DIW Diskussionspapiere
Discussion Papers

Discussion Paper No. 249

Autonomous Organization of the (International) Scientific Community Would Simplify Data Protection in the Social Sciences and Encourage Reanalysis

by
Gert G. Wagner

Berlin, April 2001

Deutsches Institut für Wirtschaftsforschung, Berlin
Königin-Luise-Str. 5, 14195 Berlin
Phone: +49-30-89789- 0
Fax: +49-30-89789- 200
Internet: http://www.diw.de
ISSN 1433-0210

Autonomous Organization of the (International) Scientific Community Would Simplify Data Protection in the Social Sciences and Encourage Reanalysis

by GERT G. WAGNER

Abstract: The re-analysis of statistical data is an effective means of protecting the public from undiscovered errors in social science research. Re-analysis is critical, as there is no difference between official data and non-official data. However, discussions concerning data protection legislation do not usually take this into consideration. Proper data protection rules must make it possible to conduct independent re-analysis of protected data. The paper discusses the possibilities of self-binding in the (international) scientific community and creating a new kind of law which would provide scientific data with a special legal status *(Forschungsdaten-Geheimnis).*

Zusammenfassung: In den Sozial- und Wirtschaftswissenschaften besteht ohne die Moeglichkeit einer Re-Analyse von statistischen Ergebnissen (gleichermaßen amtlichen wie nicht-amtlichen) die Gefahr von nicht entdeckten Irrtuemern. Mit anderen Worten: Re-Analysen sind in der Wissenschaft die „Berufungsinstanz“, ohne die es keine funktionierende Scientific Community geben kann. Der Schutz vor fehlerhaften wissenschaftlichen Ergebnissen mit Hilfe von Re-Analysen ist ein „oeffentliches Interesse“, das in der Datenschutzdiskussion und insbesondere bei der Auslegung von Datenschutzregelungen bislang zu wenig beachtet wurde. Das Wissenschaftssystem und der Gesetzgeber sind gleichermaßen aufgefordert, Re-Analysen zu ermoeglichen ohne den Datenschutz zu verletzen. Als Instrumente werden Selbstbindungen im Wissenschaftssystem und die Schaffung eines gesetzlichen „Forschungsdaten-Geheimnisses“ diskutiert.

JEL Classification: A20, C81, K39

1. Outline of the Problem

Many scientific results have influence on private and public decisions. In the social and economic sciences these results are quite often based on empirical evidence. Their influence is usually indirect on public decisions, but sometimes direct on private decisions: for example, many people all over the world adjust their everyday nutritional habits to medical-statistical results.

While data protection regulations traditionally attempt to strike a balance between the basic rights of "data privacy" and "freedom of research," it is shown that a third legal good should be taken into consideration as well: the protection of the public from methodologically poor and incorrect statistical results.

For this reason data protection must ensure that the reanalysis of official statistics remains feasible both for independent science (including commercial research) and for public administrations.

When the actual practice of data protection is considered, it is apparent that one of the main reasons why official statistical offices are able to evaluate even sensitive, personalized data is their special ethical code, known in Germany as "statistical confidentiality" (*Statistik-Geheimnis*). As yet the international scientific community, with the exception of the field of medicine, has failed to develop clear, binding obligations for itself about how to deal with sensitive data. It is proposed in this paper that the international scientific community shall develop a "code of ethics" whose violation would entail sanctions. Such self-imposed obligations could be made especially effective through a new legislation of research-data confidentiality.

2. The Importance of Reanalysis for Science

Modern science defines itself according to the dictate of "intersubjective verifiability," i.e., empirical research must be open for reanalysis. Particularly for experimental or statistical data, reanalysis is important in consideration of

- the dependence of data and results on methods,

- errors in data collection and analysis and
- fraud.

The dependence of results on methods plays an important role for statistical data and empirical analysis. Statistical data are influenced by the collection methods selected (cf. Dillman 19xx), and the statistical methods of evaluation used (cf. Fitzenberger and Speckesser 2000).

