

Working Paper No. 10
ENGLISH ONLY

**UNITED NATIONS STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE
CONFERENCE OF EUROPEAN STATISTICIANS**

**EUROPEAN COMMISSION
STATISTICAL OFFICE OF THE
EUROPEAN COMMUNITIES (EUROSTAT)**

Joint ECE/Eurostat work session on statistical data confidentiality
(Luxembourg, 7-9 April 2003)

Topic (iii): Emerging legal/regulatory issues

RESEARCH DATA CENTRES OF OFFICIAL STATISTICS

Invited paper

Submitted by the Federal Statistical Office of Germany¹

¹ Prepared by Thomas Wende (thomas.wende@destatis.de).

DEVELOPMENTS IN INFORMATIONAL INFRASTRUCTURE

On October 1st 2001 one Research Data Centre of the German Federal Statistic Office was established in Wiesbaden and one in Berlin. On April 1st 2002 RDCs in the Statistic Offices of the federal states 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 the official statistics and the empirical science.

The Research Data Centres provide a well balanced service proposition for users. They 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 visiting researchers desktops² and the supply of metadata for decentral surveys. The German Statistical Offices Research Data Centres focus is the development of Scientific and Public Use Files³, the improvement of controlled remote data processing⁴ 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 the further improvement of the informational infrastructure.

HISTORY

Beginning this essay about the work of Research Data Centres and the development of a better informational infrastructure, 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 questions resulting from this, changed the scientific interest and aggregated data was not enough anymore. The first requests for official statistics microdata by scientists were in the early 1970s. A group of scientists at the Universities of Mannheim and Frankfurt founded a research project called SPES which tried to create a “social-political decision- and indication system for the federal republic of Germany” (“Sozialpolitisches Entscheidungs- und Indikatorensystem für die BRD”) by using official microdata. From this project evolved the so called special research sector 3 (Sonderforschungsbereich 3 – SFB 3), which until today deals with matters of social policy and econometrics. This pioneering work, which showed the urge to use microdata for societal research paved the way for still ongoing changes in law and the development of an informational infrastructure for the empirical use of microdata bases.

At almost the same time a project called VASMA dealt with the comparing analysis of the social structure by population data.

LEGAL BASICS

The first legal regulation for the passing on of official microdata was made in the federal law on statistics in it's version of 1980. It allowed the passing on of completely anonymised microdata in it's §11 subsection 5. This of course left a lot of restrictions, as the anonymisation always goes along with a loss of information. But it was still an epoch-making invention, as it offered the first legal opportunity for the official statistics to give out so called Public Use Files⁵, which are absolutely anonymised datasets of social statistics, to everyone who needed such information. And it showed a way, how to go on. A more satisfying solution for empirical researchers was the next legal improvement: The federal law on statistics in it's version of 1987 brought up the so called “privilege of science”, which means, that from that point

² see VISITING RESEARCHER DESKTOP AND “ONE DOLLAR MAN”

³ see SCIENTIFIC USE FILES AND PUBLIC USE FILES

⁴ see CONTROLLED REMOTE DATA PROCESSING AND SPECIAL DATA PROCESSING

⁵ see SCIENTIFIC USE FILES AND PUBLIC USE FILES

on, scientists were allowed to receive factual anonymised microdata. Factual Anonymisation means, that the data is not absolutely anonymised, so there is a chance to de-anonymise data and to draw conclusions back to single persons or organisations, but the effort of de-anonymisation is so enormous, that it is not worth trying⁶.

Excursus 1: The Development of Anonymisation Criteria

Between 1988 and 1991 a large-scale research-project aiming for anonymisation of selected microdata has been performed. Representatives of the federal statistical office and of the statistical offices of the federal states worked alongside with representatives of the data protection registrars of the federal states and the Federation, the University of Mannheim and the ZUMA – the Centre for Survey Research and Methodology. This project was directed by Prof. Dr. Walter Müller at the University of Mannheim, mainly because this project was about the factual anonymisation of person- and household related microdata which can be of particular meaningfulness for the social sciences. In the course of this project some measures have been developed for a specific factual anonymisation of the sample survey of income and expenditure and the Microcensus. Special issues were for example coarsening of data files and drawing of sub-samples. The results of this research project culminated in two similar reports: “Textbook for the building of factual anonymised data regarding the Microcensus” and “Textbook for the building of factual anonymised data regarding the sample survey of income and expenditure”.

