Schmollers Jahrbuch 130 (2010), 393 – 401
Duncker & Humblot, Berlin

The Research Data Center (RDC) of the German Socio-Economic Panel (SOEP)

By Joachim R. Frick, Jan Goebel,
Michaela Engelmann, and Uta Rahmann

1. Introduction

Based on recommendations from the "Commission to Improve the Informational Infrastructure (KVI)", the Council for Social and Economic Data (RatSWD) has issued guidelines for Germany's new Research Data Centers (RDCs, *Forschungsdatenzentren*). The RatSWD has been tasked by the Federal Ministry for Education and Research with "improving data use and data access for empirical research".[1] While at the beginning of the process the RDCs were mainly created for agencies that collect official statistical data and administrative register data and that decided, based on these recommendations, to make their data available to the scientific community for research purposes, the SOEP Research Data Center has a somewhat different history – mainly because SOEP is driven by academic research goals.[2]

The Socio-Economic Panel Study (SOEP), the largest longitudinal survey of private households in Germany, was founded in 1982 as a subproject of the German Research Foundation's (DFG, *Deutsche Forschungsgemeinschaft*) Collaborative Research Center 3 (Sfb3) "Microanalytic Foundations of Social Policy". SOEP was and still remains a research-based survey, in contrast to the official statistical agencies, and its goal is to improve the range and quality of data available for research on a wide range of social issues and contexts. Providing and distributing user-friendly data to a broad group of German and international scholars has always been among SOEP's central tasks.

In 2001, the Bund-Länder Commission for Educational Planning and Research Promotion (BLK, now the Joint Science Conference, GWK) recognized SOEP's importance for Germany's "research infrastructure". Since 2002, SOEP – which had been located at DIW Berlin since the mid-1980s – has been funded as a service facility by the federal government and the German states (*Bundesländer*). Since that time SOEP has been an independent member of

[1] See RatSWD, 2008.

[2] This is similar to some other RDCs under academic direction, such as the SHARE Research Data Center: http://www.share-project.org/t3/share/index.php?id=75.

the interdisciplinary network of infrastructural facilities in the Leibniz Association (WGL). In early 2009, SOEP established the SOEP Research Data Center as an overarching framework to facilitate and synthesize its diverse services.

2. The German Socio-Economic Panel Study: Data Preparation, Documentation, and Access

Since 1984 the SOEP data have been collected through an annual survey of as many eligible household members as possible belonging to households randomly selected into the survey sample. In survey year 2008 – the 25th wave of the SOEP survey – almost 20,000 individuals in over 11,000 households were surveyed, mainly by means of face-to-face interviews. The fieldwork is conducted by TNS Infratest Sozialforschung, Munich.

The SOEP research staff works together with the SOEP Survey Committee to create the questionnaire, also following recommendations provided by experts in the respective fields. The annual questionnaires and thus the SOEP microdata cover a wide range of subjects including employment and professional mobility, earnings, household composition, housing, social participation, time allocation, personal satisfaction, personality traits, physical and mental health, occupational and family biographies, childcare and education participation, as well as subjects dealt with in topical modules of the survey aimed at providing more in-depth information. These modules, which are not carried out annually but at longer regular intervals (normally every five to six years), concentrate on topics such as family and social services, education and training, social security, and environmental behavior.[3] Overall, the SOEP questionnaire design is "clearly centered on the analysis of the life course and well-being", the latter having been measured since SOEP's inception using the concepts of income and life satisfaction (Wagner et al., 2007, 146).

2.1 Data Distribution

With the regular SOEP data distribution, all SOEP users receive the data on DVD. Along with the original survey data, the DVD contains user-friendly support software, extensive documentation and additionally generated variables, which can be used in a wide range of analyses as independent or dependent variables (e.g., household typology, migration status, imputed income) and are also meant to directly support longitudinal research. All of the data distributed (for all survey years) are readily available for use with the major statistical software programs (SPSS, Stata, SAS) and are completely bilingual, with German and English variable and value labels. The DVD also includes

[3] For more information, see Wagner et al., 2007.

ASCII datasets and ASCII labeling information for those who use other more specialized programs. The microdata are organized in separate wave-specific files, complemented by various data files supporting longitudinal research. However, with the 2009 data release (data from survey years 1984 – 2008), a first version of a user-friendly "long format" has been provided that further expands the opportunities for longitudinal analysis. This "long format" pools observations for all individuals and households into one common file with consistent variables over time.