Errors in evaluation are likely (cf. Williamson and Jones 1983)[1] and the possibility of fraud cannot be ruled out, as even scientists and statisticians may pursue self-serving goals (on this, cf. Finetti and Himmelrath 1999).

Society can protect itself from one-sided results (biased due to a certain method selected for data generation and/or evaluation), errors and fraud in statistical analysis only if the data used, i.e., the tables and micro-data upon which other results are based, are available for reanalysis by any scientist. Without such reanalyses, legislators, already at a disadvantage in their ability to evaluate information as compared to scientists, have no chance of protecting society from weak results and even fraud.

Theoretically, the optimum situation would be one in which a scientist wishing to reanalyze data receives not only the data collected, but also the opportunity to check how the data were generated. Thus, in the case of the social and economic sciences, it would be possible to check the fieldwork, i.e. whether the information from a survey or in a file produced by a survey actually corresponds with that provided by the subjects.

---

1 For this reason such journals as the "Journal of Political Economy" have established their own sections for reanalysis.

3. The Status Quo of Data Protection for Research Data in Germany

3.1 Definition and Concepts

This paper deals with only one dimension of data protection in the field of science, that of the anonymity of units of collection. An equally important second dimension, that of the consent of persons to the collection and processing of their data, is beyond the scope of this paper.[2]

All over the world, "personalized" data, i.e., those which reveal the personality hidden behind a record, are subject to "data privacy" and may not be collected, stored and analyzed without restriction. Such data may be generated and processed only with the explicit consent of each affected individual.

In practice, data on firms and companies are treated all over the world as personalized data. This treatment is based on the principle of the justified interest of businesses to keep data secret from competitors (although this principle is undermined in part by the legal obligation of companies to make certain data public).

Data protection within official statistics (called "statistical confidentiality" in Germany [*Statistikgeheimnis*]) is older than modern data protection. Statistical confidentiality guarantees that individual data will not be made known to third parties. For pragmatic reasons, it applies to both personalized and business-related data, as its purpose is to secure the readiness of all parties to provide information.

It is indisputable that absolutely "depersonalized" data, for which reverse-tracking for the purpose of identifying individual units is precluded with one hundred percent certainty, are

---

2 The author considers it a matter of course ethically that explicit consent is a fundamental requirement. Exceptions to this rule may be made only in few areas of medical research, where explicit consent is difficult or even impossible to obtain. The "right to data privacy" becomes particularly awkward when data produced by other processes, and for other purposes (for instance, register data subject to social law), are to be used for statistical or scientific purposes, respectively. This raises both the question as to whether this is permissible in principle and the question of who should be allowed to evaluate these data, as they were not produced for statistical purposes and were hardly collected for independent scientific studies.

not subject to data-protection legislation and may be analyzed without any further protective measures.

In practice, there is no such thing as absolutely depersonalized data. As long as a record contains even one bit of information, there is a chance that the unit of collection can be identified (at least by coincidence). In practice, the border between truly depersonalized and personalized data is blurred, since purely formal depersonalization (by omitting the name and address) generally does not suffice to make a record "truly depersonalized." In certain cases at least it is not difficult to identify unit beyond a statistical record (e.g., large companies in a certain region).

It stands to reason that the actual depersonalization of data is weakened by the practice of linking different records to increase their information content. Especially in the fields of labor economics and epidemiology, data linked in such a manner are becoming increasingly important (cf. Westergard-Nielsen 1999).

3.2 The Reality of Science and Data Protection

Agencies of official statistics are allowed to process non-depersonalized data, as all states have granted their official statistical offices a special legal status. Thus they are allowed to collect, retrieve and analyze personal data because the states believe the statistical offices are effectively able to ensure the integrity of their computers and buildings (so that nobody can steal data) and the integrity of their personnel (by means of a professional code of ethics and special legal enforcement in their work contracts).

