Schmollers Jahrbuch 125 (2005), 437 – 447
Duncker & Humblot, Berlin

European Data Watch

This section will offer descriptions as well as discussions of data sources that may be 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 will start with data from German-speaking countries that allow international comparative research. While most of the data will be at the micro-level (individuals, households, or firms), more aggregate data and meta data (for regions, industries, or nations) will be included as well. Suggestions for data sources to be described in future columns (or comments on past columns) should be sent to: Joachim Wagner, University of Lueneburg, Institute of Economics, Campus 4.210, 21332 Lueneburg, Germany, or e-mailed to ⟨wagner@uni-lueneburg.de⟩.

The Research Data Centre of the Federal Employment Service in the Institute for Employment Research

By Annette Kohlmann

Introduction

The German Federal Employment Service (FES) is the most important producer of register data on the German labour market. In order to make these data available for external researchers, the Institute for Employment Research (IAB)[1] had already undertaken several steps in the 1990s.[2] Nevertheless, the access to those official microdata had not been developed in a systematic way.

[1] The „Institut für Arbeitsmarkt- und Berufsforschung“ (IAB) is the research institute of the „Bundesagentur für Arbeit“ (FES).

[2] For an overview on the datasets available to researchers before the establishment of the Research Data Centre, (Allmendinger / Kohlmann 2005).

Following the recommendations of the "Kommission zur Verbesserung der informationellen Infrastruktur" (Commission to improve the informational infrastructure by co-operation of the scientific community and official statistics; KVI 2001), the FES and the IAB established a research data centre in spring 2004[3]. During its pilot period (until November 2006), it is funded by the Federal Ministry of Education and Research (BMBF), supplemented by means of the FES. Depending on its success, it is planned to establish the research data centre as a permanent department.

The basic aim of the research data centre of the FES in the IAB[4] is the institutionalisation of data access for researchers analysing the German social security system and the labour market, given the regulations on data protection. In order to achieve that aim, the research data centre fulfils the following tasks:

- It develops standardised and plain ways of data access for researchers.
- It provides the development of data models and their implementation, the checking and preparation of data for use as well as the updating of data.
- It develops and organises documentations on data, on statistical aspects of the data and it provides means in order to facilitate analyses of the data.
- It organises conferences on data and provides individual counselling.

The aim of this paper is to give insight into the activities of the research data centre of the FES in the IAB. We start with an overview on the data available from the research data centre, followed by a brief description of the fields of research in which these data have been used up to now. Next, we describe the different modes of access to the data in the research data centre. We conclude by describing the additional services the research data centre provides to the scientific community. Detailed information on the activities of the research data centre is available at: http: // fdz.iab.de;[5] our e-mail address is: iab.fdz @iab.de.

[3] Since 2001, other producers of official data in Germany have also set up research data centres (Zühlke et al. 2004; Rehfeld 2004). Moreover, the service centre of the Institute for the Study of Labour (IZA) was established and the service centre of the German Social Science Infrastructure Services (GESIS) at ZUMA has intensified and enhanced its focus (Lüttinger et al. 2004).

[4] Due to reasons of data confidentiality, the research data centre of the FES is settled at the IAB.

[5] The research data centre's internet site is currently in German language. An English version of the main pages will be available soon.

Data on the German labour market[6]

The research data centre of the FES in the IAB provides official micro data on the German labour market. With regard to content, those data cover various facets of the labour market: employment and unemployment covered by social insurance, establishments, participation in measures of active labour market policy, payment of social benefits and job search. These data comprise survey data, register data from the notification procedure of the social security system in Germany, as well as data coming from the employment services (e.g. data on job placing). Therefore, most of them are process-produced person-specific data on a daily basis.

The *IAB Establishment Panel* (Bellmann 2002 / Kölling 2000) is an annual survey on establishments in Germany and provides information on approx. 16,000 establishments from 1993 (West Germany; East Germany since 1996) to 2003. It covers various aspects of establishments like employment, recruitments and dismissals, firm turnover, investments, exports, innovations, wages, working hours, retraining as well as subsidies. The sample is drawn from the employment statistics, using the establishment number assigned by the employment services. Therefore, the IAB establishment panel covers exclusively establishments with at least one employee covered by social security at the time of the interview (30th June of each year). These data are available via controlled remote data access in the research data centre.

