Schmollers Jahrbuch 130 (2010), 599 – 608
Duncker & Humblot, Berlin

European Data Watch

This section offers descriptions as well as discussions of data sources that are of interest to social scientists engaged in empirical research or teaching courses that include empirical investigations performed by students. The purpose is to describe the information in the data source, to give examples of questions tackled with the data and to tell how to access the data for research and teaching. We focus on data from German speaking countries that allow international comparative research. While most of the data are at the micro level (individuals, households, or firms), more aggregate data and meta data (for regions, industries, or nations) are included as well. Suggestions for data sources to be described in future columns (or comments on past columns) should be send to: Joachim Wagner, Leuphana University of Lueneburg, Institute of Economics, Campus 4.210, 21332 Lueneburg, Germany, or e-mailed to ⟨wagner@leuphana.de⟩. Past "European Data Watch" articles can be downloaded free of charge from the homepage of the German Council for Social and Economic Data (RatSWD) at: http://www. ratswd.de.

The Sample of Integrated Labour Market Biographies

By Matthias Dorner, Jörg Heining, Peter Jacobebbinghaus, and Stefan Seth*

1. Introduction

The Research Data Centre (FDZ) of the German Federal Employment Agency (BA), which is located at the Institute for Employment Research (IAB), provides high quality micro data on the German labour market. Since its establishment in 2004, the FDZ is acknowledged in the scientific community not only as an important service-oriented data supplier but also as a producer and developer of innovative data products.

* We would like to thank Stefan Bender and Tanja Hethey-Maier for their helpful comments.

With the *Sample of Integrated Labour Market Biographies (SIAB)*[1], the FDZ offers a new data set for the analysis of individual labour market biographies in Germany. A main goal of providing the SIAB is to make well-documented high quality data on labour market biographies available to the scientific community. It replaces prior individual biography data of the FDZ, such as the IAB Employment Samples (IABS) (Bender et al., 2000) and the Sample of the Integrated Employment Biographies (IEBS) (Jacobebbinghaus / Seth, 2007). The SIAB provides a large data basis of more than 1.5 million individuals and combines the advantages of both predecessors, namely the IABS's long observation period with the IEBS's comprehensive set of variables.

Compared to the IABS and IEBS two new features enhance the analytical potential of the SIAB: First, individual labour market outcomes of the so-called Hartz Reforms on Social Code (SGB) II in 2005 can be analysed. Second, all of the establishment information of the Establishment History Panel (BHP) can be linked to the SIAB. Hence, a detailed picture of the employment structure of the firm is available for every employee.

The SIAB is not only intended to foster empirical research on the German labour market; it also represents an initial step towards the harmonisation of the administrative data at the FDZ. The harmonisation will make it easier to switch between data sets, will give our data portfolio a clearer structure, and will generate added value in terms of data management efficiency.

The remainder of the article is designated to detail a description of the current version of the weakly anonymous SIAB data set[2].

2. Outline of the Data Set

2.1 Data basis of the SIAB

The majority of FDZ data products is based on process-produced data from administrative sources such as the notification process of the social security system and from several internal procedures of the German Federal Employment Agency (BA).

Data from the notification process of the social security system are submitted in accordance with the German Data and Transmission Act (DEÜV)[3]. They comprise information on individual employment episodes subject to the social security contributions. Hence, certain groups of employees like civil servants or self-employed are not covered in the data. Marginal (part-time)

1 *SIAB: Stichprobe der Integrierten Arbeitsmarkt-Biographien des IAB.*

2 The current version of the data set is *SIAB 7508.*

3 DEÜV: *Verordnung über die Erfassung und Übermittlung von Daten für die Träger der Sozialversicherung.*

workers are featured since the general restructuring of the notification scheme in 1999. These data on employment are combined with data produced at the Federal Employment Agency. This collection of data contains information on recipients of benefits from the social security system according to Social Code (SGB) III legislation and since 2005 also on benefit recipients covered by the SGB II legislation. Moreover, participation in active labour market programmes[4] as well as job search activities officially registered at the BA are included in the data basis (Dorner et al., 2010).