Scientists do not have the same legal status as official statisticians.[3] Thus scientists are not allowed to analyze statistical data (provided by official statistics) which are "personalized" (i.e., the identification of which is possible). If no identification is possible, scientists are allowed to analyze these data but such complete anonymization is extremely rare. Data

---

3 However, medical doctors have certain privileges to handle personalized data on the basis of their special professional code of ethics.

protection in science thus refers not only to personalized data, but also deals with the issue of whether and, especially, under what conditions, data can be classified as "effectively depersonalized." A data set is considered as "effectively depersonalized" when a marginal risk of reidentification still exists, but it is not reasonable that anyone will spend the "disproportionate costs" required to identify a statistical unit. Costs can consist of time, machinery used, and/or legal penalties. Thus the estimation of "disproportionate costs" is judged according to the position of the scientist using such data. Such cost-benefit calculations can have quite different results, especially when opportunity costs (i.e., the possible consequences of the action) are taken into consideration along with direct costs. It is in fact easier for a youthful computer hacker with practically no fear of legal sanctions to re-personalize a record than it is for a scientist who has higher costs (including social costs) because he faces the risk of destroying his reputation and thus his career if he intentionally re-personalizes units of collection.[4,5]

---

4 It is not the author's intention here to claim that scientists are distinguished as especially ethical personalities, who *eo ipso* guarantee data protection on the basis of a "special ethics" of science. On the contrary: scientists, too are prone to deceit, as apparent in the increasing number of fraud scandals in sciences, and in the life sciences in particular (cf. Finetti and Himmelrath 1999). But in terms of the anonymity of units of collection, despite the temptations to which scientists are subjected no less than other individuals, one can rely on scientists' own interest in complying with data protection because scientists have nothing to gain from violating strict data protection standards. Scientists have no interest in re-personalizing individual cases, as the process of re-personalization cannot be publicized. Through such illegal behavior scientists would endanger their careers without any corresponding benefits to outweigh the risks. In the social and economic sciences, surveys representative of the population are, in a manner of speaking, effectively protected against systematic data abuse by their representativeness: these data, completely lacking any sorting characteristics, are of no interest whatsoever for marketing purposes, as the individuals surveyed are dispersed across the entire territory of a country and are thoroughly inhomogeneous. Such statistical data cannot be sold to a company which has a targeted population of customers.

5 Between these extremes are groups like employees in marketing departments and election assistants. These groups, too, face negative consequences not only from the law, but also from the "professional culture" within which they work. However, how such cultures look outside the sphere of science is not known at this time.

Because of the limitations to this everyday interpretation of "effective anonymity," many data producers who possess effectively depersonalized data are careful to supply their data only to scientists and scientific institutions who have much to lose as individuals and as organizations by not taking data protection seriously. Different data producers as the *Statistisches Bundesamt* (Federal Statistical Office (cf. Bizer 1992: pp. 398) and the *Deutsche Institut fuer Wirtschaftsforschung* (German Institute for Economic Research) (Seufert and Wagner 1998) proceed accordingly. This procedure does not constitute an obstruction of scientific research.

A special problem are "register data" which may be used only for purposes of social administration. In Germany, for instance, such data may only be evaluated when a research project is defined in the interest of a social insurance agency. This is problematic as the statistical evaluation of data by the social authorities is hardly a pure act of administration, but is rather based on a scientific foundation. In the interest of the public, data and evaluations from such agencies must be scientifically verifiable, a condition which can be fulfilled only through independent reanalysis. Only by upholding this principle of public interest, i.e., by ensuring that the relevant social authorities permit every proposed reanalysis to be performed, is it possible to eliminate the danger of weak research and even frauds.

Non-published research in commercial interest and the lack of reanalyses of such research do not constitute a problem for the public because the risk of weak research presents a risk only to the commercial enterprise. However, when commercial research is publicized in order to inform (or influence) the public, legislators should establish the right to reanalysis of the findings.