The *Employment Panel* of the FES (Koch / Meinken 2004) consists of an approx. 2%-sample of employees covered by the social security system in Germany. The source of this panel is not a survey but the integrated notification procedure for the social insurance system (Neidert 1998). These process-produced data range from 1998 to 2003. Each of the 24 quarterly waves consist of information on approx. 600,000 individuals (since the sixth wave, „marginal" part-time employees are also included). The research data centre updates the dataset annually. The employment panel provides information on employees (demographic information, education, occupational status, occupation, group of employment etc.) as well as information on their employers (size of establishment, branch, and ratios of specific groups of employees in the establishment). The data are factually anonymised in accordance with the regulations for the anonymisation of the German Microcensus (Müller et al. 1991) and are provided as a Scientific Use File. A second „weakly anonymised" version of this dataset contains the unanonymised variables while person numbers are changed (pseudonymised) in order to prevent from de-anonymisation.

---

6 This overview concentrates exclusively on data that will be made available by the middle of 2005. An overview on all data that will be accessible until the end of 2006 – subject to matters of data protection – is given in Oertel / Passenberger / Janser 2004.

The *IAB Employment Subsamples (IABS)* are also based on data of the social security pension. In contrast to the Employment Panel, the IABS samples are drawn from the longitudinal processed database of the employment notifications. Additionally, these data sets cover information on benefit recipients. In general, three different kinds of IABS samples can be distinguished: (1) the basic scientific use file (covering the years 1975 – 1995) (Bender / Haas / Klose 2000), (2) the regional scientific use file (1975 – 2001) (Hamann et al. 2004) and (3) the pseudonymised original file (1975 – 2001). All data sets cover demographic information, education, occupational codes, employment status, gross pay per day, unemployment benefits, unemployment assistance, maintenance allowance, industry, establishment size etc. (1) is a 1 % sample including approx. 560,000 persons in Germany (East Germany since 1992). It is factually anonymised in the cross-section and longitudinal. In this file, only West and East Germany can be distinguished while industry and occupational codes each have three digits. (2) is a 2 % sample with information on 1.3 Mio employees and is factually anonymised in the cross-section. The main focus is set on regional codes (343 regions) while information on industry (16 categories) as well as on occupational codes (130 categories) is rather crude. (3) is the same 2 % sample as in (2) but is only weakly anonymised, i.e. it contains most of the original variables.

Since 2005, the research data centre offers a new database that allows for simultaneous analyses of employees and the establishments they are working at, the *Linked-Employer-Employee-Dataset of the IAB (LIAB)* (Alda / Bender / Gartner 2005b). The sources for this dataset are the IAB Establishment Panel and data from the social security system. Currently, the research data centre has developed two LIAB versions: a cross-sectional version and a longitudinal version (for details see Alda / Bender / Gartner 2005a). On the personal level, the variables contain demographic information as well as occupation, daily wage, start and ending of employment notification, schooling / training, detailed regional level etc. On the establishment level, all variables of the IAB Establishment Panel are available.[7] The cross-sectional version includes information on 1.9 to 2.7 Mio employees in 3,900 to 15,000 establishments for the years 1993 to 2001. The longitudinal version offers information on 2,100 establishments and 1.8 million employees and their employment and benefit recipient histories (start and end, type of benefit approval, reasons for submitting the benefit notification, reasons for terminating the receipt of benefit etc.). In both LIAB versions, data on East and West German employees and establishments are available. For reasons of data confidentiality, data access to the LIAB versions is possible only via a guest stay at the research data centre in Nuremberg.

[7] The datasets are not connected with each other but have to be matched by the respective researcher via pseudonymised identification variables. The datasets are not factually anonymised.

The last dataset we will mention here is a joint 2 %-sample from the employment histories, unemployment histories, participants in measures of active labour market policy and unemployed persons who are searching for an employment but who are not eligible for unemployment benefits. This *Integrated Employment Biographies Sample (IEBS)* embraces data from 1990 to 2004 – however, not all of the above-mentioned information is available throughout the whole period. In total, approx. 17 Mio observations from 1.37 Mio individuals are included in the dataset. The research data centre currently is preparing those data for access by external researchers. Moreover, like in the case of the LIAB, we are planning several IEBS versions.

Research potential

The research data centre of the FES in the IAB currently focuses on providing combined datasets to the scientific community. Combined datasets are datasets that cover information on different kinds of individuals (e.g. employed and unemployed persons in the IABS and in the Employment Panel) or integrated information on different actors on the labour market (like establishments, employed and unemployed persons in the LIAB). This enhances the research potential of the data tremendously while at the same time most of the information is quite precise (on a daily basis) and therefore suitable especially for longitudinal analyses. The data are usually 1 % or max. 2 % samples. Nevertheless, due to the size of the original population, plenty of analyses still can be conducted (also with factually anonymised datasets).

