Schmollers Jahrbuch 124 (2004), 567 – 578
Duncker & Humblot, Berlin

The research data centres of the Federal Statistical Office and the statistical offices of the *Länder*

By Sylvia Zühlke, Markus Zwick, Sebastian Scharnhorst and Thomas Wende*

The complexity of economic and social change and the progress made in science and information technology have led to a fundamental change in modern societies' need for data. The data required to analyse and shape modern societies must in particular provide information on social sub-groups and allow to perform analyses of economic and social change on the basis of longitudinal data. Due to the changed information demand, it is no longer sufficient to publish results in the form of tables. To meet the requirements in terms of methodology and content, it is necessary to present statistical data in a way meeting the data demand of the scientific community. This includes providing access to anonymised and non-anonymised microdata which allow to perform more varied analyses.

In this context, an intensive discussion has been going on in Germany over the last few years on granting the scientific community access to microdata of official statistics. Commissioned by the Federal Ministry of Education and Science, the *Kommission zur Verbesserung der informationellen Infrastruktur zwischen Wissenschaft und Statistik* (KVI – Commission to improve the informational infrastructure by co-operation of the scientific community and official statistics) developed several proposals on how to improve the interaction between the scientific community and official statistics (cf. *Kommission zur Verbesserung der informationellen Infrastruktur* 2000)[1]. One of the central institutional recommendations of that Commission was that research data centres be set up as soon as possible at the location of data producers.

That recommendation has been taken up by official statistics. In 2001, the Federal Statistical Office established such a research data centre. Another

* Dr. Sylvia Zühlke and Sebastian Scharnhorst are staff of the research data centre of the statistical offices of the Länder. Markus Zwick and Thomas Wende are staff of the research data centre of the Federal Statistical Office.

1 Cf. Empfehlungen des Wissenschaftsrats zur Stärkung wirtschaftswissenschaftlicher Forschung an den Hochschulen, in: Schmollers Jahrbuch 122 (4), 635 – 652.

research data centre of the statistical offices of the Länder was set up in March 2002 as a joint facility of all statistical offices of the Länder with 16 regional locations. By establishing the research data centres, German official statistics has been intensifying its efforts to make official statistical microdata accessible for scientific analyses.

The purpose of this paper is to present the new forms and ways of using official microdata that have resulted from setting up the research data centres. As an introduction, an overview will be given of how the framework conditions of using microdata of official statistics have developed in Germany. This will be followed by describing the goals and tasks of the research data centres of the Federal Statistical Office and the statistical offices of the Länder and presenting the various possibilities of data use offered by the research data centres.

Using microdata of official statistics in Germany

The use of microdata of official statistics by the scientific community in Germany has been strongly influenced by the development of the Federal Statistics Law, which was first adopted in 1951 and was amended in 1981 and 1987.

The transmission of microdata to third parties was hardly discussed at all when preparing the Federal Statistics Law, so that the 1951 version of the law did not contain any explicit provisions on this issue. As it was not possible to process large amounts of microdata, there was little demand for microdata in the 1950s, 1960s and the early 1970s. Where the transmission of individual data was not explicitly regulated in other laws, formally anonymised microdata were provided for individual projects. Official microdata were analysed by a scientific institution for the first time in the project *"Sozialpolitisches Entscheidungs- und Indikatorensystem für die Bundesrepublik Deutschland (SPES)"* (Socio-political decision-making and indicator system for the Federal Republic of Germany), carried through from 1972 to 1978[2]. For that project, official statistics provided formally anonymised microdata from the microcensus and the sample survey of household income and expenditure. For the same project, a sample of the 1970 population census was made available on the basis of the population census law.