Data-protection legislation would allow scientists to check how data are generated if subjects were made aware of the possibility of these inquiries and agreed to cooperate (in such a case the storage of addresses would be permitted because the survey or study would not be complete after the first round of data collection).[6]

---

6 Such an option will remain an exception, however, for practical and financial reasons alone. In everyday practice it is helpful when data collectors provide data which are supplemented by "field information" on the micro-level, e.g., information about units which did not respond, the number of attempts to contact a unit, characteristics of the interviewer, etc. Edited or "imputed" information

However, even microdata which do not belong to official statistics or the administration and are generated by the scientific community itself, sometimes are not easy to analyze because the risk of re-personalization is judged to be very high (for example when regional identification such as the county or the zip code is in a data file).

The scientific community judges the better possibilities of official statistics to analyze data which are not re-personalized to bethoroughly unfair. Especially the lack of access to some "data registers" (including the registers of social security administration) are considered to be extreme limitations on science.

4. Possibilities for Improving Data Protection Practice

4.1 Autonomous Organization of the Scientific Community

Regarding the scientific community's dissatisfaction with the practice of data protection, two basic remarks are necessary:

- In the past, scientific organizations in Germany frequently have neglected to take advantage of opportunities to influence the legislative process in the ways expected in parliamentary democracies, for instance, through lobbying (cf. Wagner 1999b).

- The social and economic sciences have yet to develop any self-binding rules about compliance with data protection which can be taken as seriously as those existing in the field of medicine and in official statistical offices.

If the scientific community does not systematically attend to the national and international[7] legal formulation of data-protection legislation, it can hardly come as a surprise that legislators do not systematically incorporate the relevant scientific organizations into hearings

---

(which thus are "generated" using hypotheses rather than collected from real units) must be flagged as such. Such measures are sufficient to provide a high degree of verifiability.

7 Especially on the EU level.

for new legal frameworks. This should be changed: if data protection is to be subject to detailed regulation, detailed rules must also exist for the hearings held on such legislation.[8]

The "privileged position" of official statistical offices and medical research in handling sensitive data is based in part on a "statistics-specific" professional ethic, reinforced by the status of their staff as civil servants, which permits especially effective sanctions. This leads to the reasonable question of whether such a system might be feasible for the scientific community in general.

Considering the estimated risk of abuse of personalized and effectively depersonalized data, a binding "code of ethics" or "code of good practice" specific to the field, accepted worldwide, and taking consideration of the special interests of individual disciplines, would probably be extremely helpful. The *Ethik-Kodex* (ethical code) of the *Deutsche Gesellschaft fuer Soziologie* (German Society for Sociology) and the *Berufsverband Deutscher Soziologen* (Professional Association of German Sociologists) (1992: Sections I.B.7 and 8) is an example of such a code. However, this Kodex, like many others, is certainly not specific enough and does not provide for any (enforceable) sanctions. It suffers especially from not being integrated into a legal framework which allows severe sanctioning. Now as ever, Kaase et al. (1980: 293) is correct in asserting that professional rules cannot replace legal regulations. Nevertheless, they may well make legal regulations practicable in the area of contention between scientific freedom and data protection and are therefore a goal worth aspiring to.[9]

---

8 In addition to the level of legislation, it might also be useful for the practice of everyday administration to form "joint" decision-making bodies in which such persons and institutions as "ethics commissions" and data protection experts cooperate.

9 An effective code of ethics must also include a binding pledge by research funding institutions that recipients of funding will be required to release their data for reanalysis after an appropriate period of time. The definition of what period of exclusive data use is "appropriate" to the specific field must be included in this obligatory pledge.

4.2 "Research-data Confidentiality" as a New Legal Status

To encourage and reinforce binding pledges by the scientific community, the possibility of establishing a legal status of "research-data confidentiality" should be investigated, both on the national and the international levels (cf. also Bizer 1992: pp. 229). One objective of such a legal status is to ensure that scientists can be especially trusted to process sensitive (i.e., effectively depersonalized and personalized) data. Another requirement of such an obligation is that it protect (personalized) data collected or prepared for purposes of research from access by the state for the purpose of investigations (e.g., in cases of criminal and tax law).[10]

This could also mean that the separation of research stipulated in data protection laws, between "official statistics" on the one hand and "free scientific research" on the other, would be relaxed. This would strengthen both the field of science and the official statistical offices alike: science would receive recognition of its special professional status, while the scientific character of official statistics would become more visible.

