this way—as outlined above direct identifiers will not be provided to researchers.

Secure methods of dissemination are particularly interesting for a social science spatial data infrastructure because of the high levels of detail provided by spatial information and also, for example, because environmental measurements promise a particularly high potential for novel social scientific analyses. However, before any dissemination of merged datasets created in the project can begin, it is necessary to test the different access routes to the spatial data with selected groups. Only if tests show that data can be provided in the various access methods in compliance with privacy legislation can data be made available to all interested researchers. More complicated is the case of researchers using the individual geoservice. In this case, there is a large quantity of spatial research data that can be used in future projects.

CONCLUSION

Survey data are among the most commonly used research data in the social sciences. Enhanced with spatial data, their analytical potential can be considerably improved. Unfortunately, up to now, combined data are available in Germany only for specific survey programs featuring a limited set of spatial variables, although many other research projects would likely benefit from merging spatial data with survey data. This is particularly puzzling because government agencies (e.g., statistical offices, local and regional governments) publish spatial data that they have collected more frequently. In contrast to commercial suppliers, these data are free for reuse for scientific purposes (in accordance with the INSPIRE guideline of the European Commission) and the data generating process is documented.

Two major factors hamper more widespread use of spatial data in the social sciences. First, although spatial data are freely available, the segmented structure of data production in Germany leads to a confusing multitude of different suppliers, data formats, formats for spatial references, and so on. Second, privacy protection legislation imposes restrictive conditions for merging survey data with spatial data if individuals’ identifiers (i.e., the location coordinates of respondents) are used for georeferencing.

Our goal in this paper is to show how a spatial data infrastructure can be designed to facilitate the use of existing spatial data in social science applications. We have sketched the most important elements and procedures for a spatial data infrastructure that facilitates access to spatial data for social scientists. These measures are not about the production of spatial data. Rather, it is necessary to support researchers through the assistance of specialized data librarians in merging spatial data from different sources with
survey data, as well as assisting in related tasks such as data management and preservation.

Using the example of environmental noise data, we illustrated the hurdles that have to be overcome when spatial data are intended to be used for merging with survey data. As has been shown, many challenges occur during the acquisition of spatial data and the subsequent harmonization process. For this reason, one of the most important elements of the infrastructure described here is a database that indexes publicly available spatial data and provides information about the content, the spatial and temporal coverage, and the format of spatial references that are included in the data.

The core of the social science SDI is a service for merging survey data with spatial data addressed to social scientists. Its objective is to design processes that enable merging spatial data with survey data. This requires that direct identifiers of respondents must not be stored in the same data file as the data collected in a survey. To meet this need, after georeferencing is accomplished, individual identifiers are removed from the data. Only the spatial information, such as noise exposure measured in decibels, remains in the data. Grids, on the other hand, can be used for visualization purposes or for statistical control in later analyses of the data. However, examination through sufficiently large grids must be applied to prevent a de-anonymization of respondents. Therefore, before archiving and disseminating the data, the level of disclosiveness of the merged data file is tested. If a de-anonymization is highly improbable, the data can be released for scientific purposes. If the level of disclosure is too high, analyses can proceed in secure environments within existing data archives or data libraries.

Finally, a major component of an SDI is a repository service for depositing and disseminating spatial data as well as merged survey data. Institutions providing an infrastructure for this purpose have existed since the 1960s (e.g., the UKDA in Essex or the DAS at GESIS in Cologne, Germany) and, for this reason, they do not have to be created from scratch. However, metadata schemes used in such archives or repositories have to be extended to integrate spatial data into the existing service workflow. Documenting the geographical and temporal dimensions of spatial data guarantees searchability and access. Only with the inclusion of such documentation can the service be reused by other scholars for new studies.