When the demand for microdata grew rapidly along with the progress made in information technology, the transmission of microdata to third parties was explicitly regulated for the first time by adopting the Federal Data Protection Law in 1977 and by amending the Federal Statistics Law in 1981. The pur-

[2] Cf. Krupp, Hans-Jürgen (1973) "Sozialpolitisches Entscheidungs- und Indikatorensystem für die Bundesrepublik Deutschland", Allgemeines Statistisches Archiv 57, 380–387.

pose of introducing the so-called transmission provision in Article 11 of the Federal Statistics Law in 1981 was to allow better supply of microdata to the scientific community. That provision allowed microdata to be transmitted to users in an absolutely anonymised form. Applying that possibility in concrete projects showed, however, that the requirement to be met by such absolutely anonymised data material was so restrictive that data were actually transmitted only in exceptional cases. Due to the amendment of the Federal Statistics Law, only absolutely anonymised microdata could be provided at rather high costs for subsequent projects such as *"Vergleichende Analysen der Sozialstruktur mit Massendaten (VASMA)"* (Comparative social structure analyses by means of mass data) or the Collaborative Research Centre 3 *"Mikroanalytische Grundlagen der Gesellschaftspolitik"* (Microanalytical bases of social policy) of the *Deutsche Forschungsgemeinschaft*[3].

The population census judgment of 1983 and the relevant discussion showed that both the informational self-determination and the freedom of science, both of which are laid down in Article 5 of the Basic Law, are to be treated as important fundamental rights of equal standing. Consequently, the legislators had to ensure adequate data access. This was done by further modification of the Federal Statistics Law in 1987 and the introduction of de facto anonymity for microdata transmission to the scientific community. De facto anonymity now allowed – within the scope of the so-called privilege of science and under specific conditions – to supply microdata to the scientific community that involve a residual risk of disclosure. Subsequently, various projects concretised the shaping of de facto anonymised microdata sets. Especially the results of the project *"Die faktische Anonymisierung von Mikrodaten"* (De facto anonymisation of microdata) permitted from the mid 1990s to provide first standardised and de facto anonymised microdata sets for the area of household and person-related surveys[4].

While the above and other activities outside official statistics[5] created a new data basis especially for issues of the social sciences, it was at first not possible to make similar progress for economics; the reason is that anonymising data on enterprises and local units is more difficult. In its memorandum *"Erfolgsbedingungen empirischer Wirtschaftsforschung und empirisch gestützter*

---

3 For the projects, cf. Hauser, Richard (ed.): Mikroanalytische Grundlagen der Gesellschaftspolitik: Ergebnisse aus dem gleichnamigen Sonderforschungsbereich 1 / 2, Berlin 1994. The results of the VASMA project are documented at http: // www.gesis.org / Dauerbeobachtung / Mikrodaten / Daten / brd / literatur.pdf.

4 Cf. especially Müller, Walter / Blien, Uwe / Knoche, Peter / Wirth, Heike et al. "Die faktische Anonymität von Mikrodaten", vol. 19 of the publication series Forum der Bundesstatistik, Statistisches Bundesamt (ed.), Wiesbaden 1991.

5 Outside official statistics, some surveys have become established that are conducted regularly and are available for scientific analysis, in particular the Socio-Economic Panel (SOEP) and the Population Survey of the Social Sciences (ALLBUS).

wirtschafts- und sozialpolitischer Beratung" (Conditions for empirical economic research and empirically supported consulting for economic and social policies to be successful) the scientific community addressed the problem of not having access especially to data on enterprises and local units[6]. It was demanded there that access to microdata that are hard to anonymise should be granted at the location of the data producers. The memorandum and the symposium *Kooperation zwischen Wissenschaft und amtlicher Statistik – Praxis und Perspektiven*[7] (Co-operation between the scientific community and official statistics – practice and prospects) held in 1999 gave new impetus to the discussion on granting microdata access to the scientific community, which involved the political level, too.

Subsequently, the *Kommission zur Verbesserung der informationellen Infrastruktur zwischen Wissenschaft und Statistik (KVI)* (Commission to improve the informational infrastructure between the scientific community and statistics) developed a number of recommendations on how to improve the co-operation between the scientific community and statistics[8]. They range from involving data users in the development of collection and processing programmes and the prospects of modern education and training in statistics to the various possibilities for public data producers to grant the scientific community access to their microdata. A major institutional demand refers to setting up research data centres at the location of data producers and establishing service centres. The recommendations given by the *Kommission zur Verbesserung der informationellen Infrastruktur* are now being implemented by the *Gründungsausschuss des Rates für Sozial- und Wirtschaftsdaten* (Committee Establishing the Council for Social and Economic Data), so that a number of research data centres and service centres have already taken up their activities[9].