In order to reconcile the demands of scientific research with those of data protection in the best way possible, SOEP provides datasets with different degrees of anonymization and data protection measures. The preconditions for any data release are that the data be used exclusively for scientific research and that the user sign a data distribution contract, which essentially establishes the user's responsibility for correct use of the data under data protection law. The complete version of the SOEP data with the de facto anonymized microdata can be used in Germany and the countries of the European Economic Area (EEA); a scientific use file with a 95% random sample of the full version is provided to users outside these countries.[4]

Above and beyond data use for scientific research, the use of SOEP data in the college classroom is encouraged as well, but requires that the observations and variables in the dataset be further modified according to specific rules.[5]

2.2 Regional Data

The SOEP offers a wide range of possibilities for analyzing regional or geo-referenced data ranging all the way from the state (*Bundesland*) to the neighborhood level (see Figure 1). By using the regional classifications or codes assigned to a household, regional indicators on the levels of the states or *Bundesländer* (corresponding to NUTS-1)[6], the spatial planning regions (NUTS-2), the official county codes (NUTS-3), and the official municipality key (LAU-2)[7], postal codes, and street sections can all be linked with the

4 All procedures used in protecting the SOEP data are summarized in Frick et al., 2010.

5 The website of the RDC SOEP describes how to create such a teaching dataset (http://www.diw.de/en/diw_02.c.222839.en/soep_in_the_college_classroom.html). A completely anonymous SOEPcampus dataset is currently under preparation that will comply fully with data protection laws and also be of reduced complexity.

6 NUTS is the abbreviation of the french term *"nomenclature d'unités territoriales statistiques"*, a nomenclature of territorial units for statistics. It is a geocode standard for referencing the subdivisions of countries within the EU for statistical purposes.

7 LAU is the abbreviation of "Local administrative unit" and is a low level administrative division of a country. Within the EU geocode standard LAUs are basic components of the NUTS regions.

SOEP data. Since the available regional identifiers (with the exception of postal codes and street sections) refer to official geographic units of the Federal Republic of Germany, there are no technical obstacles to linking SOEP data with the official statistical data.

Level	Available Since	Data Access	Data Protection
States (Bundesländer)	1984	Standard SOEP dataset (Scientific Use File)	Data distribution contract
Municipal size classes (e.g., Boustedt)	1984	Standard SOEP dataset with special password	Expanded data distribution contract on the use of municipal size classes & data protection concept
Spatial planning regions (geocodes)	1985	Standard SOEP dataset plus SOEP geocode disk	Expanded data distribution contract on the use of geocodes & expanded data protection concept
Official county codes (KKZ)	1985	SOEPremote (online access to county-level regional data) or at the SOEP Research Data Center at DIW Berlin	Expanded data distribution contract on the use of SOEPremote & SOEPremote access form
Official municipality key, postal codes, Microm neighborhood data	2000 1993 2000	Use of data only at the SOEP Research Data Center at DIW Berlin	Only by personal arrangements in the framework of our SOEP in residence program

Figure 1: Regional datasets

Given the heightened sensitivity of these data under data protection legislation, however, varying levels of security precautions have to be taken depending on the specificity of the regional data in question. Users outside the EEA can use regional data below the state level only via remote access[8] or on the premises of the SOEP Research Data Center at DIW Berlin.

[8] This form of data access, referred to as SOEPremote, is available for data at the county level and is provided by a remote execution system (http://www.diw.de/documents/dokumentenarchiv/17/diw_01.c.95266.de/soepremote_jobsub.pdf).

2.3 Documentation

To make it easier to use the SOEP data – whose complexity has steadily increased with the survey's long duration – the SOEP Research Data Center offers a wide range of documentation that is also included with the data distributed annually on DVD. These documents include codebooks, papers describing the structure of the SOEP data, the generation of variables, the imputation procedures applied to correct for missing data arising from certain types of non-response, the weighting factors, the development of tests and questionnaires, and much more. There are also two databases (SOEPinfo and SOEPlit) and statistical series based on SOEP data (SOEPmonitor) that facilitate work with the SOEP data.[9]

The core of the SOEP Research Data Center documentation and a strongly recommended reference tool when embarking on any new research project is SOEPinfo, an interactive web-based application. It provides information on all variables for cross-sectional and longitudinal analysis (item correspondence lists) including unweighted frequencies as well as the numbers of observations available for balanced panel datasets (ranging from two to 25 waves of data). With the aid of diverse search functions by keyword, subject, or variable, users can quickly and easily identify and compile data for the research question at hand. Additionally, all variables are electronically linked to the original questionnaires in German and English, thus offering information on the context in which a given question was asked. SOEPinfo helps generate SPSS, SAS, and Stata command files in order to retrieve the desired subset of data from the more comprehensive SOEP database which currently consists of several hundred single data files.

