German Council for Social and Economic Data (RatSWD)

RatSWD

Working Paper Series

Working Paper No. 82

Data protection and statistics – a dynamic and tension-filled relationship

Peter Schaar

March 2009

Federal Ministry of Education and Research

Working Paper Series of the Council for Social and Economic Data (RatSWD)

The *RatSWD Working Papers* series was launched at the end of 2007. Since 2009, the series has been publishing exclusively conceptual and historical works dealing with the organization of the German statistical infrastructure and research infrastructure in the social, behavioral, and economic sciences. Papers that have appeared in the series deal primarily with the organization of Germany's official statistical system, government agency research, and academic research infrastructure, as well as directly with the work of the RatSWD. Papers addressing the aforementioned topics in other countries as well as supranational aspects are particularly welcome.

RatSWD Working Papers are non-exclusive, which means that there is nothing to prevent you from publishing your work in another venue as well: all papers can and should also appear in professionally, institutionally, and locally specialized journals. The *RatSWD Working Papers* are not available in bookstores but can be ordered online through the RatSWD.

In order to make the series more accessible to readers not fluent in German, the English section of the *RatSWD Working Papers* website presents only those papers published in English, while the the German section lists the complete contents of all issues in the series in chronological order.

Starting in 2009, some of the empirical research papers that originally appeared in the *RatSWD Working Papers* series will be published in the series *RatSWD Research Notes*.

The views expressed in the *RatSWD Working Papers* are exclusively the opinions of their authors and not those of the RatSWD.

Data protection and statistics – a dynamic and tension-filled relationship

Peter Schaar

Der Bundesbeauftragte für den Datenschutz und die Informationsfreiheit
(peter.schaar[at]bfdi.bund.de)

Abstract

New statistical methods have been developed for the longer-term storage of microdata. These methods must comply, however, with the fundamental right to informational self-determination and the legal regulations imposed by the Federal Constitutional Court. Thus it is crucial to develop effective and coherent methods for protecting personal data collected for statistical purposes.

Recent decisions by the Federal Constitutional Court are likely to result in the outlawing of comprehensive, permanent statistical compilations comprised of microdata from a wide range of sources and updated regularly. However, aside from such comprehensive methods, there are certainly other ways of using microdata that cannot be dismissed from the outset as violating constitutional legal norms.

Internet access to statistical microdata is likely to take on increased importance for scientific research in the near future. Yet this would radically change the entire landscape of data protection: the vast amount of additional information now available on the Internet makes it almost impossible to judge whether individuals can be rendered identifiable. In view of this almost unlimited information, individual data can only be offered over the Internet if the absolute anonymity of the data can be guaranteed.

Keywords: Right to informational self-determination, census ruling of December 15, 1983, longer-term storage of microdata, primary statistics, secondary statistics, statistical confidentiality, absolute anonymisation, de facto anonymisation, additional information, pseudonymisation, personal data profiles.

Introduction

Statistics traditionally deals with the collection and evaluation of data on the personal or material situations of a large number of individuals or organizations:

> Statistics means ... activity aimed both at measuring mass phenomena, combining the data into groups and publishing them.[1]

A more recent textbook contains the following definition,

> Statistics is a combination of mathematical methods used to assess mass phenomena. The data collected serves to describe the environment numerically and/or in the event of uncertainty to use this data as a decision-making aid.[2]

The purpose of data protection is, according to the Federal Data Protection Act, "to protect the individual against impairment of his right to privacy through the handling of his personal data."[3] "Personal data means any information concerning the personal or material circumstances of an identified or identifiable individual (the data subject)."[4]

If and when statistics are used merely to evaluate information relating to institutions (government agencies, companies) or natural phenomena (e.g., weather), data protection issues are irrelevant. However, the situation is much more complex with data on personal circumstances, as the heated debate on the 1983/87 census showed. This "individual data" is linked to the data subject at least during the data collection phase, and may also involve personal data. Only in the course of further data processing and evaluation is the personal reference eliminated partially or completely. In the final analysis, data may only be published if it can be ruled out in all likelihood that conclusions can be drawn about individuals. Personal data are therefore rendered (de facto) anonymous within the framework of traditional statistics.

Data protection requirements are changing as new statistical methods focused on the longer-term storage of individual data (microdata) become available (especially in the context of longitudinal studies). This means the described method of data collection - in which the data are rendered anonymous, preventing access to personal data and publishing only aggregated results - is no longer adequate under all circumstances, and thus no longer practicable.