To summarize, a spatial data infrastructure for the social sciences can facilitate the use of spatial data by supporting researchers in merging data from different sources. In particular, a social science SDI contributes to overcoming burdens of merging survey data with spatial data in compliance with privacy protection legislation and provides a framework for sharing spatial data within the scientific community. To serve the increasing demand for spatial data in the social sciences we have discussed how it is therefore appropriate and advantageous to develop such an SDI.
Although the SDI at the DAS is not yet fully implemented because this project has only recently started, there has been strong and active progress due to the in-depth analysis and detailed concepts as described in this article. Furthermore, the concept is transferable to other data archives and repositories and could therefore be adapted and used to offer researchers in other countries the option to work with spatial survey data.

NOTES

1. http://www.gesis.org/en/institute/gesis-scientific-departments/data-archive-for-the-social-sciences/
2. http://www.gsdio.org/
3. http://inspire-geoportal.ec.europa.eu/
4. http://www.geoportal.de/DE/Geoportal/geoportal.html?lang=de
5. An ecological fallacy occurs when a correlation between contextual units is interpreted to be a correlation at the individual level. For example, the fact that social-democratic parties have higher vote shares in electoral districts with higher shares of blue collar workers does not necessarily mean that all or even most blue collar workers did in fact cast a ballot for social-democratic parties.
6. http://geo.data-archive.ac.uk/
7. See the German data protection law “Bundesdatenschutzgesetz (BDSG)” §3 (6).
8. http://www.div.de/en/div_02.c.222519.en/regional_data.html
9. http://www.microm-online.de/
10. http://infas360.de/
11. http://www.gfk.com/de/Seiten/default.aspx
12. http://inspire.ec.europa.eu/index.cfm/pageid/48
13. www.geoportal.de
14. An overview of the data provided at a regional and local level is available on the homepage of regional statistics of the federal statistical office: https://www.regionalstatistik.de/genesis/online/data;jsessionid=32EAD4373887ACFF5F59DF096D1062E3?operation=statistikenVerzeichnis. A major shortcoming is that this information is not available in English.
15. http://cdr.eionet.europa.eu/
16. In Switzerland, harmonizing and merging is done by the Federal Statistical Office for the whole country.
17. We are referring to the 2014 update of the environmental noise data in the CDR that was the most recent by the time we started our data research.
18. https://developers.google.com/maps/terms?hl=de
19. http://www.geodatenzentrum.de/geodaten/gdz_rahmen.gdz_div?gdz_spr=eng&gdz_akt_zelle=6&gdz_anz_zelle=3&gdz_user_id=0
20. Both the grids and their metadata will be designed to be compliant with INSPIRE.
21. http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=53798
22. http://inspire.ec.europa.eu/
23. https://dbk.gesis.org/dbksearch/index.asp?db=e
24. http://www.gesis.org/sdc

REFERENCES

Ainsworth, James W. 2002. Why does it take a village? The mediation of neighborhood effects on educational achievement. Social Forces 81(1): 117–152, doi:10.1353/sof.2002.0038.
Blatt, Amy J. 2012. Ethics and privacy issues in the use of GIS. Journal of Map & Geography Libraries 8(1): 80–84, doi:10.1080/15420353.2011.627109.
Bocquier, Aurélie, Sébastien Cortaredona, Céline Boutin, Aude David, Alexis Bigot, Basile Chaix, Jean Gaudart, and Pierre Verger. 2013. Small-area analysis of social inequalities in residential exposure to road traffic noise in Marseilles, France. *The European Journal of Public Health* 23(4): 540–546, doi:10.1093/eurpub/cks059.

Bodin, Theo, Jonas Björk, Evy Öhrström, Jonas Ardö, and Maria Albin. 2012. Survey context and question wording affects self reported annoyance due to road traffic noise: A comparison between two cross-sectional studies. *Environmental Health* 11(1): 14, doi:10.1186/1476-069X-11-14.

Chhetri, Prem, and Robert J. Stimson. 2014. Merging survey and spatial data using GIS-enabled analysis and modelling. In *Handbook of Research Methods and Applications in Spatially Integrated Social Science*. Cheltenham, Northampton: Edward Elgar Publishing, Inc.

Clark, Lara P., Dylan B. Millet, and Julian D. Marshall. 2014. National patterns in environmental injustice and inequality: Outdoor NO2 air pollution in the United States. Zhang, Yinping (ed.). *PLoS ONE* 9(4): e94431, doi:10.1371/journal.pone.0094431.

