Harmonisation of statistical confidentiality in the Federal Republic of Germany

Maurice Brandt, Anja Crößmann and Christopher Gürke [*]

[*] Federal Statistical Office, Research Data Centre, 65180 Wiesbaden, Germany
maurice.brandt@destatis.de, anja.croessmann@destatis.de, christopher.guerke@destatis.de

Abstract: The Federal Statistical Office and the Statistical Offices of the Federal States in Germany provide a wide range of statistical information, information services and access to microdata for scientific research. To guarantee a consistent standard of official data, in respect of statistical confidentiality, harmonisation in different aspects is necessary. In this paper, the German way in harmonisation of statistical confidentiality between the Federal Statistical Office and the Statistical Offices of the Federal States is presented. Beside general legal regulations, the harmonisation of access to German microdata is focused. Therefore guidelines for anonymisation of microdata files as well as for data laboratories, confidentiality of output, output checking and tabular confidentiality will be exemplarily discussed.

1 Federal statistical system in Germany

Similar to the European Statistical System (ESS), due to the federalism, Germany affords several national public data producers. According to the federal structure of the state and the administration in the Federal Republic of Germany federation-wide surveys of official statistics, so called "federal statistical surveys", are organised in cooperation between the Federal Statistical Office (FSO) and the statistical offices of the 16 federal states.

For this reason the majority of official statistics in Germany are elevated in a decentralised manner, in each federal state. This requires common guidelines in which statistical methods are defined and harmonised to guarantee a consistent standard of data collecting, processing, publication and dissemination to third parties. A division of labour between the Federal Statistical Office and the Statistical Offices of the Federal States is necessary.

The responsibilities of the various statistical operations are regulated by the German Law on Statistics for Federal Purposes (FSL)[1]. Duties of the Federal Statistical Office (FSL, Art. 3) are primarily the preparation and further development of the federal statistics programme. In closely collaboration, the statistical divisions of the FSO and the federal states define methodological and technical needs of official

[1] Federal Statistics Law – FSL of 22 January 1987

surveys. In examination of the Federal Statistical Office data collection and data processing of decentralised statistics are predominantly carried out by the Statistical Offices of the Federal States. Thus the Statistical Offices of the Federal States realise the legal required statistics locally, which will be combined by the Federal Statistical Office.

The main function of the Federal Statistical Office in this collaboration is the mutual coordination of federal statistical surveys. These have to be without overlaps, complied with the defined standard methods and fulfilled within the time schedule. The Federal Statistical Office is in part or entirely entrusted with the statistical processing of the data (FSL, Art. 8). After collecting it from the states, the Federal Statistical Office compiles the regional data to produce one common federal survey which can be published and presented on a federal level for general purposes. Furthermore – in frame of European community statistics – the Federal Statistical Office transmits the common German microdata to Eurostat.

2 Harmonisation of statistical confidentiality

Principal five of European Statistics Code of Practice[2] describes that "the privacy of data providers (households, enterprises, administrations and other respondents), the confidentiality of the information they provide and its use only for statistical purposes must be absolutely guaranteed". Generally, data confidentiality is one of the most important duties of official data producers. Therefore the statistical offices of the national states have to take organisational and procedural precautions to counteract the risk of the right to privacy being violated. The main aims of statistical confidentiality are:

- The protection of every respondent against disclosure of its personal circumstances and its material situation
- Preservation of mutual trust between the respondent and the statistical offices
- Volition of response and guarantee of response reliability

These requirements of statistical confidentiality need an establishment of standardisation and harmonisation between the statistical offices. The following describes some selected aspects in harmonisation of statistical confidentiality between the Federal Statistical Office and the Statistical Offices of the Federal States in Germany.

[2] European Statistics Code of Practice for the national and community statistical authorities, 2005.

2.1 Legal aspects of harmonisation

To assure the confidentiality requirements Germany has defined a set of legal standards which have to be fulfilled by both, the Statistical Offices of the Federation and the Federal States. These standards are fixed in Law on Statistics for Federal Purposes. Therein harmonisation of statistical confidentiality is regulated in various fields. From collecting to publishing, every step of producing official data takes place in respect of standardised confidentiality. For example before publishing regional or national results, summary tables are checked for confidentiality by the statistical offices to protect the release of sensible data.