---

1 Meyers Konversationslexikon (Meyer's Conversational Encyclopedia), 1907, volume 18, under the term "Statistik."
2 Bücker, Statistik für Wirtschaftswissenschaftler (Statistics for Economists), 3rd Edition, 11.
3 BDSG (Bundesdatenschutzgesetz) 1, 1.
4 BDSG 3, 1.

The right to informational self-determination

With its census ruling of December 15, 1983, the Federal Constitutional Court (BVerfG) formulated the "fundamental right to informational self-determination":

> Under the terms of modern data processing, the protection of the individual against the unlimited collection, storage, use and passing on of his/her personal data comes under the general right to free development of one's personality set forth in Article 2 para. 1 of the Basic Law in conjunction with Article 1 para. 1 of the Basic Law [inviolability of human dignity]. The basic right warrants [...] the capacity of the individual to determine in principle the disclosure and use of his/her personal data.[5]

The ruling defined the requirements that need to be met in order to ensure that personal data are processed in accordance with the German Constitution (Grundgesetz). It ruled that any risks of misuse must be taken into account when data are processed. Even data that may seem irrelevant in an isolated context has the capacity to become relevant in a different context (by data fusion and data matching). The Federal Constitutional Court hence ruled that "considering the fact that individual data can be stored without any technical restraint with the help of automatic data processing ... there is no longer any such thing as irrelevant data."[6] According to the Constitution, the collection and processing of data must be justified by reasons of compelling public interest; the prerequisites and scope of data processing must be regulated comprehensibly for citizens, and the principle of proportionality must apply.

Last but by no means least, the further processing of data must, in principle, be limited to the purpose for which it was originally collected, which is particularly relevant for the collection of data for statistical purposes. Contrary to the collection of personal data for a specific administrative task, the need to collect data for statistical purposes can only be described in abstract terms, as the results can and indeed should be used for multiple purposes. It is hence all the more important to ensure that statistical data processing is separated strictly from the processing of data for administrative tasks. The envisaged use of data to correct information in the identity register in the 1983 census was the main reason for the negative ruling by the Federal Constitutional Court.[7]

One of the major risks in terms of data protection is that personal data profiles can emerge that are capable of presenting a complete picture of an individual. Any such personality profiles are incompatible with the Basic Law. The Federal Constitutional Court already established this in 1969, in its microcensus ruling on personality profiles:

---

5 BVerfGE 65,1, 1.
6 BVerfGE 65, 1, 45.
7 BVerfGE 65, 1, 63.

> It would be incompatible with human dignity if the state claimed the right to register and catalogue people in their whole personality coercively, even if the data collected in a statistical survey was rendered anonymous, as this would treat people like objects that are accessible for data collection in every respect.[8]

Pursuant to this case ruling, it is now compulsory to protect personal data collected for statistical purposes in an effective and coherent manner. As such, it is important that the measures taken be oriented to the concrete threat situation and take the risks associated with rapid technological advancement into account.

Technological change

The main regulatory approaches to data protection originate from the time of mainframe computers, when electronic data processing took place at remote computing centers in accordance with rigid principles. Storage units the size of refrigerators, punch cards and continuous printing paper dominated the scene when the Federal Constitutional Court issued its census decision in 1983.

Three main changes have taken place that are important in this context: the dramatic increase in storage capacities, the flexible evaluation possibilities, even of huge databases ("data mining"), and the "liberation" of computers from computing centers, offering 24/7 access to databases via networks, particularly the Internet.

In view of these trends, certain protection concepts that date back to the 1980s and 1990s are no longer realistic in today's world. This applies, for instance, to the approach of physically sealing off the use of statistical data processing from processing for other purposes. Nowadays, statistical data can, of course, be processed on separate systems.

When data users are to be offered the benefits of computer technology, it is virtually impossible to do so without giving them electronic access to statistical data - e.g., through networks. It is difficult, if not impossible, to explain to users in the scientific and political communities why they are confined to rigid evaluations in the form of statistical aggregates and why they are denied access to microdata. After all, it is precisely microdata that offer a wide range of opportunities for obtaining new information. Nonetheless the risks associated with these convenient types of use must be considered carefully. If data are processed outside the "walls" of statistical offices, it is virtually impossible to control how it is used - for instance, whether it is being used in combination with other databases.

8 BVerfGE 27, 1, 6.

What is needed are concepts that develop new, flexible possibilities for utilization that meet the expectations of data users and that simultaneously safeguard effective, modern data protection.

