UNITED NATIONS

E

Economic and Social Council

Distr.
GENERAL

CES/2003/30
15 May 2003

ENGLISH ONLY

STATISTICAL COMMISSION and ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Fifty-first plenary session
(Geneva, 10-12 June 2003)

NEW WAYS OF ACCESS TO MICRODATA OF THE GERMAN OFFICIAL STATISTICS

Supporting paper submitted by the German Federal Statistical Office[1]

I. HISTORY

1. Before going further in this paper about the development of a better informational infrastructure and the work of Research Data Centres, it may be useful to take a short glance at the history of microdata use in Germany. In the past, it was seen as sufficient for data users to work with aggregated data-like tables and indexes given out by the statistical offices. But the accelerating change of society and the increasing amount of new societal questions resulting from this changed the scientific interest and aggregated data was no longer sufficient. The first requests for official statistics microdata by the scientific community were made in the early 1970s. A group of researchers at the Universities of Mannheim and Frankfurt founded a research project called SPES (Sozialpolitisches Entscheidungs- und Indikatorensystem für die BRD) that tried to create a social-political decision- and indication system for the Federal Republic of Germany by using official microdata. From this project evolved the so-called Special Research Sector 3 – "Microanalitic Basics of Society Politics" (Sonderforschungsbereich 3 – SFB 3 – Mikroanalytische Grundlagen der Gesellschaftspolitik), which dealt with matters of social policy and econometrics. This pioneering work, which demonstrated the trend to use microdata for societal research, paved the way for ongoing changes in law and the development of an informational infrastructure for the empirical use of microdata bases.

2. At almost the same time a project called VASMA (VergleichendeAnalysen der Sozialstruktur mit Massendaten) dealt with the comparative analysis of the social structure by population data.

[1] Prepared by Tom Wende.

II. LEGAL BASIS

3. The first legal regulation for the use of official microdata was included in the federal law on statistics in 1981. It allowed the passing on of completely anonymised microdata in §11 (5) BStatG. This, of course, resulted in a lot of restrictions, as complete anonymisation always involves a certain loss of information. Nevertheless, this was a ground-breaking law offering the first legal opportunity for official statistics to distribute so-called Public Use Files[2], which are completely anonymised datasets of official statistics. It also paved the way to continue.

4. A more satisfying solution for empirical researchers was the next legal improvement: The federal law on statistics in 1987 – specifically in § 16(6) BStatG - introduced the so-called "Privilege of Science", which means that from then on, scientists were allowed to receive factually anonymised microdata[3].

Excursus 1: The Development of Anonymisation Criteria

5. Between 1988 and 1991, a large-scale research project aimed at anonymisation of selected microdata was performed. Representatives of the German statistical offices worked alongside representatives of the data protection registrars and of empirical science, e.g. the University of Mannheim and ZUMA – the Centre for Survey Research and Methodology. In the course of this project some measures were developed for a specific factual anonymisation of the Microcensus and the Sample Survey of Income and Expenditure. The results of this research project culminated in two reports: "Textbook for the building of factually anonymised data regarding the Microcensus" and "Textbook for the building of factually anonymised data regarding the Sample Survey of Income and Expenditure".

End of Excursus

III. NATIONAL AND INTERNATIONAL DATA REQUEST

6. In the previous section, the access so-called Public (PUF) and Scientific Use Files (SUF) was described. You may ask yourself why that is such an important development: the example of international data access will illustrate this. Before PUF and SUF, access to official microdata was almost impossible. After 1981 – with the implementation of Public Use Files - data access was possible for everyone, but with a lot of restricted and non-accessible information. After 1987 the researcher´s requirement for less restricted data was solved with the invention of Scientific Use Files, which are today only provided to German nationals. But what if a researcher from a foreign country wants access

[2] see SCIENTIFIC USE FILES AND PUBLIC USE FILES

[3] Factual Anonymisation denotes data that is not absolutely anonymised, so there is a chance of disclosure for a potential intruder, but the expense of disclosure is much higher than the use for the intruder; also see SCIENTIFIC USE FILES AND PUBLIC USE FILES

to German official data? Today, this is very difficult due to the lack of international data access regulations. First solutions are proposed by the EC-Regulations 322/97 and 831/2002, which give access possibilities to common microdata. The 831/2002 commision act specifies this access for four European surveys at least for members of the EC: the CVTS (Continuing Vocal Training Survey), CIS (Community Innovation Survey), LFS (Labour Force Survey) and ECHP (European Community Household Panel). Possibilities offered are, for example, the controlled remote data processing[4] or the possibility for a visiting researcher to work in the protected area of the German statistical offices[5].