Among them are the two research data centres of official statistics. Although the two institutions are independent of each other, they closely coordinate their activities to make a common offer to the scientific community for improved data access. Following the KVI recommendations, which specify that setting up an efficient data infrastructure is a task of research promo-

---

6 Cf. Hauser, Richard / Wager, Gert / Zimmermann, Klaus: "Erfolgsbedingungen empirischer Wirtschaftsforschung und empirisch gestützter wirtschafts- und sozialpolitischer Beratung: Ein Memorandum", Allgemeines Statistisches Archiv 82, 369 – 379.

7 The results of the symposium are documented in: Müller, Walter / Schimpl-Neimanns, Bernhard / Krupp, Hans-Jürgen / Wiegert, Rolf et al. "Kooperation zwischen Wissenschaft und amtlicher Statistik – Prais und Perspektiven – Beiträge zum Symposium am 31. Mai / 1. Juni 1999 in Wiesbaden", Forum der Bundesstatistik, Vol. 34, Statistisches Bundesamt (ed.), Wiesbaden 1999.

8 Cf. footnote 1

9 For an overview cf. Lüttinger, Paul / Schimpl-Neimanns, Bernhard / Wirth, Heike / Papastefanou, Georg: "Mikrodaten (German Microdata Lab): Das Servicezentrum für amtliche Mikrodaten bei ZUMA", ZUMA Nachrichten 5, 2003, 153 – 172.

tion, both the Federal Statistical Office and the statistical offices of the Länder have submitted a promotion request to the Federal Ministry of Education and Research. With the help of such funds, it is intended to offer a comprehensive range of services and data.

Goals and functions of research data centres

The major goal of the research data centres of the Federal Statistical Office and the statistical offices of the Länder is to facilitate access to microdata of official statistics for the scientific community by establishing various ways of data use. A major prerequisite for achieving that goal is a fundamental improvement of the data infrastructure by setting up a system (centralised in terms of subject-matter) of data maintenance for selected statistics and by establishing a metadata information system.

In Germany, most of the statistical surveys are conducted in a decentralised manner by the statistical offices of the Länder, which means that over 90 percent of all microdata of official statistics are collected, processed and stored there. Since, however, scientific analyses generally refer to several Länder or the entire Federal Republic, the statistical offices of the Länder are currently setting up a system of data maintenance that is centralised in terms of subject-matter. This will allow to use the microdata of official statistics for all Länder at all regional locations of the research data centres.

Scientific users wishing to analyse and interpret the microdata of official statistics also need comprehensive information on the data sets as well as on data collection, processing, and quality. Therefore, the research data centres of the Federal Statistical Office and of the statistical offices of the Länder will develop a web-based metadata information system available for users to obtain information on surveys of official statistics.

Ways of data use

To allow the scientific community to have access to the entire range of information of official statistics, the research data centres of the Federal Statistical Office and of the statistical offices of the Länder have been establishing various ways to access their microdata. This has been providing users with additional and more enhanced possibilities of analysing microdata of official statistics than has been possible so far.

The idea underlying such additional ways of use is that avoiding reidentification of respondents should be ensured not only through modifying the data material but also by controlling data access. The various ways of use thus are the result of different combinations of data anonymisation and access control.

a) Absolutely anonymised microdata sets

Absolutely anonymised data are modified by aggregation or by deletion of individual variables to an extent making it impossible – as far as anyone can judge – to identify respondents. Official statistics offers absolutely anonymised microdata in the form of so-called Public Use Files (PUF). They may be made available to anyone interested.