First of all, official employees which are entrusted with the operation of official statistics are specially sworn in for public services (FSL, Art. 16, Para. 1).

In principle, after data collection, the plausibility of conclusiveness and completeness of a survey and its auxiliary characteristics, like personal number or address, will be checked. The auxiliary characteristics will be separated from the survey characteristics by the Statistical Offices of the Federal States and, in case of periodical surveys, stored separately (FSL, Art. 12). The transmitting of individual microdata between the Statistical Offices of the Federation and the Federal States takes place under strongly controlled encrypted data lines.

Individual data is strictly confidential, unless otherwise stipulated by a special legal provision (FSL, Art. 16, Para. 1). Such as individual data for the transmission or publication of which the respondents have given their written approval, individual data from generally accessible sources, individual data which have been summarised with such of other respondents as well as individual data if it is not possible to allocate respondents or concerned persons. The last case gives the opportunity to create absolutely anonymised microdata files of common surveys, so called Public Use Files (PUF). These can be used for national or aboard researches by every interested person or, in form of especially for academic teaching created CAMPUS Files, at institutions of higher education for statistical methods training.

Preconditioned that a disclosure of microdata is possible only with an excessive amount of time, expenses and manpower, the Statistical Offices of the Federation and the Federal States can transmit microdata to institutions of higher education or other institutions entrusted with tasks of independent scientific research (FSL, Art. 16, Para. 6). For this purpose the statistical offices create de facto anonymised microdata files in form of standardised Scientific Use Files (SUF) for off-site use to users from the scientific community. These Scientific Use Files are similar to the so called “Microdata Under Contract” (MUC). The recipients of SUF must prior to the transmission be committed to confidentiality (FSL, Art. 16, Para. 7). The data can be used only for a predefined scientific project and as soon as the project has been completed, data has to be deleted. At agencies to which SUF are transmitted, it must be warranted by means of organisational and technical measures that only especially authorised persons are recipients for using the individual data (FSL, Art. 16, Para. 8).

The statistical offices have to keep records on contents, recipient agency, forwarding date and purpose of transmission and preserve them for a minimum of five years (FSL, Art. 16, Para. 9).

To protect personal information of respondents it is prohibited to return individual data to the administrative authorities. Furthermore it is not allowed to match individual data from federal statistics or to combine such with other information for establishing a reference to persons, enterprises, establishments or local units for other than the legal statistical purposes or of a legal provision ordering a federal statistics (FSL, Art. 21).

It is important to assure the data security during the production procedure in the statistical offices. The probably most difficult part is to preserve the confidentiality of official statistics where the data are accessed by scientific research. On the one hand this needs safe (anonymised) data or a safe environment to get data access and on the other hand the results have to be kept safe until they are checked for confidentiality that they can be released to the public.

2.2 Access to German microdata

For the access to German microdata there are regulations and guidelines which are implemented to assure the confidentiality of the data. There are various ways to keep the access to microdata for the scientific community as comfortable as possible under strict observation of the secureness of the data.

Access to microdata of official statistics in Germany is possible via Scientific Use Files (SUF), Public Use Files (PUF), via data laboratory[3] and remote execution through the Research Data Centres (RDC) of the statistical offices. Due to the federalism the Federal Statistical Office and the Statistical Offices of the Federal States build up two RDCs, the RDC of the Federal Statistical Office and the RDC head office of the Statistical Offices of the Federal States, which is organised in fourteen field offices. Besides there are two more RDCs which are publicly funded and named here for completion: The Research Data Centre of the Federal Employment Agency at the Institute for Employment Research (FDZ-BA) and the Research Data Centre of the German Pension Insurance (FDZ-RV). These RDCs provide special microdata located in the field of employment and pension and are governed by diverse laws than the Statistical Offices of the Federal States and Federation[4].

[3] A data laboratory is a secured room in a statistical office especially designed for researchers granted with right of access to microdata. Such a room is equipped with special features preventing the transmission of any kind of information to the outside world.