Eventually, these comprehensive data on individual labour market activities are comprised in the Integrated Employment Biographies (IEB) of the IAB.

The SIAB and its predecessor, the IEBS, are both based on the comprehensive IEB data with complete individual labour market biographies. Figure 1 depicts the full spectrum of data sources comprised in the SIAB data basis.

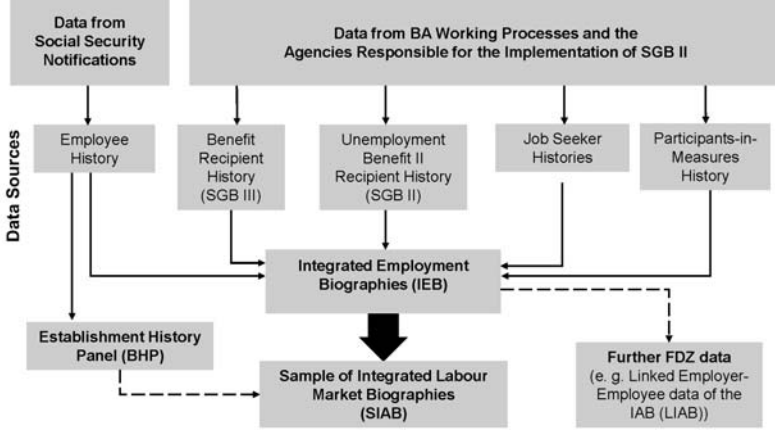

Figure 1: Administrative data basis of the SIAB

Since the procedure of data generation contains sources which either were implemented by law or were made available just in recent years (e. g. data sources on the reformed SGB II), older FDZ data up to the release of the SIAB were based on differing intermediate data sources and data generation procedures.

In fact, data sets like the IABS incorporated only employment records and data on benefit recipients covered by the SGB III. Even though these data were available from 1975 onwards, they lacked of complete information on the in-

[4] These data are not yet included in the current version of the SIAB, but will be featured in the updated version available in 2011.

dividual labour markets status such as job search or participation in active labour market programmes.

The IEBS resolved this particular shortcoming of the IABS by introducing further data sources to reproduce more comprehensive labour market biographies. However, the data could only be utilised for a relatively short observation period compared to the IABS.

Given these shortcomings of either the IABS or the IEBS, the SIAB was designed to combine the advantages of both data sets in only one data product. Furthermore, the SIAB is the first FDZ micro data to include administrative data sources which were established as a result of the reforms of the Social Code legislation in 2005. Hence, the SIAB incorporates the most recent information available on unemployment benefits and enables researchers to distinguish between benefit receipt according to the legal context of the SGB II and SGB III (Dorner et al., 2010).

2.2 Sample Design and Observation Period

The SIAB is a two percent random sample drawn from the full population of the IEB. This means that inferences can be drawn from the sample not only with respect to (marginal part-time or regular) employees but also with respect to job searchers and benefit recipients. The way the sample of the SIAB was drawn can be regarded as an enhancement compared to the IABS, which was an employment based sample, and hence, inferences on additionally linked benefit recipience was limited.

The current version of the SIAB contains employment histories of 1,659,024 individuals, documented in 40,501,525 data records. The SIAB comprises data reaching as far back as 1975. However, it must be kept in mind that not all of SIAB's data sources provide information since 1975, and the information from some sources is more readily available than the information from other sources.

Data on employment are featured from 1975 until 2008[5]. Unemployment benefits covered by the SGB III legislation are available from 1975 until 2009. Data on job search and participation in active labour market programmes are only contained from 2000 onwards. Information on benefit recipients and job search of individuals covered by the SGB II is not available until 2005. Figure 2 visualises for which respective timeframe the data sources in the current SIAB version provide information.

[5] Employment records are only available with a certain delay as the notification procedure of the social security administration allows retroactive reports from the employers. For further information please refer to Dorner et al. (2010).