IV. RESEARCH DATA CENTRES (RDC) OF THE OFFICIAL STATISTICS

7. In 1999, the Federal Ministry on Education and Research (Bundesministerium for Bildung und Forschung - BMBF) created a Commission for the Improvement of the Informational Infrastructure (Kommision zur Verbesserung der Informationellen Infrastruktur – KVI). It was the constitutional task of this Commission to revise the informational infrastructure of the Federal Republic of Germany (BRD) and to work out new concepts for the exchange of data between the scientific community and data producers. The KVI worked out a number of recommendations which are elaborately prescribed in their final report[6].

8. One first elementary recommendation of the KVI was the establishment of so-called Research Data Centres (RDC). Implementation was almost immediate. On 1 October 2001, a Research Data Centre of the Federal Statistical Office of Germany (Statistisches Bundesamt) was established in Wiesbaden. On 1 April 2002, RDCs in the Statistical Offices of the Federal States (Statistische Ämter der Länder) with one location in each federal state were founded. The Research Data Centres offer a lot of opportunities for microdata access and thus an extraordinary improvement of the informational infrastructure between official statistics and empirical science.

9. The Research Data Centres provide a well-balanced service proposition for users. The RDC of the Federal Statistical Office and of the Statistical Offices of the Federal States are independent but co-operate closely with each other. The main focus of the Federal States Research Data Centres is centralised data storing, a widespread web of Safe Scientific Workstations[7] and the supply of metadata for decentralised surveys. The Research Data Centres of the German Statistical Offices focus on the development of Scientific and Public Use Files[8], the improvement of Remote Controlled Data Processing[9] and the supply of metadata for central surveys. Together, all Research Data Centres are keen on developing a high quality metadata system, consulting data users and compelling further improvement of the informational infrastructure.

---

[4] see CONTROLLED REMOTE DATA PROCESSING AND SPECIAL DATA PROCESSING
[5] see SAFE SCIENTIFIC WORKSTATION
[6] KVI (HRSG.) 2001: WEGE ZU EINER BESSEREN INFORMATIONELLEN INFRASTRUKTUR. BADEN-BADEN: NOMOS VERLAGSGESELLSCHAFT
[7] see SAFE SCIENTIFIC WORKSTATION
[8] see SCIENTIFIC USE FILES AND PUBLIC USE FILES
[9] see CONTROLLED REMOTE DATA PROCESSING AND SPECIAL DATA PROCESSING

10. The main functions of the RDCs are:

- Continuing the further development and implementation of the advice given by the KVI;
- Serving as an interface between official statistics and the scientific community;
- Providing consulting and service for the use of official microdata;
- Creating and providing possibilities for access to microdata with a lower level of anonymisation.

11. The invention of the RDC is a great improvement for the informational infrastructure because, for the first time, there is one elaborate option for the use of official microdata. There exist already different ways of access to official microdata like Controlled Remote Data Processing and Safe Scientific Workstations[10]. The RDCs also offer consulting and service for the use of official microdata.

12. Let's now move on to the Research Data Centres´ work in practice.
As was already mentioned, the RDC offer different ways of microdata access:

- Scientific and Public Use Files;
- Safe Scientific Workstations;
- Remote Controlled Data Processing;
- Special Data Processing.

V. SCIENTIFIC USE FILES AND PUBLIC USE FILES

13. One possibility for microdata use is the purchase of Scientific or Public Use File. Different surveys are already available in this format. For example, the Microcensus, the Sample of Income and Expenditure or the Statistics of Road and Traffic Accidents and many more can be obtained as SUF. Available as Public Use Files are, for example, the Time Use Survey or the Social Welfare Statistics.

14. One important aim of the Research Data Centres is to significantly broaden the range of PUF and SUF in the near future.