The SOEPlit online literature database provides a comprehensive list of all publications known to us that are based on SOEP data. This database is yet another means by which the SOEP Research Data Center makes SOEP results more easily accessible – both to the public at large and to researchers seeking to build on previous work or compare their findings with others from the literature. In addition, SOEPlit which contains entries for more than 5,000 publications gives a comprehensive overview of the diversity of research opportunities that the SOEP data offers.

The SOEPmonitor provides time series on selected indicators based on SOEP data disaggregated for East and West Germany. The SOEPmonitor thereby fulfills two functions: first, it provides benchmark results for external users working with SOEP data. Second, it serves as quotable information (e.g., for journalists reporting on time trends in income inequality and poverty).

[9] The SOEP documentation is available online at: http://www.diw.de/en/diw_02.c.222735.en/documentation.html. For a more detailed description of the individual documents provided, cf. Anger et al., 2008.

SOEPmonitor provides information at the *household* level, for example, on housing conditions, and at the *individual* level, on the labor market, education, income, health, and various subjective indicators.

3. SOEP as Part of the International Research Infrastructure

International comparisons are a widely used tool in scientific research – not just for seeking best practices but also for obtaining a deeper understanding of human behavior and for studying different responses to social change. To support comparative research, SOEP contributes data to a range of cross-nationally harmonized databases. SOEP longitudinal microdata are part of the European Community Household Panel (ECHP) 1994 – 2001 directed by EUROSTAT, the Consortium of Household Panels for European Socio-Economic Research (CHER) coordinated by CEPS / INSTEAD in Luxembourg,[10] and – most importantly – the Cross-National Equivalent File (CNEF) coordinated by our US-based fellow researchers at Cornell University in Ithaca, NY[11]. At present, the CNEF contains microdata from the US Panel Study of Income Dynamics (PSID), the German Socio-Economic Panel (SOEP), the British Household Panel Study (BHPS), the Household, Income and Labour Dynamics in Australia Survey (HILDA), the Canadian Survey of Labour and Income Dynamics (SLID), and the Swiss Household Panel (SHP) (Frick et al., 2007).

For the (near) future, plans exist to include Asian datasets from the Russian Federation and China, and initial discussions have already taken place with the South African panel study NIDS. Furthermore, work is underway to include longitudinal data from the Korea Labor and Income Panel Study (KLIPS) in the upcoming data distribution.

Cross-sectional data from SOEP are already included in the Luxembourg Income Study (LIS) and the Luxembourg Wealth Study (LWS). The Luxembourg Income Study (LIS) database combines datasets from some 35 countries and contains income data supplemented by demographic and labor market information (Smeeding et al., 2002). The first release of the Luxembourg Wealth Study (LWS) database contains comparable wealth information from ten countries including data for Germany based on the SOEP wealth module collected for the first time in 2002, including multiple imputations in case of missing data due to non-response (Sierminska et al., 2006). Future extensions of LWS will also include the replication of this wealth module in 2007.

[10] http://www.ceps.lu.

[11] http://www.human.cornell.edu/che/PAM/Research/Centers-Programs/German-Panel/cnef.cfm.

4. Other Services Provided by the SOEP Research Data Center

4.1 SOEP in Residence

With the founding of the SOEP Research Data Center, the *SOEP in Residence* program was created to ensure a solid institutional footing for the diverse existing possibilities for guest researchers visiting the SOEP group at DIW Berlin. This program is designed to offer incentives for research on SOEP data at all levels of academic training, and to allow external researchers to benefit from the wide-ranging expertise of the SOEP staff in Berlin.[12]