Analyses on the basis of the IAB Establishment Panel have been concentrating on employment trends in establishments, (further) training in establishments, flexibility of establishments, productivity and innovations in establishments and analyses on specific industries. An overview on publications based on the IAB-establishment panel is given at http: // betriebspanel.iab.de / publikationen.htm.

As well, analyses based on the IABS versions are quite heterogeneous. They comprise investigations on wage structures and their development, regional mobility of employees, occupational mobility, effects of institutional changes on the labour market, labour turnover, unemployment durations, labour market performance of distinctive groups, etc. An older overview on publications based on the IABS is given in Bender / Haas (2002).

Up to now, a rather small number of scientists have used the Employment Panel. Projects using the Employment panel dealt with immigration and integration on the labour market, international comparisons of labour markets, skill needs, labour supply and labour demand, etc.

Data from the LIAB versions (or previous versions) have been used for analyses on wage gaps with regard to gender, industries, occupational groups and

establishments, for the estimation of wage equations taking into account firm-specific and individual-specific characteristics, failure on completing company vocational training, effects of technological and organisational changes on mobility, etc. An overview on publications using older or current LIAB versions is given in Alda / Bender / Gartner (2005a).

Data access

The Social Code X (§ 75) and the Social Code III (§ 282) form the legal basis for data access to the data mentioned above by defining the requirements for access to not anonymised and anonymised data from the social insurance sector (for detailed information cf. Allmendinger / Kohlmann 2005). The main aim is to protect data from disclosure. Accordingly, the need for protection of the respective data determines the ways and limits of data access. In practice, the general rule is the more detailed the data, the more restrictive is the access to them. We can distinguish three modes of data access:

(1) Data access via *scientific use files* is possible for factually anonymised data. This holds for the IABS-files (basic file and regional file) as well as for the factually anonymised Employment Panel. These datasets are available for delivery to researchers from Germany as well as for researchers from other EU-countries. Scientific use files can be handed over only to full scientists; they may not be used in teaching or for commercial research interests. German researchers turn to the Zentralarchiv für Empirische Sozialforschung (ZA)[8] of the University of Cologne that distributes the data. Researchers send an application for delivery of the scientific use files to the ZA. In this application, they have to specify a project related to labour market research as well as a settled timeframe and the names of the persons who will be in contact with the data. Moreover, the modes of providing data security have to be described. The ZA hands over the application to the research data centre, which decides on it. Then, a contract is filed between the ZA and the research institution at which the researcher is employed. After this, the ZA sends the data to the respective researcher. As an administrative charge, 50 € have to be paid to the ZA. Currently, an international contract in English language is under development. In the meantime, researchers from the EU can turn to the research data centre of the FES in order to get access to the Scientific Use Files without a fee.

(2) Data that cannot be anonymised without losing basic research potential may be accessed via *controlled remote data access.* This applies to the original IAB Establishment Panel and the pseudonymised Employment Panel. This way of data access is open to full researchers as well as students from Ger-

[8] Central Archive for Empirical Social Research, cf. http: // www.gesis.org / en / za / index.htm.

many and from abroad. Data access for teaching or commercial research interests is not possible. Researchers send a form[9] that includes information on the applicant, the aim of the analyses and their content to the research data centre, which decides on the application. Then the researcher sends syntax files (STATA, SPSS or SAS) to the research data centre and they are processed with the original data. We anonymise the results and then send them back to the researcher. In order to allow for flawless syntax files, the research data centre provides additional information on the internet. This service includes questionnaires, codebooks, and syntaxes for dealing with specific problems of the data. Moreover, artificial test data files that mimic the structure of the original files are available. Test data files allow the researcher to check his syntax before he sends it to the research data centre in case there is no scientific use file available.[10] The research data centre charges no fee for this way of data access.