[4] For further information see Bender et. al. 2009

Access to German microdata is coordinated between the Federal Statistical Office and the Statistical Offices of the Federal States for each request.

The researcher can fill out a common request form to provide information on:

- the institution submitting the request/the individual data users
- the microdata requested
- the form of data access
- the research project

This request form can be send to both Research Data Centres of the German federation or the federal states. The form will be circulated and coordinated between all offices.

After the RDCs agreed upon the request form the researchers are able to access the microdata in a way they have chosen in the request form, namely by SUF, data laboratory or remote execution.

The researchers can use the microdata for example in any desired data laboratory of the federal states or the Federal Statistical Office. This is possible because the federal states and the federation agreed on guidelines for data laboratories which regulate the security requirements and organisational issues. The rules for the data laboratories are described in Chapter 2.4.

2.3 Guideline for anonymisation of microdata files

Due to the federal organisation German microdata is divided in centralised and de-centralised statistics. However, as a result of the harmonisation of the RDCs researchers have the opportunity to analyse both types of de facto anonymised microdata at the Federal Statistical Office in Wiesbaden, Bonn or Berlin or at one of the fourteen Statistical Offices of the Federal States.

De facto anonymity is achieved not only by an anonymisation of the data but in combination with a controlled data access. This is why these data may contain much more detailed information than the Scientific Use Files submitted in the form of data files.

Microdata are called de facto anonymised if deanonymisation cannot be ruled out completely but the data can be allocated to the respective statistical unit only with an excessive amount of time, expenses and manpower (Art. 16 Para. 6 Federal Statistics Law). Pursuant to that law, de facto anonymised data may be made available only to scientific institutions and only for the purpose of scientific projects.

De facto anonymisation mainly aims at reducing the possibilities of allocating the values of a variable to the respective statistical units by careful information reduction and information modification while preserving the informational value in statistical

terms. The cost and benefit of deanonymisation have to be analysed for each individual survey. Therefore the RDC has to anonymise microdata specific to every research project. Referring to this the local kind-of-activity unit develops an anonymisation concept in close collaboration with the researcher. The concept is specified in such a way that researchers can conduct their full analyses while the used microdata meet confidentiality. Upon completion the anonymisation concept will be approved by the RDCs of the Federal Statistical Office and the Statistical Offices of the Federal States. The review focuses on the compliance with the definition of de facto anonymisation and local specific disclosure risks.

Afterwards the anonymised microdata will be provided at the data laboratory, where the researcher grants access to a special account.

The daily business in dealing with research applications is the coordination of the data request and the anonymisation concepts for Scientific Use Files for off-site and on-site usage. To accelerate and simplify the process of coordination guidelines for the management of requirements for de facto anonymous individual data for the off-site usage and for on-site usage in a data laboratory are developed.

The guideline contains the following steps:

1. Receipt of the request for microdata use in the Research Data Centre

   The RDC is for processing the research application and involving the concerned organisational units and maintains contact with the user.

2. Legal analysis of access according to § 16 paragraph 6 FSL

   For the application of new institutions the RDC involves the legal unit.

   In collaboration with lawyers of the Statistical Offices of the Federal States the legal unit verifies, whether this institution performs independent scientific research within the meaning of § 16, paragraph 6 FSL. If the legal prerequisites are not fulfilled the RDC informs the data users, and indicates to alternative forms of access (remote execution on a absorbed cost bases, special evaluation or Public Use File)

3. Creation of the anonymisation concept

   The RDC creates the anonymisation concept based on the specific anonymisation basic guideline to anonymous individual statistical data in consultation with the relevant statistical division. The concept includes the performance of anonymity measures and relevant tests. If subsamples

should be used for anonymity, the change of weighting factors has to be coordinated with the division for mathematical methods.

4. Participation of the RDC of the Statistical Offices of the Federal States

   For decentralised statistics, the RDC of the Statistical Offices of the Federal States is to integrate for coordination of the anonymisation concept with all regional locations. Central statistics are simply told to note.

5. Contract

   The legal department will sign a contract with the data users. Part of the contract is the anonymisation concept. A copy of the completed contract is forwarded to the RDC.