Statistical confidentiality as a special data protection requirement

When developing data protection measures, it is crucial that the various legal, organisational, and technical measures are well coordinated. As such, the starting point is the obligation to maintain the statistical confidentiality, which aims first and foremost at ensuring - like all other regulations governing secrecy[9] - that only authorised "insiders" have access to the data and that the data are safe from use by unauthorised persons. The regulations governing the obligation to maintain statistical confidentiality represent special data protection regulations that override the general data protection legislation. They are intended not only to take the special sensitivity of the respective data into account, but also to build trust between the data subject and those who collect the data, who are obligated maintain the confidentiality of statistical data on individuals without the individual having to fear negative consequences. According to the Law on Statistics for Federal Purposes (BStatG):

> Individual data on personal circumstances or the material situation provided for federal statistics shall not be disclosed by the incumbents and the persons specially sworn in to public service who are entrusted with the operation of federal statistics, unless otherwise stipulated by a special legal provision.[10]

In principle, personal data may only be used for certain tasks defined by law. It is prohibited and a punishable offence to use data for any purposes other than those expressly permitted by law. The same applies to passing on data to third parties outside the respective area. However, the principle of purpose limitation does not apply to statistical results that do not contain any personal reference. Individual statistical data may also be used for scientific purposes under certain circumstances:

> For the purpose of scientific projects, the Federal Statistical Office and the Statistical Offices of the Länder may transfer microdata to institutions of higher education or other institutions entrusted with tasks of independent scientific research if an allocation of the individual data are possible only with an excessive amount of time, expenses and manpower, and if the recipients are elected officials, persons specially sworn in for public service, or persons obligated according to subsection 7.[11]

Contrary to this exemption for scientific purposes, the BStatG does not contain any explicit obligation to render individual data anonymous when this data are stored at statistical offices; however, this obligation arises implicitly from the jurisprudence of the Federal Constitutional

9 Examples: the duty to treat medical records confidentially, confessional secrets, secrecy of postsostal and telecommunications secrecy.
10 BStatG, section 16, subsection 1, sentence 1.
11 BStatG, section 16, subsection 6.

Court, particularly in its census ruling. The legislator took these terms of reference into account by issuing detailed regulations on the rendering anonymous of data in a large number of individual statistical regulations on the deletion of calculation input features. After all, section 10 of the BStatG defines certain minimum (albeit merely geographical) requirements specifying precisely what individual data can be stored for extended periods by saying exactly what is prohibited and by prohibiting the use of precise address details. Finally Section 21 of the BStatG stipulates that it is prohibited to match individual data from federal statistics or to combine any such individual data with other information:

> It is prohibited to match individual data from federal statistics or to combine such individual data with other information for establishing a reference to persons, enterprises, establishments or local units for other than the statistical purposes of this Law or of a legal provision ordering a federal statistics.

Statistical methods and data protection

Even though traditional statistics is based, by and large, on data that refers to individual survey units, it does not rely on having retroactive access to individual data - with the exception of checks carried out during the data collection and data processing phase (to ensure the data are complete, plausible and, to a limited extent, correct). Rather, statistics involves aggregates, namely numerical values that are analysed and matched, with comparison and evaluation of any changes in this data over time. In principle, any such aggregates do not contain personal data unless it is possible to trace the results back to persons indirectly. It may, for instance, be possible to draw indirect inferences about individuals from statistical results if the respective table cell relates to a small number of people. The same applies to special characteristic values - for instance, if all members of a group have the same characteristic values.

Statistics are not matched at the case level. Only when statistics are published must it be ensured that none of the above-mentioned scenarios occur and that relevant countermeasures have been taken (for instance, combining survey units, less distinctive characteristic values). As a rule, the loss of information associated with this rendering anonymous of data does not have any serious consequences and can certainly be tolerated (as long as different tabulations are not restricted, so to allow flexible tabulations).

The further development of statistical methods has led to heightened data protection requirements. The evaluation of statistical aggregates is supplemented by a more detailed analysis of patterns of individual statistical units at so-called “micro-level.” Group patterns are traced back to patterns in the lives of individuals, who may have been observed over an

extended period of time. To this end, the data on the individual needs to be collected and, if applicable, matched over time (into a longitudinal data set). The annual microcensus surveys that are carried out on the households under review at regular intervals over four consecutive years are based on this model.