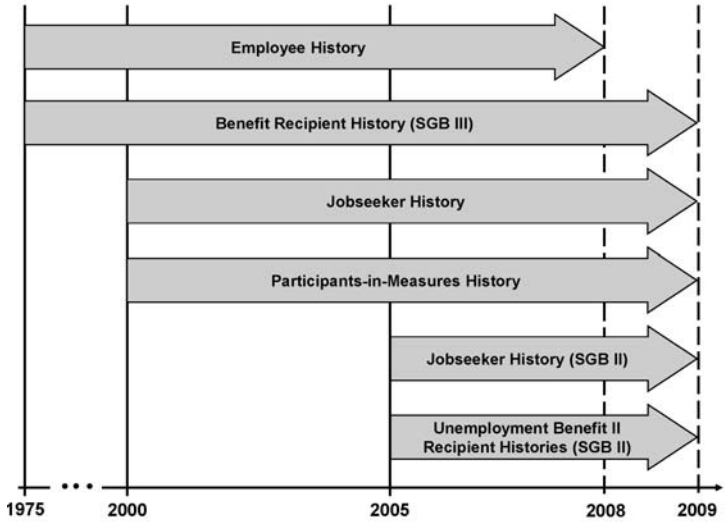

Figure 2: Timeframes of SIAB's data sources

2.3 Features of the Data

The SIAB is provided as spell data. Parallel information is recorded, i. e., it is visible in the data when people have several jobs at the same time, or receive employment benefits while being on job search, or take part in certain active labour market programmes triggering both an employment spell and a "policy" spell, etc.[6]

All spells in the SIAB are recorded on a daily basis. Hence, exact daily information on the *labour market status* is available in the data. For employment spells, it indicates the type of the actual employment: e. g., standard employment, apprenticeship, marginal (part-time) employment. For benefit spells, the type of benefit according to the respective social code legislation is reported. Job search spells are classified conditional on current employment or non-employment. For policy spells, the respective variable signifies the type of the labour market programme an individual attends.

Further socio-demographic characteristics of individuals and basic information on the employing establishment enhance the data substantially. On the basis of regional codes either representing administrative classifications of the or the Federal Statistical Office (*Statistisches Bundesamt)*, individuals can be localised both at their place of work and place of residence.

[6] Several "technical" variables are included to facilitate data handling. For further information see Dorner et al. (2010).

The subsequent list shows a basic classification and selection of the contained variables. For a complete list of the variables in the SIAB and a detailed description please refer to Dorner et al. (2010):

- *Identifiers:* person-id, establishment-id.
- *Labour market status:* prior, during and after the spell.
- *Employment and job search:* e. g., gross daily pay, occupation, occupational status / working time.
- *Benefit receipt:* e. g., daily benefit rate, type of benefit and willingness to seek employment throughout Germany.
- *Socio-demographic characteristics:* e. g., year of birth, sex, citizenship, school education, training, severe disability status and marital status.
- *Basic characteristics of the employing establishment:* e. g., industry code, first and last appearance of establishment(-id), number of (full) employees and median gross daily pay.
- *Regional codes:* place of residence and place of work.

As several other FDZ data originate from the same data sources, they are basically compatible with each other. While the SIAB only includes basic information on the employing establishment of an individual, the IAB Establishment History Panel (BHP) comprises extensive aggregated cross sectional data on the firm level, stemming from the same administrative data basis (Spengler, 2008; Hethey-Maier / Seth, 2010). The SIAB and the BHP both include identical identifiers for establishments, hence making both data sets suited for combined analysis (see Figure 1). All variables from the rich set of the BHP can be merged with the SIAB, which enhances the analytical potential significantly.

2.4 Data Quality Issues

Generally, the data comprised in the SIAB can be considered to be very reliable. This is particularly true for information collected not exclusively for statistical purposes: for instance, the data on remuneration are used by the German Statutory Pension Insurance (*Deutsche Rentenversicherung*) to calculate pension claims. As another example, information on education is solid in job seeking spells as it is highly relevant for job placement, and the respondent has an incentive to assure correctness. However, the educational variable is less reliable for employment spells (originating from employment register data) because an incorrect information there neither hurts the employer (who gives the information) nor the employee (Fitzenberger et al., 2006; Scioch, 2010).