6. Data delivery and invoicing

   The RDC offers the de facto anonymised individual data for the off-site use on a Disk (CD-ROM, DVD) or for the on-site use in a data laboratory. Invoices are issued by the RDC.

The improvement of harmonisation is shown exemplary for personal and household as well as for business statistics.

2.3.1 Personal and Household Statistics

The statistical offices decided to produce a standardised on-site file for the microcensus, which is the most often used official data in the field of personal and household statistics in Germany. This is due to the fact that the anonymisation concept for on-site use is developed in close cooperation with the researchers and can therefore turn out very time consuming and elaborate. In addition to that the coordination itself delays the microdata access for the researchers. Therefore the benefit of a standardised on-site file is, that the anonymisation concept needs to be coordinated only once and the microdata can be used after that by all researchers via a data laboratory or remote execution without any additional adaptation. The on-site file contains the full coverage of observations and all important characteristics for most research questions. All identifiers are removed and the detail of regional level is limited. Users have to document and comment the analysis program and the output. For the structure of analysis programs there are also unified guidelines how to write a code. This makes it easier for the employees to reproduce the results.

In terms of off-site use via SUF the anonymisation concept of the relevant statistics has to be coordinated between the Statistical Office of the Federation and the Federal States. For a standardised SUF this needs to be done only once per statistic.

2.3.2 Business Statistics

To produce Scientific Use Files i.e. for business statistics, there were two projects in the past where the Statistical Offices of the Federal States and the Federal Statistical Office took part. Within these projects the possible different views of the partners are already reflected. Goal of these anonymisation projects was to set up a guideline how to anonymise business statistics in an efficient way that the confidentiality and the analysis potential of the data can be assured. One first project was conducted for cross section business statistics and a following up project transferred the collected knowledge to longitudinal enterprise microdata. Each project aimed to consider the specialities of the federal states and the federation to create a common product.

After the anonymisation guideline is agreed in the Federal Statistical Office, it has to be sent to each single statistical office of the federal states to give the opportunity for remarks, feedback and changes. One substantial advantage of the incorporation of the federal states in decision making is the situation of the functional responsible location. This means that almost each federal state has a functional responsibility for a certain statistic. The regional offices have more detailed knowledge about their region. They know their companies and can rather tell which company could be under risk even in a Scientific Use File. Furthermore federal states are not only involved in process of decision making but also get the function of double checking to improve the quality of statistical products in aspects of confidentiality.

After this process a standardised file for scientific use can be released to the research community. The developed anonymisation guideline is then used for following up statistics of different years and waves. This makes it much faster to create Scientific Use Files of other business statistics. The knowledge shall be communicated within the Federal Statistical Office and also to the Statistical Offices of the Federal States.

2.4 Guideline for Data Laborаties

The Federal Statistical Office and the Statistical Offices of the Federal States notice considerable requests by researchers for access to confidential data sets. Therefore the statistical offices established harmonised data laboratories, which are special PC workplaces in the RDCs, where domestic and foreign researchers can analyse de facto anonymised microdata.

The researcher has no direct access to the internal production network of the statistical offices. This restriction prevents the possibility of researchers to access other sensitive information. It also allows RDC staff to supervise every step of analysis at all times.

The data laboratory itself consists of a secure hermetic working and data storage environment in which the confidentiality of the data for research can be ensured. It also avoids feeding the de facto anonymised micro data with further information.

A separate PC workplace with an internet connection is available for e-mail communication and World Wide Web searches.

From the administrative point of view, the data laboratories comply with defined standards. The following aspects are taken into account:

- The data laboratories are located in the premises of the Federal Statistical Office or in the premises of the Statistical Offices of the Federal States,
- legal measures have to be taken when allowing access,
- only authorised users should be able to make use of this facility,
- the use of laptops, mass storage and picture recording devices (e.g. digital cameras, camera phones) is prohibited in the data laboratory,
- the RDC staff members are permitted at all times to examine the activities of the researchers including their working materials.