Crowder, Kyle, and Scott J. South. 2011. Spatial and temporal dimensions of neighborhood effects on high school graduation. *Social Science Research* 40(1): 87–106, doi:10.1016/j.ssresearch.2010.04.013.

Diekmann, Andreas, and Reto Meyer. 2010. Demokratischer Smog? Eine empirische Untersuchung zum Zusammenhang zwischen Sozialschicht und Umweltbelastungen. *KZfSS Kölner Zeitschrift für Soziologie und Sozialpsychologie* 62(3): 437–57, doi:10.1007/s11577-010-0108-z.

Dietz, Robert D. 2002. The estimation of neighborhood effects in the social sciences: An interdisciplinary approach. *Social Science Research* 31(4): 539–575, doi:10.1016/S0049-089X(02)00005-4.

Dittmann, Joerg, and Jan Goebel. 2010. Your house, your car, your education: The socioeconomic situation of the neighborhood and its impact on life satisfaction in Germany. *Social Indicators Research* 96(3): 497–513, doi:10.1007/s11205-009-9489-7.

Esri. 2015. *ArcGIS Desktop: Release 10.3*. Redlands, California: ESRI–Environmental Systems Research Institute.

GESIS–Leibniz Institute for the Social Sciences. 2015. ALLBUS/GGSS 2014 (Allgemeine Bevölkerungsumfrage Der Sozialwissenschaften/German General Social Survey 2014). *GESIS Data Archive*. http://dx.doi.org/10.4232/1.12209.

Helbig, Marcel. 2010. Sind Lehrerinnen für den geringeren Schulerfolg von Jungen verantwortlich?. *KZfSS Kölner Zeitschrift für Soziologie und Sozialpsychologie* 62(1): 93–111, doi:10.1007/s11577-010-0095-0.

Karschnick, Oliver, Fred Kruse, Stephanie Töpker, Thomas Riegel, Marco Eichler, and Sven Behrens. 2003. The UDK and ISO 19115 Standard. In *Proceedings of the 17th International Conference Informatics for Environmental Protection*, 475–81. Cottbus.

Knies, Gundi. 2010. Income comparisons among neighbours and life satisfaction in East and West Germany. ISER Working Paper Series, No. 2010-11.

Kohlhuber, Martina, Andreas Mielck, Stephan K. Weiland, and Gabriele Bolte. 2006. Social inequality in perceived environmental exposures in relation to housing conditions in Germany. *Environmental Research* 101(2): 246–255, doi:10.1016/j.envres.2005.09.008.
Kounadi, Ourania, and Michael Leitner. 2014. Why does geoprivacy matter? The scientific publication of confidential data presented on maps. *Journal of Empirical Research on Human Research Ethics* 9(4): 34–45, doi:10.1177/1556264614544103.

Lersch, Philipp M. 2013. Place stratification or spatial assimilation? Neighbourhood quality changes after residential mobility for migrants in Germany. *Urban Studies* 50(5): 1011–1129, doi:10.1177/0042098012464403.

MacKerron, George, and Susana Mourato. 2013. Happiness is greater in natural environments. *Global Environmental Change* 23(5): 992–1000, doi:10.1016/j.gloenvcha.2013.03.010.

Metschke, Rainer, and Rita Wellbrock. 2002. *Datenschutz in Wissenschaft und Forschung*. Berliner Beauftragter für Datenschutz und Informationsfreiheit.

Meyer, Reto, and Heidi Bruderer Enzler. 2013. Geographic Information System (GIS) and its application in the Social Sciences using the example of the Swiss Environmental Survey. *Methoden, Daten, Analysen* 7(3): 317–146, doi:10.12758/MDA.2013.016.

Monnot, Christophe. 2013. *Croire ensemble. Analyse institutionelle du paysage religieux en Suisse*. Zürich/Genève: Editions Seismo.

