German Microdata Lab

gesis

Microdata access and confidentiality issues in Germany

Heike Wirth
GESIS – ZUMA
Germany

The Samples of Anonymised Records
Census Microdata: findings and futures

Humanities Bridgeford Street, University of Manchester

1 - 3 September 2008

Overview

1. Microdata Access in Germany - „chi va piano va sano“

2. Modes of microdata access

    - Scientific use files (off site access)
    - On site access
    - Remote access
    - Public use files

3. Concluding notes

1 Microdata Access in Germany - „chi va piano va sano“

Access to official microdata in Germany:

- has been strongly influenced by a vigorous public debate on data privacy in the 1980s initiated by the German population census of 1983 (which did not take place until 1987)
  - widespread public anxiety over all kinds of abuse of the census data by the government or third parties who might have access to the data
- As a consequence of this debate the population census was temporarily stopped by a famous verdict (Volkszählungsurteil, 1983) of the German Federal Constitutional Court

1 Microdata Access in Germany - „chi va piano va sano“

In 1983 the German Federal Constitutional Court (Volkszählungsurteil) stated:

- the ‘**right to data privacy**’ as an inherent part of personal rights
  - ‘right to data privacy’ (*Recht auf informationelle Selbstbestimmung*) means the individual authority to decide if and under which conditions to expose one’s personal data
  - the ‘right to data privacy’ can be only constraint in case of a predominant, well justified public interest

1 Microdata Access in Germany - „chi va piano va sano“

In 1983 the German Federal Constitutional Court (Volkszählungsurteil) also stated:

- an **academic privilege** regarding the access to official microdata based on following assumptions:
  - **professional interest**: academic research when analysing microdata – as a rule - is not interested in the respondent as an individual person but in the respondent as a statistical unit.
  - in academic research personalized data necessary to assign anonymized microdata to individuals are rather not available.

=> This verdict of the Constitutional Court established the theoretical and legal base of access to microdata by the academic research community in Germany

1 Microdata Access in Germany - „chi va piano va sano“

- Nevertheless due to the pronounced public concern over data privacy in the subsequent years the statistical offices adopted a very reluctant attitude regarding the release of microdata for research purposes based on following reasoning :
  - the trust of respondents is an important precondition regarding the cooperation in data collections
  - principle of data confidentiality = key element of that trust
  - If the respondents believe that the confidentiality of their data is not guaranteed, they are less likely to cooperate or provide accurate information
  - any incidence, in particular if it receives strong media attention, could have a significant impact on respondents cooperation and therefore on the quality of official statistics

1 Microdata Access in Germany - „chi va piano va sano“

- This line of argument misses the fact that researchers in a similar way as statistical agencies rely on the trust of respondents
  - gathering own data as well as sharing microdata (e.g. in terms of data archives) has a long tradition in social research
  - some of the data gathered by research are much more sensitive than official statistics
  - any breach of confidentiality whether it affects social research data, official microdata (or other types of personal data) might effect the willingness of respondents to cooperate in surveys
  - ⇒ Might harm the foundation of social research

=>Thus not only statistical agencies but also the research community have an all-embracing interest in the issue of data confidentiality

1 Microdata Access in Germany - „chi va piano va sano“

- All the same the **implementation of the academic privilege** (although explicitly included in the Federal Statistical Law, 1987) into practice was
  - strongly characterized by the fact that statistical agencies
    - could – but are not obliged to - make their data available as microdata for research purposes
    - did not get any additional ressources in terms of manpower or money for the creating of microdata sets
  - a rather slow and tedious process (attended by a number of research projects and advisory boards concerned with ‘identification risks’)
  - initially limited to a few types of microdata (mainly household or person surveys)

1 Microdata Access in Germany - „chi va piano va sano“

- However according the lines “chi va piano va sano” there was a considerable improvement over the last decade due to the set-up of a so-called **German Data Infrastructure** (founded and co-financed by Federal Ministry of Education and Research, BMBF) which is composed of:

  - **German Council for Social and Economic Data**
  - **Research Data Centres (organized by data producers)**
  - **Data Service Centres (organized by research service centres)**