4.2 SOEP as Reference Data

Household panels in general, and the German Socio-Economic Panel (SOEP) in particular, are useful as reference data for researchers whose primary datasets do not represent the full diversity of the population of interest (e.g., datasets obtained from clinical trials, intervention studies, laboratory and behavioral experiments, and cohort studies) (Siedler et al., 2009). As yet another unique service, the SOEP Research Data Center provides tailored advice to researchers who want to use SOEP as reference data or as a control sample for their research. In those cases, the SOEP Research Data Center also advises researchers on how to design their own longitudinal studies.[13]

4.3 SOEP Archive for the Reanalysis of Published Findings

Data protection issues are of utmost importance to SOEP and CNEF users. First, data protection comprises part of the (implicit) contract between the survey organization and the respondent. Second, in order to access the data, users are required to ensure full compliance with data protection legislation. Ultimately, all these precautions are crucial to achieve continued participation by panel respondents. Making SOEP and CNEF data available for reanalysis while maintaining the highest possible levels of data protection ("de facto anonymization", or *faktische Anonymisierung*) therefore presents a major challenge. Whenever such a microdata set is not considered completely anonymous (*absolut anonymisiert*) from a legal point of view, we – as data producers – cannot allow archiving subdata sets without establishing and monitoring compliance with clear-cut access regulations.

SOEP is committed to improving the statistical infrastructure for reanalysis and replication of findings using SOEP data. To this end, the SOEP Research

12 http://www.diw.de/en/diw_02.c.222617.en/soep_in_residence.html.

13 http://www.diw.de/soep-as-reference-data.

Data Center now offers users the opportunity to make their "SOEP working dataset" available to other researchers. This includes all databases associated with SOEP, such as CNEF, ECHP, LIS, and LWS. Most importantly, we offer a solution to those situations in which we cannot allow archiving of SOEP microdata in a journal's editorial office because the microdata in question are not considered "completely anonymized", minimizing the chances of disclosing the identity of a respondent by criminal means (because the data is considered as only "de facto anonymized").

4.4 Direct SOEP User Support

Above and beyond the comprehensive documentation and the various user support programs, the SOEP Research Data Center publishes the quarterly SOEPnewsletter, containing the latest updates on data, conferences, and related information, and distributes it by email to the constantly growing international SOEP user community. Additionally, the SOEP hotline at *SOEPmail@diw.de* offers individualized user support.

For further details on how to use the SOEP data and on the various services of the SOEP Research Data Center, please visit the SOEP Research Data Center website at http://www.diw.de/SOEPrdc.

References

Anger, S. et al. (2008): The SOEP user support and information system, in: B. Headey / E. Holst (eds.), SOEP Wave Report 1 – 2008. A Quarter Century of Change: Results from the German Socio-Economic Panel. Berlin: DIW Berlin, 114 – 119.

Frick, J. R. / *Goebel,* J. / *Haas,* H. / *Krause,* P. / *Sieber,* I. / *Engelmann,* M. (2010): Procedures for Controlling Access to the Confidential Microdata of the German Socio-Economic Panel Study (SOEP), DIW Data Documentation (forthcoming). Berlin: DIW Berlin.

Frick, J. R. / *Jenkins,* St. P. / *Lillard,* D. R. / *Lipps,* O. / *Wooden,* M. (2007): The Cross-National Equivalent File (CNEF) and its Member Country Household Panel Studies, Schmollers Jahrbuch 127 (4), 627 – 654.

RatSWD (2008): Kriterien des Rates für Sozial- und Wirtschaftsdaten (RatSWD) für die Forschungsdaten-Infrastruktur. http://www.ratswd.de/download/publikationen_rat/RatSWD_FDZKriterien.PDF.

Siedler, Th. / *Schupp,* J. / *Spiess,* C. K. / *Wagner,* G. G. (2009): The German Socio-Economic Panel (SOEP) as Reference Data Set, Schmollers Jahrbuch 129 (2), 367 – 374.

Sierminska, E. / *Brandolini,* A. / *Smeeding,* T. M. (2006): The Luxembourg Wealth Study – A Cross-Country Database for Household Wealth Research, Journal of Economic Inequality 4 (3), 323 – 332.

Smeeding, T. M. / *Jesuit*, D. K. / *Alkemade*, P. (2002): The LIS / LES Project Databank: Introduction and Overview, Schmollers Jahrbuch 122 (3), 497 – 517.

Wagner, G. G. / *Frick*, J. R. / *Schupp*, J. (2007): The German Socio-Economic Panel Study (SOEP) – Scope, Evolution and Enhancements, Schmollers Jahrbuch 127 (1), 139 – 169.