Data on individuals covered by the SGB II legislation is either collected by the BA or the local authorities (*zugelassene kommunale Träger*) in charge of the implementation. Due to problems in the data collection and transmission

many records have not been stored appropriately in the first two years. However, as of the year 2007, the quality of the data on benefit receipt in the legal context of SGB II improved significantly and can be considered as good (Dorner et al., 2010).

To account for these shortcomings of the data, the FDZ assures the transparency of data quality issues by providing extensive documentation (Dorner et al., 2010). Furthermore, the FDZ offers a wide range of adequate working tools for the SIAB such as publicly available test data and publications in the *FDZ-Methodenreporte* series, which address, among other things, data quality or practical issues of the FDZ data.[7]

Researchers are required to spend some pre-research effort to study our data documentation and make the data suitable for their analysis. The actual amount of effort depends on which data sources and variables are eventually used to achieve the research goal.

3. Research Topics

The SIAB includes comprehensive information on individual employment biographies, provides a rich set of potential covariates and an extended observation period distinguishing it from its predecessors. Hence a wide range of potential research questions from various academic fields can be addressed adequately with just one data set.

Major topics of research with the data include foremost individual labour market outcomes such as wages (e. g., rigidities, discrimination, dispersion, effects of education), the duration of employment or unemployment and mobility (e. g., regional, occupational). Numerous studies have been conducted based on the analysis of the IABS or the IEBS.[8] With the availability of additional data and the possibility to include further information on the firm level, existing research can be extended in many ways. This is particularly the case for evaluation studies, which up to now faced either the limitation of only a short observation period (IEBS) or a lack of detailed information on the individual labour market status in non-observed periods (IABS).

Moreover, the so-called Hartz reforms, which became effective in 2005, brought fundamental changes to the German labour market. These reforms have not only affected the institutional structures but also the status of certain

[7] *FDZ-Methodenreporte* and related publications cover issues like the implementation of various definitions of unemployment in FDZ data (Kruppe et al., 2008) or the cleansing of recorded dates of active labour market programmes (Waller, 2008). Synthetic SIAB test data as well as *FDZ-Methodenreporte* are available on our website. http:// fdz.iab.de.

[8] For an up to date list of publications please refer to our website http:// fdz.iab.de.

groups of individuals. The SIAB is the first administrative data set which fully accounts for these reforms and thus opens up research opportunities for a description and yet an evaluation of the effects of this reform.

4. Data Access

The legal background of providing data access to the SIAB is § 75 book X of the German Social Code. All person-specific information collected by the German Federal Employment Agency in order to provide payments of unemployment benefits, participation in active labour market programmes or assistance in job placement, are so-called social data (*Sozialdaten*). For them strict confidentiality rules apply.[9] Since the data by law are considered as very sensitive, the ways the FDZ provides access to the weakly anonymous SIAB have to implement high confidentiality standards.

The FDZ currently offers access to the SIAB via on-site use at the FDZ and subsequent remote execution. Before data access is granted, an application form has to be filled by the researcher, approved by the Federal Ministry of Labour and Social Affairs (BMAS), and a contract with the FDZ has to be signed. Scientific use of social data requires the following conditions to be met and stated in the original request for data usage:

– Scientific research regarding social security (§ 75 SGB X).
– Prevailing public interest.
– Permission of the Federal Ministry of Labour and Social Affairs (BMAS).

The FDZ coordinates the whole application process of researchers. Specific application forms, guidance and further information on the different ways of data access can be found on our website. Based on the data use agreement, researchers is provided direct on-site access at the FDZ in Nuremberg.[10] To promote data access for researchers from non-German-speaking countries at the FDZ, the IAB provides financial support in terms of a partial payment of accommodation and/or travelling costs to lower their costs of on-site data access. Regarding the results produced in on-site use, our staff checks the generated output that contain the results and erases information suited to identify individuals in the SIAB.

9 Factual anonymous scientific use files are exempt of this legislation. Similar to the approach with the IABS Regional File we currently implement a factual anonymous version of the SIAB which be available for off-site use.

10 In the future, on-site data access will be provided at further German research data centers which participate in the RDC-in-RDC programme and in a research data centre established at the University of Michigan in Ann Arbour, US (Heining, 2010). Thereby, geographically more disperse and more cost efficient data access is provided both for international and national users.