1 Microdata Access in Germany - „chi va piano va sano“

- **German Council for Social and Economic Data**
  - comprised of 6 representatives of academic empirical research and 6 representatives of the main producers of official data and social research institutes
  - Mainly responsible for the assessment and further improvement of the data infrastructure in Germany
- **Research Data Centres (organized by data producers)**
  - RDC of the Federal Statistical Office
  - RDC of the Statistical Offices of the Länder (German states)
  - RDC of the Federal German Pension Insurance
  - RDC of the Federal Employment Agency in the Institute for Employment Research
- **Data Service Centres (organized by research service centres)**
  - Institute for the Study of Labor (IZA)
  - GESIS Service Centre for Microdata: German Microdata Lab (GML)

1 Microdata Access in Germany - „chi va piano va sano“

- Most essential progress which has been achieved by the establishment of the German Data Infrastructure up to now:
  - **closer cooperation** between the data producers and the research community
  - the **perception of the identification risk has become much more realistic** => common knowledge that uniqueness of records in a data file is by no means a sufficient condition for disclosure
  - **self-commitment** of data producers to make all types of official statistics (e.g. population, business, health, social insurance, environment) as microdata available for research purposes
  - the extension and increasing **standardization of data access** in terms of
    (a) scientific use files (off site access)
    (b) on site access
    (c) remote access
    (d) public use files (off site access)

2. Modes of Microdata Access - Scientific Use Files

Scientific Use Files = off site access

- microdata anonymized according to the concept of **reasonable anonymity** (*faktische Anonymität*)

- **reasonable anonymity** (based on the Federal Statistical Law, 1987) stipulates that the

  > microdata must be anonymized in such a way that any identification of individuals is only possible by inordinate expenditures of time, costs, and personnel

- The use of Scientific Use Files is **restricted to the academic research community** and a contract is signed before the data are provided to the researcher

2 Modes of Microdata Access - Scientific Use Files

- measures of anonymity are applied in such a way that while taking account of data confidentiality the statistical information in the data remains high, take e.g. the Mikrozensus (1% of the total population):
  - coarsening of geographic information (e.g. **common classification of territorial units for statistics:** NUTS-1: 16 German States)
  - each category of each variable in the data should represent a minimum of cases in the population (e.g. 5,000) => otherwise the category is recoded
  - drawing of a sub-sample (70 % of the original data)
- in addition contractual commitments

2 Modes of Microdata Access - Scientific Use Files

(1) commitment according to the principles of the German criminal code (Strafgesetzbuch; StGB) with regard to the

- breach of personal privacy (Privatgeheimnisse; §203 StGB)
- utilisation of external privacy information (fremde Geheimnisse; §204 StGB)

=> Precondition that any violation of data confidentiality by the obliged individual can be prosecuted (money fine or prison term)

2 Modes of Microdata Access - Scientific Use Files

(2) Additional contractual commitments, e.g.:

- the microdata will only be used for the specified research projects (upgrades possible)
- no attempt will be made to identify particular persons or organizations
- the microdata will be only provided to persons (upgrades possible) who
  - are working in the research projects
  - are mentioned by name to the data producer
  - incurred the aforementioned commitment
- point of time when the data have to be deleted (prolongation possible) as well as a written notification of the deletion to the data producer

2 Modes of Microdata Access - Scientific Use Files

- Violation of the commitments could be sanctioned by
  - money fine, and
  - deletion of the data, and
  - exclusion from further data access, and
  - claim for indemnification

- Moreover the standard contract includes recommendations concerning data protection arrangements, e.g.
  - allocation of data to entitled users only
  - protection of the media
  - no access to the data from external computers
  - data access only by dedicated terminals

$\Rightarrow$ in essence scientific use files should be only used in the office in a safe environment but e.g. not at the workplace at home or a laptop

2 Modes of Microdata Access - Scientific Use Files

- Except for the constraint of using the data only at the “office”, Scientific Use Files offer the flexibility needed when doing empirical research

  – work with data is not limited to specific times. Literature, statistical programmes, etc. are available. Problems and findings can be immediately discussed with colleagues …

= > Scientific Use Files are the first choice of microdata access

2 Modes of Microdata Access - Scientific Use Files

Scientific Use Files = standardized data files

- some drawbacks:
  - some specific research interests might not be covered because of anonymity measures (in particular all kind of small area studies)
  - not all types of data (e.g. business data) can be anonymized by using measures which do not harm the information included in the data at large
  - the preparation of a scientific use file requires considerable efforts which are not worthwhile if there is only an exceedingly small demand for the data