The computers for analyses are subject to restrictions in order to prevent disclosure of individual persons or entities and to meet confidentiality. Therefore it is not possible to

- print documents,
- copy data to diskettes, USB sticks, CD-ROM's, DVDs or Zip drives,
- copy data to the local hard disk,
- connect recording devices to the serial, parallel and USB ports,
- connect a laptop to the network,
- use E-mail,
- make Internet connections,
- install hardware (the PC is locked) or to take out,
- nor boot the PC from floppy, CD-Rom, DVD-Rom or any other media.

Before releasing intermediate and final outputs the RDC staffs have to check every output concerning confidentiality (see above).

2.5 Confidentiality of output

Due to the fact that researchers can access official statistics from RDCs in different federal states and different locations of the Federal Statistical Office it is necessary to harmonise the rules of output checking. This is important to avoid the case of unequal treating of the same analysis and the same output by different researchers in aspects of confidentiality because they access data from different points. For that reason there are statistic overlapping confidentiality rules. For example the p%-rule which parameters are harmonised between federation and federal states. Furthermore the minimum frequency of at least three cases in a table is agreed as well. There are also recommendations like double-checking the output in the RDC and also in the related statistical division.
For the standard publication of the statistical divisions it is inevitable to coordinate the confidentiality of tables between the Federal Statistical Office and the Statistical Offices of the Federal States. To manage this complex intention the committee for organisation and implementation decided to harmonise the confidentiality regulations for the turnover tax statistics. This means the standard tables of the federal states and the federation shall be coordinated in a central cooperation model on federal states level. Synchronising the primary and secondary cell suppression for tabular data is a complex problem. One the one hand the federal states have to produce and publish their own tables and have to take care of confidentiality, on the other hand the Federal Statistical Office has to produce and publish tables aggregated on a federal level. If some cells of some federal states are already suppressed it is barely possible to publish results on a federal level, because the values can be recalculated. If the Federal Statistical Offices perform secondary cell suppression by suppressing the values of a certain federal state, the federal state again can't publish the own tabular data. This can be solved by using automatic tabular data protection in τ-Argus, whereas the federal states and the federation using the same unified concept. The adoption of the secondary cell suppression on federation and federal states level is mandatory for the respective publication of the Federal Statistical Office and the Statistical Offices of the Federal States.

It is planned to develop and transfer the concept also to other statistics like accommodation statistics.

3 Summary and Outlook

Considering the examples in harmonisation of statistical confidentiality between the Federal Statistical Office and the Statistical Offices of the Federal States presented above, there are already progress and results in this field. But there is also plenty of

space for optimisation. In a decentralised statistical system it has to be taken into account that not only the conduction of a survey needs to be harmonised. Also the confidentiality rules for tabular data in the different statistical divisions and the output checking rules in the RDCs of the several statistical offices in the federal states and the locations of the RDCs of the Federal Statistical Office need to be harmonised to treat each research output equal and to keep track of various results produced by researchers. Because of the different laws by which publicly funded RDCs in Germany are governed aspiring an unified approach is even more important.

One improvement could be a shared user management database where the employees of each RDC have always an insight which researcher is using which data for what reason.

Harmonisation in a decentralised system is inevitable to create comparable and high quality results. But this process needs a lot of resources to coordinate and unify procedures. It is important to optimise the process especially on the interfaces between the related bodies. The decentralised statistical system in Germany can be a model for the European system because represents a kind of microcosm for the European situation even if the legal framework is different. For both systems better methods and instruments for harmonisation are necessary.

References

Bender, S., Himmelreicher, R., Zühlke, S. & Zwick, M.(2009): *Improvement of Access to Data Set from the Official Statistics.* Working Paper Series of the Council for Social and Economic Data (RatSWD), No. 118.

CENEX SDC (2007): *Handbook on Statistical Disclosure Control, Version 1.01.* http://neon.vb.cbs.nl/casc/SDC_Handbook.pdf .

Federal Republic of Germany (1987): *Law on Statistics for Federal Purposes (Federal Statistics Law - FSL) of 22 January 1987*.

Statistical Programme Committee (2005): *European Statistics Code of Practice, 24. February 2005*.

Zühlke, S., Zwick, M. Scharnhorst, S. & Wende, T. (2005): *The research data centres of the Federal Statistical Office and the statistical offices of the Länder*. Schmollers Jahrbuch 4, p. 567 ff.