Generally speaking, these new methods involve microdata that can be linked multifunctionally and can be evaluated over time (as, for example, in clinical studies). There are numerous ways of accessing so-called "microdata files" - for example, through personal references in case scenarios in which the data are linked by a general personal identifier that can be used in a wide range of surveys.[12] However, there is no doubt that any such personal identifier is incompatible with the above-mentioned requirements of the Federal Constitutional Court in Germany.[13] For this reason, the court is likely to declare comprehensive, permanent microdata statistics comprising regularly updated data from a wide range of sources to be unlawful. However, aside from these comprehensive methods, there are certainly ways of using microdata that cannot be dismissed from the outset as violating the constitutional requirements.

Measures aimed at safeguarding data protection

It goes without saying that the traditional method of rendering data anonymous and deleting individual statistics based on a type of "stage model" is not compatible with a method that links microdata. It may be possible to trace the individual data back to the data subject even at micro-level in the long term, which means the data does actually represent personal data in the majority of cases.

As such, one very interesting option would be to randomly link data collected within the framework of completely different statistical surveys in order to gain new information. In addition to the additional informative value such a method would yield, another argument in favor of it is the flexibility of the results it would generate.

In terms of data protection, any such method would involve major risks, given the apparent difficulty - if not impossibility - of rendering data anonymous in order to prevent inferences being drawn about identifiable statistical units. This risk could be mitigated by effectively ensuring that the data are protected against unauthorised access. However, whether this could achieve adequate protection is questionable, at least where particularly

12 Lenz, R. and Zwick, M.: "Integrierte Mikrodatenfiles—Methoden zur Verknüpfung von Einzeldaten" (integrated microdata files - methods of linking individual data). In: Statistische Bundesamt (Ed.): Statistik und Wissenschaft 10, 100.
13 Cf. n. 6, BVerfGE 27, 1, 6.

comprehensive or diverse microdata files containing personal features are concerned.

The type and size of the database is important when it comes to gauging the risk of abuse. Generally speaking, it can be said that the more comprehensive the database and the more sensitive the data, the greater the risk. This explains why censuses (which cover full populations) must be rated differently than surveys in which small random samples of an entire population are taken. Data abuse also occurs with random sampling, albeit only to the extent of the sampling units included. Thus, the "abuser" needs to know who is included in the survey.[14]

It is also important to distinguish between primary and secondary statistics. It is not possible to state simply which of the two methods is the more data protection-friendly. Occasionally, it is claimed that primary statistics, which collects data on the data subject is much more intrusive than secondary statistics, which does not collect any "new" data. This is only partly true. With secondary statistics, data are used and linked that were generally collected for another purpose altogether. This explains why most secondary files go hand in hand with data being used for a different purpose. Besides, the data subjects are "unaware" of the fact that their data are being used. Thus, they never gave informed consent. They are hence unable to check whether the data collection is lawful, and are unable to influence the process. In data protection terms, reference is made to deviation from the "Principle of Primary Collection" (ethical principle of informed consent). After all, more comprehensive secondary statistics - for instance the census envisaged for 2011 - presupposes that it is possible to linke data from different sources in which a particular type of infrastructure is needed. The question must therefore be raised how it can be prevented that infrastructures set up to collect statistics can also be used to link databases outside of statistics, with potentially far-reaching consequences for the data subject.

Rendering persons anonymous: absolute or de facto?

During the census debate of the 1980s, the most important question raised was: when does data lose its personal reference and when is it deemed anonymous or at least "de facto anonymous".[15] Only data that is completely anonymous contains no personal reference whatsoever, and therefore does not come under data protection regulations, whereas with de

14 It is easy to make this impossible by deleting some cases that were sampled from the file that is available for analysis.

15 Fischer-Hübner (1986): "Zur Anonymität und Reidentifizierbarkeit statistischer Daten" (Anonymity and reidentifiability of statistical data), Mitteilungen des Fachbereichs Informatik der Universität Hamburg 143; Brunnstein, K. (1987): "Über die Möglichkeit der Re-Identifikation von Personen aus Volkszählungsdaten" (The possibility of reidentifying persons from census data). In: Appel, R. (Ed.): Vorsicht Volkszählung! 2nd Edition, Cologne.

facto anonymous data it cannot be ruled out that the personal reference can be made/restored, if relevant "additional information" is available. Additional information describes the information needed to identify a person even if neither the person's name nor any other direct personal data (e.g., telephone number) can be linked with other information that uniquely identifies the person. With individual statistics, it is possible to restore the personal reference if certain characteristic values are disclosed and if these characteristic values can be associated with the data subject. As such it must be borne in mind that the boundaries between personal and anonymous data are fading in view of the rapid increase in data volumes, as ever more powerful computers are making it easier and easier to restore the personal reference retroactively.[16]

With fully anonymous data**,** there is no case scenario in which third parties can associate data with a person. Complete or genuine rendering anonymous hence means that personal data is altered in such a way as to ensure that the data can no longer be assigned to the person (even if there is additional information available). Only data that has been rendered fully anonymous contains no personal reference whatsoever.