After a research stay at the FDZ, researchers can decide to continue data processing via remote data execution. Remote execution means that the researcher uses the data documentation and the publicly available test data to prepare statistical code[11] and sends it to the FDZ by e-mail. The FDZ staff executes the code and checks the generated output so that information suited to identify individuals is deleted from the statistical output. The remaining anonymised results are returned to the researcher by email.

5. Prospects

The SIAB is intended to become the core data product of the FDZ for empirical research on individual labour market biographies. Therefore, the weakly anonymous version of the SIAB will be updated on a regular basis, the next update is scheduled for 2011. One focus will be to further improve the data quality, especially with regard to the SGB II records. Any feedback on data quality issues is greatly appreciated.

The FDZ would like to broaden the ways in which the SIAB may be accessed by researchers. However, Social Code legislation restricts the delivery of sensitive social data such as the SIAB directly to the data user at places outside of the FDZ. Therefore, the FDZ is working on a concept of a factual anonymised scientific use file of the SIAB, which accounts for the trade-off between the requirements of data confidentiality and research potential of the data. This successor of the IABS Regional File will be available for offsite use.

So far the FDZ's administrative data sets slightly differ in the way the variables are prepared although the variables originate from the same administrative processes and contain the same information. One reason for this is that the data were drawn from different intermediate data bases. Since the IEB has been extended to contain employment and benefit records since 1975, it is now possible to base all administrative data on individuals on the same data base. Hence, future FDZ data sets will be harmonised with regard to the naming and coding of their variables. The conventions applied in the SIAB will be the standard for all other FDZ data sets such as the Linked Employer-Employee Data of the IAB (LIAB) (Alda et al., 2005). This harmonisation will make it easier to switch between data sets as programmes written for one data set can also be used for others. Second, our data portfolio will get a clearer structure and last but not least, data production and documentation will be more efficient.

11 We currently accept code written in Stata, SPSS, Matlab, Gauss, TSP.

References

Alda, H. / *Bender,* S. / *Gartner,* H. (2005): The Linked Employer – Employee Dataset created from the IAB Establishment Panel and the Process-Produced Data of the IAB (LIAB), Schmollers Jahrbuch 125 (2), 327 – 336.

Bender, S. / *Haas,* A. / *Klose,* C. (2000): The IAB Employment Subsample 1975 – 1995, Schmollers Jahrbuch 120 (4), 649 – 662.

Dorner, M. / *Heining,* J. / *Jacobebbinghaus,* P. / *Seth,* S. (2010): Sample of Integrated Labour Market Biographies (SIAB) 1975 – 2008, FDZ Datenreport (01 / 2010).

Fitzenberger, B. / *Osikominu,* A. / *Völter,* R. (2006): Imputation Rules to Improve the Education Variable in the IAB Employment Subsample, Schmollers Jahrbuch 126 (3), 405 – 436.

Heining, J. (2010): The Research Data Centre of the German Federal Employment Agency: Data Supply and Demand between 2004 and 2009, Journal for Labour Market Research 42 (4), 337 – 350.

Hethey-Maier, I. / *Seth,* S. (2010): Das Betriebs-Historik-Panel (BHP) 1975 – 2008, FDZ Datenreport (04 / 2010).

Jacobebbinghaus, P. / *Seth,* S. (2007): The German Integrated Employment Biographies Sample IEBS, Schmollers Jahrbuch 127 (2), 335 – 342.

Kruppe, T. / *Müller,* E. / *Wichert,* L. / *Wilke,* R. (2008): On the Definition of Unemployment and its Implementation in Register Data – The Case of Germany, Schmollers Jahrbuch 128 (3), 461 – 488.

Scioch, P. (2010): The Impact of Cleansing Procedures for Overlaps on Estimation Results – Evidence for German Administrative Data, FDZ Methodenreport (4 / 2010).

Spengler, A. (2008): The Establishment History Panel, Schmollers Jahrbuch 128 (3), 501 – 509.

Waller, M. (2008): On the Importance of Correcting Reported End Dates of Labor Market Programs, Schmollers Jahrbuch 128 (2), 213 – 236.