End of Excursus

NATIONAL AND INTERNATIONAL DATA REQUEST

In the past chapter the way to so called Public (PUF) and Scientific Use Files (SUF) was described. You will maybe ask yourself, why that is such an important development. The example of international data access will show you. Before PUF and SUF the access to microdata was hardly possible at all. After 1980 – with the invention of Public Use Files - data access was possible for everyone, but with a lot of restricted and non-accessible information. After 1987 the scientists need for less restricted data was answered with the invention of Scientific Use Files, which are only provided to german scientists by now. But what if a scientist from a foreign country wants access to german official single data? By now it's almost impossible. One first Solution is the EC-Regulation 831/2002, which regulates the data access possibilities for members of the EC, such as controlled remote data processing⁷ or the possibility for a guest researcher to work in the safe area of the statistical offices⁸. For Non-EC-Countries the legal situation is still not satisfying at all.

RESEARCH DATA CENTRES (RDC) OF THE OFFICIAL STATISTICS

Due to this exigency, resulting from the conflict between data-protection and the eminently reasonable interest in microdata access by the science community, the federal ministry on education and research (BMBF) originated a “commission for the improvement of the informational infrastructure between science and statistics” (KVI). It was the constitutional task of the commission to revise the informational infrastructure of the federal republic of Germany (BRD) with respect to it's capacity and to work out new concepts on the exchange of data between science and statistics. The KVI worked out a number of advice which are elaborately prescribed in their final report⁹.

⁶ see SCIENTIFIC USE FILES AND PUBLIC USE FILES

⁷ see CONTROLLED REMOTE DATA PROCESSING AND SPECIAL DATA PROCESSING

⁸ see VISITING RESEARCHER DESKTOP AND “ONE DOLLAR MAN”

⁹ KVI (Hrsg.) 2001: WEGE ZU EINER BESSEREN INFORMATIONELLEN INFRASTRUKTUR. BADEN-BADEN: NOMOS VERLAGSGESELLSCHAFT

One first elementary advice of the KVI was the establishment of so called Research Data Centres (RDC). The implementation of this advice almost immediately started. The main functions of the RDCs are:

- a) carrying on the further development and implementation of the advice given by the KVI.
- b) serving as an interface between official statistics and science.
- c) providing consulting and service for the use of official microdata.
- d) creating and providing possibilities for access to microdata with a lower level of anonymisation.

The invention of the RDC is a great improvement for the informational infrastructure because for the first time, there is one place for every service related to official microdata access. There already are different ways of access to official microdata like controlled remote data processing and visiting researcher desktops¹⁰. The RDCs also offer consulting and service for the use of official microdata.

Let's now talk about the Research Data Centres work in practice.

As it was already mentioned, the RDC offer different ways of microdata access:

- Scientific and Public Use Files
- Visiting Researcher Desktop
- Controlled Remote Data Processing
- Special Data Processing

SCIENTIFIC USE FILES AND PUBLIC USE FILES

The first possibility for a scientist to get access to microdata is to purchase a Scientific or Public Use File. Different surveys are already available in that format. For Example, you can get different waves of the Microcensus, the Sample of Income and Expenditure or the Statistic of road and Traffic Accidents and many more as SUF. Available as Public Use Files are for example different waves of the Time Use Survey, the Wage and Income Statistics or the Social Welfare Statistics.

One important aim of the Research Data Centres is to enormously broaden the range of PUF in the near future.

Scientific and Public Use Files are anonymised with different grades of anonymisation. The Public Use Files offer no way to draw conclusions about single characteristic carriers anymore. The Scientific Use Files do theoretically offer that possibility, but the expense is much higher than the use of de-anonymising the factually anonymised data¹¹. The rights to use Scientific Use File are reserved – as the name implies – to scientists (at the moment nearly unexceptional to german scientists). That is another confidentiality function of these files, because in case of a breach of confidentiality the scientist can be prosecuted by law. The advantage of giving out anonymised files is, that the scientist is able to work with his own Software on his own PC, the disadvantage is the loss of information resulting from anonymisation and following from that the difficulty to close from the sample to the complete population.