15. Scientific Use and Public Use Files are anonymised with different grades of anonymisation. The Public Use Files no longer offer possibilities to draw conclusions about single cases in the surveyed population. The Scientific Use Files do theoretically offer this possibility, but the expense is much higher than the disclosure of the factually anonymised data[11]. The rights to use Scientific Use Files are restricted by the German Statistics Law – as the name implies – to the scientific community. That is another confidentiality function of these files, because in the case of a breach of confidentiality, the German researcher can be prosecuted by law. The advantage of giving out anonymised files is that a researcher is able to work with his own Software on his own PC; the disadvantage is the loss of information resulting from anonymisation.

---

[10] see CONTROLLED REMOTE DATA PROCESSING AND SPECIAL DATA PROCESSING
11 §16(6) BStatG

16. The Research Data Centres offer further possibilities of data access, which in combination with the supply of Public and Scientific Use Files close the circle of informational infrastructure and in combination with each other are able to provide a good balance between empirical research interests and data confidentiality. Specifically, there are the Safe Scientific Workstations and the option of Controlled Remote Data Processing (Special Data Processing), which will be described in the following chapters.

VI. CONTROLLED REMOTE DATA PROCESSING AND SPECIAL DATA PROCESSING

17. If a researcher requires more information than a Public or Scientific Use File can offer, or if there is no standardised SUF or PUF yet available for a certain survey, there are ways to work with less or even non-anonymised data via the Research Data Centres. One way is to work as a first step with the anonymised dataset, for example a standardised Scientific or Public Use File - or if a SUF is not available with a so-called structural dataset, which corresponds to the original dataset in all structural attributes but not in content attributes – and as a second step to send the thus produced syntax for Software like SAS, SPSS or STATA back to the RDC, where it is processed under internal control over the original data. This is called Remote Controlled Data Processing. A special form of Remote Controlled Data Processing is Special Data Processing. In this form, a proposer informs a representative of the Statistical Office of his research interests, and the representative does the empirical work.

18. One advantage of Remote Controlled and Special Data Processing for data confidentiality is that the computing process is not beyond control and the representatives of the Research Data Centre know exactly what information is given to the researcher. Another advantage is that the output is not microdata but aggregated data in the form of tables, which can be anonymised more easily. The advantage for the researchers is that they have the possibility to make an exact predication about the whole population with a lower standard error and in general a low error variance. Further advantages are that the consulting function of the Research Data Centres can be engaged and there is a possibility to work with company data, a possibility which didn't exist before. The disadvantages are that these processes mean a lot more work and cost for both the researchers and the representatives of official statistics and, as a result, require a lot more time.

VII. SAFE SCIENTIFIC WORKSTATION

19. The RDCs provide another new method of data access in the protected area of the German Statistical Offices. A visiting researcher has the possibility to access microdata over sealed-off computers at especially equipped Safe Scientific Workstations in the statistical offices. By working on a Safe Scientific Workstation, the researcher obtains on-site access to factually anonymised data. The difference in anonymisation between an On-Site Scientific Use File and the given out standardised SUF, which is also factually anonymised, is that the anonymisation criteria in the on-site case is lower, because of other means of confidentiality control, like the fact that the visiting researcher is given no way to transfer data, except for his aggregated output in the form of tables. To further improve this method of

microdata access, the production of many more On-Site SUF is planned.

20. It has become clear in previous sections that the new method of data access by the RDC guarantees confidentiality by means of well-balanced anonymisation. There exists only one way of access to non-anonymised microdata: the cooperation of a statistical office with a researcher in a single research project initiated by the statistical office, also known as a "One-Dollar-Man-Contract". Such a contract can be completed if there is a research interest, which is primarily useful for official statistics and secondarily for an external researcher. It is possible for a researcher then to sign a terminable employment contract (with the symbolic payment of one dollar) with a statistical office and to work with microdata as an employee of that statistical office, therefore bound to confidentiality like every employee of the statistical offices. But that method of microdata access is an exception, which is only used if a statistical office needs external help in a single project. It is mentioned here just to complete the menu of microdata access possibilities in Germany. At the end of this paper, let´s have a brief glance at the projects planned for the further improvement of German microdata access.

VIII. PROJECTS IN THE FUTURE

21. Prospectively the Research Data Centres are working on the expansion of low cost microdata access in the form of standardised Scientific and Public Use Files. Remote Controlled Data Processing will be simplified and improved in the future. Later on there will be an improvement in consultancy capacity for visiting researchers on Safe Scientific Workstations and researchers who use Remote Controlled or Special Data Processing. The RDCs are working on the central availability of all official microdata and also on the elaboration of a widespread metadata-system for all official data.