So far, such data sets have been compiled for statistics of public assistance and for the time use survey. In that segment, too, the research data centres of the Federal Statistical Office and of the statistical offices of the Länder have been increasing their efforts to extend the range of data offered. Another major target area of Public Use Files is higher education. The research data centres are currently developing so-called Campus Files that may be used at institutions of higher education for teaching purposes. Campus Files of the Microcensus, the statistics on public assistance and cost structure in the crafts sector are already available for free download at www.forschungsdatenzentrum.de. Campus Files of the sample survey of household income and expenditure and the wage and income tax statistics will follow in 2005.

b) De facto anonymised microdata sets

The disadvantage of absolute data anonymisation is that it involves a considerable loss of statistical information. Microdata are referred to as being de facto anonymised when – although disclosure cannot entirely be ruled out – the data can be matched with the relevant unit only by making unreasonable efforts in terms of time, cost, and labour[10]. Consequently, the main goal of de facto anonymisation is to reduce the number of possible matches of characteristics items with the relevant units by carefully reducing and modifying the information while, at the same time, not much affecting the statistical information content. Costs and benefits of disclosure must be analysed for every individual survey, and different disclosure methods may be applied[11]. It is laid down in the Federal Statistics Law, however, that de facto anonymised data can be made available only to scientific institutions and only to carry through scientific projects.

De facto anonymity thus results not only from the real information content of the data but also from the existing disclosure possibilities. Therefore, the question of whether a microdata set may be referred to as de facto anonymous

10 That regulation is based on Art. 16 para. 6 of the Federal Statistics Law.

11 An overview of anonymisation methods is given in Köhler, Sabine: "Anonymisierung von Mikrodaten in der Bundesrepublik und ihre Nutzung – Ein Überblick." Forum der Bundesstatistik, Vol. 31, Statistisches Bundesamt, 1999, 133 – 144.

depends especially on the framework conditions under which the data are processed. It is of crucial importance here what additional knowledge is available and where the data are used. Depending on whether the microdata are used outside or within the statistical offices, de facto anonymity can be achieved by modifying the information more or less strongly.

The scientific community has frequently, and very clearly expressed their wish to use microdata in an anonymised form at their own workstations. De facto anonymisation allows to transmit such microdata that are not fully anonymised for external (off-site) use in scientific institutions. As, however, the mere fact of transmitting the data involves a higher disclosure risk than does the use within a statistical office, data anonymisation is rather intensive. The data sets created for that type of use are referred to as Scientific Use Files (SUF).

As regards person-related data, official statistics already offers a wide range of data as Scientific Use Files, that is the microcensus, the sample survey of household income and expenditure, the wage and income tax statistics and the time use survey. The research data centres of the Federation and the Länder endeavour to successively extend that range. A Project is currently being carried out to anonymise the statistics of diagnoses. What is planned for next year is the first standardised anonymisation of a so-called employer-employee data set, that is the structure of earnings survey. With its project "De facto anonymisation of microdata of economic statistics", official statistics endeavours to develop, jointly with scientific users, anonymised standard files also in this area[12].

c) Project-related de facto anonymisation for on-site use

Where demand for specific statistics is low and where microdata are hard to anonymise, it will in many cases not be reasonable to create standardised Scientific Use Files in a highly complex procedure. Project-related data anonymisation is more reasonable here; its advantage is that what is anonymised is not the entire statistics but only the variables required.

Project-related anonymisation also creates de facto anonymity. However, the relevant data can only be analysed on the premises of the research data centres of the Federation and the Länder, using so-called safe scientific workstations. As in this case the microdata will not leave the premises of official statistics, as the users cannot use the data in any way they like, and they can

[12] Cf. Sturm, Roland: "Wirtschaftsstatistische Einzeldaten für die Wissenschaft" in: Wirtschaft und Statistik 2, 2002, 101 – 109 and the results of the workshop "Anonymisierung wirtschaftsstatistischer Einzeldaten" held in Tübingen. The results are documented at http: // www.uni-tuebingen.de / iaw / fawe-nutzertagung.html.

hardly combine them with additional information, that form of use involves another major advantage. As de facto anonymity is achieved here already with much smaller modifications to the data material than is necessary for the creation of Scientific Use Files for off-site use, more information will remain in the data material.