Oberwittler, Dietrich, and Per-Olof H. Wikström. 2009. Why small is better: Advancing the study of the role of behavioral contexts in crime causation. In *Putting crime in its place*. Weisburd, David, Wim Bernasco, and Gerben J. N. Bruinsma, (eds.), 35–59. New York: Springer.

Öhrström, Evy. 2004. Longitudinal surveys on effects of changes in road traffic noise-annoyance, activity disturbances, and psycho-social well-being. *The Journal of the Acoustical Society of America* 115(2): 719–129.

Passchier-Vermeer, Willy, and Wim F. Passchier. 2000. Noise exposure and public health. *Environmental Health Perspectives* 108(1): 123–131, doi:10.2307/3454637.

QGIS Development Team. 2015. *QGIS Geographic Information System*. Open Source Geospatial Foundation Project.

Rat für Sozial- und Wirtschaftsdaten–RatSWD. 2012. *Endbericht der AG “Georeferenzierung von Daten” Des RatSWD–Bericht der Arbeitsgruppe und Empfehlungen des RatSWD*.

R Core Team. 2015. *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Sager, Lutz. 2012. Residential segregation and socioeconomic neighbourhood sorting: Evidence at the micro-neighbourhood level for migrant groups in Germany. *Urban Studies* 49(12): 2617–2632, doi:10.1177/0042098011429487.

Saib, Mahdi-Salim, Julien Caudeville, Florence Carre, Olivier Ganry, Alain Trugone, and Andre Cicoletta. 2014. Spatial relationship quantification between environmental, socioeconomic and health data at different geographic levels. *International Journal of Environmental Research and Public Health* 11(4): 3765–3786, doi:10.3390/ijerph110403765.

Sampson, Robert J., Jeffrey D. Morenoff, and Thomas Gannon-Rowley. 2002. Assessing “neighborhood effects”: Social processes and new directions in research. *Annual Review of Sociology* 28(1): 443–478, doi:10.1146/annurev.soc.28.110601.141114.
Sampson, Robert J., Stephen W. Raudenbush, and Felton Earls. 1997. Neighborhoods and violent crime: A multilevel study of collective efficacy. *Science* 277: 918–924.

Schräpler, Jörg P. 2009. *Verwendung von SGB II-Dichten als Raumindikator für die Sozialberichterstattung am Beispiel der “sozialen Belastung” von Schulstandorten in NRW–Ein Kernel-Density-Ansatz*. Vol. 57. Statistische Analysen und Studien Nordrhein-Westfalen.

Siegfried, André. 1913. *Tableau politique de La France de l’Ouest sous la Troisième République*. Paris: Collin.

Stimson, Robert. 2014. A spatially integrated approach to social science research. In *Handbook of Research Methods and Applications in Spatially Integrated Social Science*, 13–25. Cheltenham, United Kingdom: Edward Elgar Publishing Inc.

Tumba, Anthony G., and Anuar Ahmad. 2014. Geographic information system and spatial data infrastructure: A developing societies’ perception. *Universal Journal of Geoscience* 2(3): 85–92.

Vardigan, Mary, Pascal Heus, and Wendy Thomas. 2008. Data documentation initiative: Toward a standard for the social sciences. *International Journal of Digital Curation* 3(1): 107–113, doi:10.2218/ijdc.v3i1.45.

Vardigan, Mary, and Cole Whiteman. 2007. ICPSR meets OAIS: Applying the OAIS reference model to the Social Science Archive Context. *Archival Science* 7(1): 73–87.

Voigtländer, Sven, Ursula Berger, and Oliver Razum. 2010. The impact of regional and neighbourhood deprivation on physical health in Germany: A multilevel study. *BMC Public Health* 10(1): 403, doi:10.1186/1471-2458-10-403.

Wirth, Heike. 2013. Das faktisch anonymisierte Mikrozensus-Regionalfile. In *Regionale Standards*, 229–254. GESIS Schriftenreihe 12.

Zwickl, Klara, Michael Ash, and James K. Boyce. 2014. Regional variation in environmental inequality: Industrial air toxics exposure in U.S. cities. *Ecological Economics* 107: 494–509, doi:10.1016/j.ecolecon.2014.09.013.