(3) The third possibility is a stay as a *guest researcher* in the research data centre. This mode of access is appropriate if it is not possible to anonymise the data factually and especially if the data are complex. Since 2005, the research data centre offers access to the two LIAB versions, the IAB Establishment Panel, the pseudonymised Employment Panel, as well as to the pseudonymised IABS in this way. One advantage of a guest stay in the research data centre is, among others, the access to full information on the content of the variables (only identification variables are anonymised). Researchers from Germany and abroad can apply for a guest stay using a form available on the internet. In this application, researchers have to give the name of the data set, the names of the persons who will be in contact with the data as well as a description of the research design of the respective project. Further information to accomplish the legal regulations specified in the Social Code X (cf. Allmendinger / Kohlmann 2005) is required. The research data centre asks the Federal Ministry of Economics and Labour (BMWA) to authorise data access by the applicants. Then, a contract between the research data centre and the researcher is filed. This contract allows access to the specified data in the research data centre under strict observance of regulations on data protection. Prior to the guest stay, the researcher develops syntax files (STATA, SPSS or SAS) on the basis of artificial test data files. A maximum stay of two weeks at the research data centre is possible. We do not charge any fees for a guest stay.

[9] For the IAB-Establishment Panel, cf. http: // doku.iab.de / fdz / iabb / Anfrageformular.pdf.

[10] Artificial test data files are not usable for content-related analyses. They merely depict the structure and the organisation of the original datasets.

Services to the scientific community

The research data centre provides access to official data from the FES and the IAB on the labour market in Germany subject to the regulations on data protection. Moreover, the research data centre gives additional support for researchers who are interested in the data.

(1) The research data centre's major task is to make new data sets available to the scientific community. On the one hand, this demands the development of new data models and the implementation of those models. On the other hand, the research data centre is updating older data. Our strategy in the current period is to provide standard data models in order to allow for a broad variety of research opportunities. However, we also plan to enable the implementation of highly specialised data wishes in the future as well.

(2) The data of the FES and the IAB are complex, especially since most of them are generated in administrative processes. Therefore, documentations as well as standardized support (program syntaxes, FAQ, etc.) are crucial for an efficient use of the data. The research data centre has established internet pages where researchers can get an overview on the available data and publications based on them, data documentations, information on the access to data, and conferences organised by the research data centre. Moreover, in 2005 the research data centre started two series of publications (FDZ-Datenreporte and FDZ-Methodenreporte[11]) that contain data documentations and information on statistical aspects of the datasets.

(3) In addition, the research data centre provides individual assistance to the scientific community. This relates to actual users of the scientific use files, controlled remote data access and guest researchers in the research data centre as well as to researchers who need counselling on the suitability of data for various research projects.

The research data centre provides neither service for commercial research projects nor does it offer aggregate data. It does not develop processing programmes on behalf of researchers and it does not tranfer non-anonymised data outside of its facilities.[12]

Prospects

Only one year after setting up the research data centre of the FES in the IAB, we can conclude that this was a major step into the right direction.

[11] Data Reports of the Research Data Centre (FDZ) and Methodological Reports of the Research Data Centre.

[12] For these purposes, researchers can turn to the Data Centre of the Statistics Department of the FES (Service-Haus.Statistik-Datenzentrum@arbeitsagentur.de).

Although this period was marked mainly by setting up the basic structures of the research data centre, during the first twelve months more than 170 researchers have already taken advantage of our services. One of the reasons for this relatively fast progress is, of course, previous work done by the FES and the IAB since the middle of the 1990s.

Nevertheless, we aim to reach more milestones during the next years. To develop various samples from integrated databases like the IEBS and to develop factually anonymised versions of them (scientific use files) will be challenging tasks. These tasks are not achievable without the help from the scientific community, of course. Therefore, the research data centre aims at getting scientists from the IAB as well as external scientists involved in its considerations on the development of integrated databases. One example from the past are the LIAB versions that had been developed and implemented by the research data centre. Subsequently, we discussed them with scientists from the IAB in the LIAB working group and with external scientists at our first workshop in November 2004. The first user conference on the data of the FES and the IAB took place in July 2005. Here again, feedback from scientists about our data products played an essential role in helping the research data centre improve its products.

The strategy of developing standardised datasets has two big advantages: the technical implementation as well as getting the permission from legal authorities result in quite a good cost-benefit ratio in terms of effort and time. Moreover, this strategy allows for reliable knowledge about the data quality. Nevertheless, we cannot cover all research interests by implementing standardised data models. Therefore, in a longer perspective, the plan is to provide project-specific datasets, i.e. datasets developed to meet a single project's specific requirements. This could be projects that are analysing very small regional units or projects that require a specific treatment of variables. Beyond the advantages of project-specific datasets, one has to take into account that due to legal and technical requirements, gaining access to project-specific data takes much more time than getting access to standardised files.

Another focus to be concentrated on in the future is the data access for researchers from other EU countries, given the legal regulations on data confidentiality according to the German Social Code. Since most of the FES and IAB data are administrative, process-produced data, a basic understandi