According to the definition of the Federal Data Protection Act, "rendering anonymous" means the modification of personal data so that the information concerning personal or material circumstances can no longer be attributed to an identified or identifiable individual, or only with a disproportionate amount of time, expense or labour.[17] This statutory definition is confined to rendering data de facto anonymous. Pursuant to Section 3a of the BDSG, the data controller must implement data reduction and data economy measures. Pursuant to Section 3a (2) of the BDSG, use is to be made of the possibilities for aliasing and rendering persons anonymous, insofar as this is possible and the effort involved is reasonable in relation to the desired level of protection to be achieved. This also applies to statistics.

Pseudonyms as an expedient?

The use of pseudonyms is appropriate in cases where it is necessary to identify a person but where an assumed identity is sufficient, namely when the real personal details do not need to be known and when, on the other hand, (actual or absolute) rendering anonymous is not possible. This type of case scenario frequently arises in statistics if data are stored at micro level (for instance for longitudinal analyses).

16 Mattern and Langheinrich: Allgegenwärtigkeit des Computers – Datenschutz in einer Welt intelligenter Alltagsdinge (Omnipresence of computers – data protection in a world of intelligent everyday objects), 13.
17 BDSG, 3, 6.

Aliasing means replacing a person's name and other identifying characteristics (e.g., name, account number or personnel number) with a label in order to preclude identification of the data subject or to render such identification substantially difficult.[18] Data stored under an alias generally contains some kind of personal reference - albeit indirect. As such, it is important to distinguish between various types of aliases that are used in the various contexts:

With reference aliasing, the allocation feature is assigned to the data subject using a reference list (or reference file). Reference aliases can always be deleted by the data controller using the reference list. With disposable aliases, the assignment features are derived from personal data using special computing functions (hash functions). The methods used must be selected to ensure that inferences cannot be draw from the result about the individual persons or the identification features used. Disposable aliases are particularly suitable for longitudinal analyses in scientific research projects and statistics. With this type of aliasing, however, the data stored under the alias can only be assigned to the person using the alias if the original data used to create the alias is known and if the attacker knows how the alias was created (“Brute-Force Attack”).

Research Data Centers

The Research Data Centers (RDCs) run by the Federal Statistical Office, the Statistical Offices of the Federal States, the Institute for Employment Research of the Federal Employment Agency, and the Statistics of the Federal German Pension Insurance Association have for a number of years made an attempt to balance data protection requirements against the interests of the scientific community in using microdata. The statistical offices give scientists direct access to individual data, observing general data protection requirements.

The RDCs focus on microdata that has been cleared for remote data access.

However, the scientists do not access the statistical raw data or individual data managed by the offices direct, they access microdata sets, so-called scientific use files (SUFs) generated for various purposes in which only virtually or fully anonymous data are stored.

SUFs can be accessed on-site or off-site. With on-site access, the data are accessed in the protected facilities of the research data centers, whereas with off-site use, the data are accessed outside the research data center for a specifically defined research project.

As the statistical offices have no way of ensuring the data are used properly, extreme caution must be taken when rendering data files anonymous, taking all of the additional

18 BDSG 3, 6a.

information available to science into account. Access to official individual data are hence subject to the provisions set forth in the Law on Statistics for Federal Purposes.

Intensive use is being made of the newly established RDCs.[19] Yet this is certainly not where developments will end, as there continues to be a keen interest in making the utilisation of data even more flexible and above all in facilitating access from any location. Access via the Internet will likely be of key importance in the future. However, this would change the whole environment in terms of data protection, as it would no longer be possible to estimate the additional knowledge that might have been used to render individuals anonymous. In view of the unlimited amount of additional information available, individual data can hence only be used for uncontrolled Internet access if their absolute anonymity can be guaranteed. Anonymity "of a lesser quality" is not sufficient in view of the unlimited possibilities that exist for linking the widest range of databases.

19 Federal Data Protection Commissioner, TB 21, no. 7.6.