To make on-site use of microdata even more attractive and to ensure the regional availability of that type of use, the research data centres set up safe scientific workstations in all statistical offices. There, the data may be analysed by means of standard programmes for statistical analysis.

d) Using official microdata through controlled remote data processing

Using protected data stocks through controlled remote data processing is a rather recent development, whose importance will grow in the future[13]. That method allows scientists to use the information potential of microdata material that is not anonymised, or only formally anonymised, without having direct data access themselves. The scientists develop analysis programmes (syntax scripts) which are then applied to the original data by the staff of the research data centres. That service is currently offered for the programmes SPSS, SAS and STATA. In contrast to Scientific Use Files, controlled remote data processing is not restricted to a specific group of persons, thus enabling also foreign scientists and non-scientific persons to use microdata of official statistics.

For practical application of controlled remote data processing, the research data centres of the Federal Statistical Office and the statistical offices of the Länder make data structure files available, allowing the users to tailor their analysis programmes to the structure of the original data. Such data structure files represent the data structure of the original data set, without transporting subject-related information. The material thus is identical to the original material as regards the structure of variables, number of data record places, and length of data record. Through a technical process, however, the data have been falsified, so that only synthetic data sets without content are available. Data structure files are currently available for the microdata of the microcensus and the wage and income tax statistics.

Today, controlled remote data processing is a rather time-consuming procedure because, first of all, the programme syntax has to be checked for possible disclosure strategies and the data output must subsequently be checked for cases where data have to be kept secret. Those work steps still have to be done

[13] First experience has been acquired here in the field of tax statistics. Cf. Zwick, Markus: "Individual tax statistics data and their evaluation possibilities for the scientific community" in: Schmollers Jahrbuch 121 (4) 2001, 639 – 648.

manually. Although first automated procedures are available now for such checks, it is not possible yet even with those approaches to fully automate controlled remote data processing.

It is therefore a major goal of the research data centres to develop methods allowing to further automate controlled remote data processing. Starting points are offered especially by the LIS / LES database and by an online data processing method applied in Denmark.[14] The LIS / LES database permits direct Internet-based access to the microdata of the Luxembourg Income Study / Luxembourg Employment Study[15]. By giving a project-related password, the transmission of SAS, SPSS or STATA files allows to automatically start data analysis. LIS / LES is designed to include limited checks of syntax and results.

The Danish model, however, enables scientists at their workstations to directly access a server that is set up and maintained by the statistical office for research purposes[16]. Scientists are granted access to a directory where only those data are stored that they need for their research. They may then copy the data into a working directory specifically created for them; from there the analyses are carried through. The results are sent automatically by e-mail. The statistical office can check any time during the entire process of data processing whether the contractual rules are being adhered to.

The Federal Statistical Office of Germany is currently running a pilot project to develop a client-server solution for access to official microdata from outside the statistical offices. That prototype is called SAM "Server for Access to Microdata" and is based on the experience two employees of the Federal Statistical Office, Jobst Heitzig and Tom Wende, have made with the Danish system.

The various ways of data use of the research data centres as presented here may be combined with each other. Irrespective of which kind of data access is chosen, however, data provision is subject to earmarking. This means that the microdata may not be made available for general research purposes, but only for a well defined research project that is limited to a specific period of time.

---

14 The programme μ-Argus offers the possibility that data that are available in the form of tables are kept secret in an automated manner. The programme was developed by the Statistical Office of the Netherlands for Eurostat and was extended for cross-table confidentiality by the Land Office for Data Processing and Statistics of North Rhine-Westphalia. It is currently undergoing substantial testing in Germany.

15 Cf. in detail Smeeding, Timothy M. / Jesuit, David K. / Alkemade, Paul: The LIS / LES Project Databank: Introduction and Overview. In: Schmollers Jahrbuch, Zeitschrift für Wirtschafts- und Sozialwissenschaften 122 (3), 2002, 497 – 517.

16 The model is described in the report "Access to Microdata in the Nordic Countries", which was published in 2003 by the Statistical Office of Sweden.

Special processing

Apart from the above ways of data use, any group of users may of course order some special processing. This refers to data evaluations that are tailored to the special information needs of a specific user and where using other ways of data use is not possible or insufficient. In contrast to controlled remote data processing, the processing programmes are not developed by the users but by official statistics. For that purpose, official statistics and users discuss the data requirements in order to make them concrete, thus permitting to develop a processing programme. Then the data material, which has been anonymised just formally or not at all, is evaluated by means of the developed programme. After having been checked for cases where data have to be kept secret, the results are transmitted. Thus users do not get in direct contact with microdata that have been anonymised just formally or not at all.