=> In these instances ‘on site access’ or ‘remote data processing’ might be the appropriate mode of access

2 Modes of Microdata Access: On-site access

On-site access to official microdata is possible at the Research Data Centres

- access only for the academic research community
- Much less convenient for researchers in terms of time, flexibility and costs (e.g. travel expenses) than scientific use files but e.g. appropriate for:
  - Microdata which can not be anonymized according to the principle of reasonable anonymity without seriously damaging the usefulness of the statistical information
  - Microdata for which no scientific use files are provided because they are of interest only to very few researchers
  - Researchers who need more detailed data as available in the scientific use file (e.g. very fine geographical information or full sample size)
  - Researchers who like to use official microdata but do (yet) not possess the skills, knowledge or facilities required to analyze the data

2 Modes of Microdata Access: Remote Access

The ideal:

- researchers can use their local computer facilities at any time to do statistical analysis with official microdata which are physically stored at the research data centres and without actually “seeing” the data
- Statistical output is checked (automated or manually) for any breaches of confidentiality

Indeed remote access is still on a rather elementary stage in Germany:

- The analysis script (SPSS, Stata, SAS) is send by email to the RDC
- In the RDC the script is applied to the detailed microdata
- The statistical output is manually checked for confidentiality risks
- The output is send back to the researcher

2 Modes of Microdata Access: Remote Access

Advantages of the current model:

- no travel expenses
  - precondition: the researcher knows the data very well and has already a very clear idea of the kind of statistical analysis needed to do the job
- preparation for a deepening analysis using on-site access

2 Modes of Microdata Access: Remote Access

Disadvantages of the current model:

- rather time-consuming for the researcher as well as for the RDC
  - From sending in the script to receiving the statistical output some time might lapse away (e.g. depending on the manpower resources available in the RDC)
  - any inconsistencies in the analysis most often can be seen only by a closer inspection of the findings => revision of the script, sending in ….sending back …
- rather low flexibility, e.g.
  - during the research process upcoming empirical questions could not or only with a time lag sort out
  - explorative nature of social research is rather pushed back

2 Modes of Microdata Access: Public Use Files/Campus Files

- Public Use Files/Campus Files are microdata that are disseminated for general public use via CD-ROM or internet

- Data are anonymized in such a way that identification of individuals is impossible, e.g.

  – suppression of variables
  – high level of collapsing categories
  – small sub-samples

  => remaining statistical information is rather low

- However Public Use Files/Campus Files are valuable for teaching purposes simply because they can be easily used by the students to gain first experience in dealing with official microdata, to “replicate” published findings, doing first steps in research … etc.

3 Concluding notes

- The access to microdata is an ongoing project, they are still frictional losses regarding e.g.
  - the assessment of identification risks for certain types of microdata
  - the time it sometimes takes to prepare scientific use files (especially in case of new data types: e.g. paneldata or small area files, but also the regular upgrading of existing scientific use files)
  - according to which priority microdata are provided as scientific use files (e.g. what is more important? Providing ‘old’ data or the latest data?)
  - why do the research community also need high-quality campus-files?
  - (…)

3 Concluding notes

II The right to data privacy and data confidentiality are important issues in Germany and might even become more important

- the more data are collected, matched with other data sources and stored by governmental agencies and in the private sector
- the less the public trusts that information supplied for statistical purposes does not have any direct consequences for the supplier of the data
- and last but not least: because of the latest come to know breaches of data confidentiality in the private sector (selling personal data including bank account information which were then used by third parties in terms of criminal acts)

=> but these developments should be frankly discussed and in the interest of the respondents, statistical agencies and researchers joint solutions should be developed

3 Concluding notes

- from the researcher's point of view the most favourable solution is the provision of scientific use files

- only if after a thoroughly audit - and in due consideration that the main interest of social research is definitively not in identifying individuals - it becomes evident that there is no way to protect data confidentiality without losing too much statistical information other modes of access should be preferred

- If in future remote data access enables data handling with the same flexibility as using the data locally it might be an alternative

  – however one should not forget that sometimes the researcher has to see the data (or to be precise: single records) only if to check for data inconsistencies

THANK YOU FOR YOUR ATTENTION !