

Working Paper No. 10 (Summary)

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: THE FUTURE OF COOPERATION BETWEEN  
EMPIRICAL SCIENCE AND OFFICIAL STATISTICS**

**Invited paper**

Submitted by the Federal Statistical Office of Germany<sup>1</sup>

---

<sup>1</sup> Prepared by Thomas Wende (thomas.wende@destatis.de).

## **Research Data Centres: The future of co-operation between empirical science and official statistics**

The unimpeded use of official microdata should be a self-evident part of empirical research. Unfortunately the reality in the last years showed a different face. The discrepancy between data-confidentiality-laws and researcher's interests caused many bottle-necks in the informational infrastructure - especially in questions of international interest.

The *privilege of the empirical science*, which was prescribed by the data protection law (§ 16 Abs. 6 BStatG), brought a first remedy. It gave the official statistics the possibility to offer so-called factually anonymised data sets to the empirical researchers. That was a first important step but it was of course only a single step and not enough to already solve the problem.

The pressure, resulting from the conflict between data protection law and the empirical researcher's self-evident interest in microdata has caused an intense effort to find ways for an improvement of informational infrastructure in the last years. The German federal ministry of education and research founded a commission for the improvement of the informational infrastructure (KVI) between science and statistics.

- **Research Data Centres (RDC)**

This commission came up with suggestions on how to improve the interaction between scientific research and statistics. For the realisation of these suggestions so called Research Data Centres (RDC) were established. The German federal statistic office set up a Research Data Centre in 2001 and the statistical offices of the federal states set up a joint RDC in 2002. Let's now look at the RDCs work and how they are able to offer the necessary service to improve the informational infrastructure.

The Research Data Centres (RDC) offer a new effective way of providing official microdata to the empirical oriented science by guaranteeing a high safety in data protection and confidentiality. The different instruments to make this possible are: Providing of factual anonymised "Scientific Use Files" (SUF) and "Public Use Files" (PUF), controlled remote data processing and a visiting researcher desktop in each RDC.

- **Scientific Use Files (SUF) and Public Use files (PUF)**

Scientific Use Files (SUF) contain factually anonymised microdata. I.E. for an attacker the effort to deanonymise information about single persons or households is more lavish than the use of the de-anonymisation. Public Use Files (PUF) are fully anonymised. PUF can be accessed by everyone for example via internet or CD-ROM. The disadvantage of PUF is, that they are often too anonymous to base a reasonable research on them without a high loss of information.

Yet –mainly because of often small sample-sizes - it is unfortunately not possible to create SUF for enterprise or organisation samples, but a research project of the German statistical office is about to find a solution to that point. The difference between SUF and PUF is, that the SUF are less anonymised but only accessible by scientists who are under contract and prosecutable by law in case they injure confidentiality. Accordingly the lack of an internationally prosecutable law for breach of contract makes

it difficult – especially for non-EU researchers to access German official data. One possible solution for that problem could be the so-called „controlled remote data processing“:

- **Controlled Remote Data Processing**

The idea is very simple: To prevent disuse of data the scientist gets a structural dataset, which has exactly the same structure as the original dataset (same variable a.s.o.) but is filled with no information. With it he can write his analysis program at his own workstation with a standard software program (as SPSS or SAS) and then send it to the RDC. There the program is checked for Trojans or other confidentiality risks, applied to the original, non-anonymised dataset and finally the aggregated results are send back to the researcher.

- **Visiting Researcher Desktop**

If the microdata can't come to the expert, the expert can come to the microdata in future. Therefore special visiting researcher desktops are installed both in the federal statistical office and in the statistical offices of every single federal state. Here the researchers are able to work with microdata on especially sealed-off computers, that can not be transferred as SUFs for use outside the statistical offices. The advantage of this system is: It makes it possible for the scientist to work with the original non-anonymised dataset without injuring data protection, because no original data leaves the statistical office.

- **Servicecentre**

Last but not least the RDC will be a place of service and consultancy in questions concerning microdata, data protection, availability of data and many other things. Researchers are not forced to deal with a lot of partners within the Federal Statistic Office to plan and organize a microdata based analysis or to get detailed information about methods and approaches of surveys anymore; the RDC acts as mediator between the different statistical experts and the (external) researchers.