Data demand of the scientific community and preferred uses

As the research data centres of the Federal Statistical Office and the statistical offices of the Länder intend to further develop the services they offer, taking account of the demand of the scientific community, they regularly ask potential users for their preferences. To be able to take account of the concrete data demand of the scientific community when developing the services to be offered, the research data centre of the statistical offices of the Länder conducted a user survey in summer 2002[17]. The purpose of the survey was to contact potential users of the data to be offered in the future by the research data centres and to identify their concrete data needs. Respondents also had the opportunity to comment on the various types of use, the analysis programmes they use and on whether they are interested in events planned by the two research data centres.

Among the 700 scientists interviewed, almost 600 indicated that they use, or will need, microdata within the scope of their scientific activities. Altogether, the results of the user survey show that the scientific community is very much interested in using microdata of official statistics. As the data demand as indicated in the survey covers a very wide range of subjects, it will not be possible for the research data centres to focus on making available just a few statistics; over the medium term, they will have to offer a wide range of data.

---

17 Cf. in detail Zühlke, Sylvia / Hetke, Uwe: "Datenbedarf und Datenzugang: Ergebnisse der ersten Nutzerbefragung des Forschungsdatenzentrums der statistischen Landesämter", in: Allgemeines Statistisches Archiv 3, 321 – 334.

As regards the ways of data use offered, the survey shows a clear preference for using de facto and absolutely anonymised data at the user's own workstation. Working at safe scientific workstations and controlled remote data processing did not meet with much interest by the scientists when the survey was conducted. It will however not be possible to offer the results of all relevant surveys of official statistics in the form of anonymised data sets. Therefore, the research data centres are planning to considerably improve the attractiveness of the other ways of use by shaping them in a way better meeting user requirements. In particular setting up safe scientific workstations at all regional locations of the research data centres will considerably improve the regional availability of the services offered by official statistics, thus making microdata access much easier.

Prospects

Further developing the data infrastructure and establishing various ways of data use, as described above, will dramatically improve the microdata basis for scientific analysis. However, by setting up research data centres at the location of various public data producers, the discussion about access of the scientific community to the information potential of public data producers is not finished. Future issues of data access will regard general regulations and, in particular, the availability of internationally comparable microdata.

For the discussion about a general regulation of data access, the Committee Establishing the Council for Social and Economic Data uses the term "research data confidentiality". According to the proposal of the Committe, scientists should be considered equal to the staff of data producers, thus improving their rights to use microdata. At the same time, a right for scientists to refuse to give evidence and a prohibition of seizure are planned to prevent unauthorised access to microdata stored by scientists.

A major problem in using microdata for international comparisons in scientific research is that regulations on microdata access are highly different across countries[18]. Consequently, obtaining national microdata from various countries involves considerable efforts. First steps were taken to harmonise access to data obtained through surveys in the European Union by adopting Regulation No. 322/97 and, based on it, Regulation No. 831/2002 for Community statistics[19]. The purpose of those Regulations is to make microdata of

[18] A presentation of the various national approaches to data access is contained in the papers contributed to the international "Workshop on Microdata" held on 21/22 August 2003, http://www.mirco2122.scb.se.

[19] Commission Regulation (EC) No. 831/2002 of 17 May 2002 implementing Council Regulation (EC) No. 322/97 on Community Statistics – concerning access to confidential data for scientific purposes.

all EU countries available, regarding the Labour Force Sample Survey, the European Household Panel, the continuing vocational training survey, and the Community innovation survey. However, implementing the Regulations is currently complicated by the fact that some of the provisions are in conflict with national provisions in European Union countries.

Considering the developments presented above, it is expected that the opportunities of the scientific community to have data access will further improve. The research data centres of the Federal Statistical Office and the statistical offices of the Länder will continue to take an active part in the process by making proposals of their own as to how such access might be shaped.