Excursus 2: Anonymisation Procedures

For better understanding of the problems it is necessary to know, how anonymisation is realised in practise. Basically there are three ways of anonymizing data: Enlarging the pitch, clearing critical data and drawing of samples. For better understanding you will be given an example: Imagine an offender gets paid by a company for finding out strategic information about competitors of this company (This example was chosen because of its clearness. Actually it is not yet possible to anonymise company data to an extend which is good enough to give out a Scientific Use File, but right now specialists from the official statistics are working together with highly decorated scientists to find a solution for that problem). First

¹⁰ see CONTROLLED REMOTE DATA PROCESSING AND SPECIAL DATA PROCESSING

¹¹ §16(6) BStatG

of all de-anonymisation can occur, if an offender connects different information – maybe different variables in the data set or external information about the population. With Scientific Use Files de-anonymisation is most of the time impossible without additional knowledge about the population, because the internal critical variables are already anonymised. So the offence scenario bases on the thought, that a data offender is able to connect knowledge which he receives from the data set with knowledge from other sources like the Internet or other Scientific Use Files of earlier surveys about the same topic. Let's see ways of reacting: Enlarging the pitch means for example if you have a very small high turnover class of companies in one region - maybe less than three - you can easily find out who these three are by Press or Internet and then research concrete data in the dataset. De-anonymisation is very easy preventable if you subsume high and very high turnover in one category or if you enlarge the geographic area, i.e. if you subsume two or three small regions to one like the federal states of the different points of the compass into north, south, east and west. Another possibility is to cut off critical data, like the turnover numbers, but that goes along with a high loss of information and research quality. And the third way is to draw a sample of the original file with the result, that the offender does not really know, if all companies are in the sample and therefore can not exactly estimate if all very high turnover companies are in his sample.

End of Excursus

If SUF and PUF were the only ways to research microdata, a lot of empirical questions would remain unanswered. But the Research Data Centres offer some more 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 more satisfying balance between empirical research interests and data confidentiality. Particularly there are the visiting researchers desktop and the option of controlled remote data processing respectively special data processing, which will be described in the following chapters.

CONTROLLED REMOTE DATA PROCESSING AND SPECIAL DATA PROCESSING

If the researcher needs more information, than a Public or Scientific Use File can offer, there is a way to work with less or even non-anonymised data via the Research Data Centres. One way is to work in a first step with the anonymised dataset, for example a Scientific Use File - or if a SUF is not available with a so called structural dataset, which corresponds with the original dataset in all structural attributes but not in content attributes – and in a second step send the so 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 Controlled Remote Data Processing. A special form of Controlled Remote Data Processing is Special Data Processing, where the scientist tells his research interest to a representative of the Statistical Office and the representative does the empirical work. As you can imagine, Special Data Processing is more cost-intensive than simple Controlled Remote Data Processing. And beyond it is supposed to be dissatisfying for a scientist to let others do his work.

One advantage of Controlled Remote and Special Data Processing for data confidentiality is, that the computing process is not beyond control and the representatives of the Research Data Centre exactly know what information is given to the researcher. Another advantage is, that the output is not microdata but aggregated data in form of tables, which can be anonymised easier. 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, which wasn't given before. The disadvantages are, that these processes mean a lot more work and cost for both the scientist and the representative and as a result need a lot more time.

VISITING RESEARCHER DESKTOP AND “ONE DOLLAR MAN”

The RDCs provide another new way of data access in the protected area of the German Statistic Offices. The empirical researcher gets the possibility to access microdata over sealed-off computers at the visiting researchers desktop in the statistical offices. Generally there are two different way of data access for a visiting researcher. One way – the Visiting Researcher Desktop – is the future, the other way – the “One-Dollar-Man” – is the past. In the past, it was possible for a scientist to sign a terminable employment contract (with the symbolic payment of one Dollar) with a statistical office and work with microdata in the area of the statistic offices as an employee and therefore bound to confidentiality like every employee of the statistical offices. But that was a very inordinate solution and is now replaced by the regulated way of the visiting researcher desktop, where the researcher stays employee of his actual institution and gets on-site-access to factually anonymised data as a guest of the Research Data Centre. The difference in anonymisation to the given out Scientific Use File, which is also factually anonymised, is that the anonymisation criteria in that case is lower, because of other means of confidentiality control, like the fact that the guest researcher is only able to take aggregated data – in form of tables – out of the statistical office. Also he is given no way of data transfer, except for his aggregated output. The researcher is given a special password protected folder, where he is given the ability to save his research data for limited time in the statistical office for further work.

PROJECTS IN THE FUTURE

Prospectively the Research Data Centres are working on the Expansion of low cost microdata access in form of Scientific and Public Use Files. Also the production of PUF and SUF for on-site-use will be forced. Controlled Remote Data Processing will be simplified and improved in the future. Further on there will be an improvement of consultancy capacity for visiting researchers and researchers who use controlled remote 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.