"Research-data Confidentiality" could be operationalized by stipulating that a Master of Science degree is the "license" required for any person wishing to work with sensitive data. This requirement would have to be linked to the practicability of trusting a licensed scientist to extract such a pledge from the staff which works with him, such as that long since granted to medical doctors and attorneys (cf. § 203 Paragraph 3 of the German penal code [*Strafgesetzbuch*]).

Anyone who violates the rules of research-data confidentiality loses his "license to analyze" and moreover is legally punished. Such licenses could be granted in the context of an autonomous professional administration.

A central component of research-data confidentiality would be an obligation to learn about data-protection measures and opportunities while earning the professional degree. Thus data protection must be incorporated either into training for all relevant scientific disciplines, or

---

10 Thus it would constitute a ban on confiscation and a right to withhold witness to the police, judicial authorities and intelligence services, as well as to authorities who provided register data for scientific analysis (and may be interested in the research findings, for example in transfers on the individual level).

into the statistics training which is standard for a number of disciplines. Data protection experts should participate in practical instruction.

Research-data confidentiality inevitably raises the question of how scientific research is defined. This question can be answered in principle using two methods: the first in terms of qualifications, i.e., research is that which is performed by a trained researcher with a Master of Science degree; the second is "institutional," i.e., research is that which takes place in independent research institutions (i.e., in institutions which are legally defined to serve scientific research).

If research is defined institutionally, the risk arises that data protection might be abused as a check and control instrument for research desired (or not desired) by the government; if research is defined in terms of qualifications, the risk is that the constitutionally chartered "freedom of research" might be violated, as research would be permitted only for licensed researchers, and not for anyone else. Freedom of research is a precious gift, however. For example, in the German constitution (*Grundgesetz*), is written: "Art and science, research and instruction are free" (Artikel 5, Paragraph 3 GG).

The commandment of freedom of science is, of course, interpreted by such bodies as the German Constitutional Court (*Bundesverfassungsgericht*) to concern the "serious, systematic attempt to determine truth" (according to Bizer 1992: 45). This means that not every dilettante can claim to be a scientist, but that the "seriousness and systematicity" of their work can be investigated by a third party. In this respect, "freedom of science" is already restricted today. And those who work with new methods would not be excluded on principle, they need only possess a Master of Science degree earned on the basis of standardized methodological requirements (such training is a good basis for the development of new paradigms in any case).

An optimum solution for how to define "science" is apparently not easy to find. In a democratic constitutional state, all solutions which could lead to checks and censorship should be avoided. Thus the freedom of research remains most likely to be guaranteed if scientific research is defined in terms of qualifications.

5. Summary and Prospects

To date, judgements of the data-protection component of research projects attempt to strike a balance between only the basic rights of “freedom of research” and “data privacy.” In this paper it is argued that the “protection of the public from falsifications and deficient results” also must be taken into consideration.

It is clear from the perspective of science theory that the quality of empirical results based on data can only be determined by means of reanalysis. For this reason data sharing by producers of data and other scientists is a command central to the scientific project. This statement is also valid for data from official statistics, as these also are based on scientific methods (just as for commercial research to the extent that data are published in order to inform or influence the public).

Scientists (including official statisticians) and data protection experts therefore should work out generally valid, clear regulations for the manipulation and reanalysis of personalized and (effectively) depersonalized data. Such regulations must recognize especially that there is in practice no such thing as absolutely depersonalized data about individuals and business enterprises. In imitation of tried and true regulations within official statistics offices, the scientific community should develop a binding code of handling and analysis which is also required to be included in the course of study and final exams for students of the scientific professions. Then the division between “official statistics” and “free research” stipulated in data-protection regulations would become a thing of the past.

A new legal status of “research-data confidentiality” to ensure the anonymity of research data, beginning with the scientific community’s own binding pledge, submits for discussion a legally anchored research-data confidentiality, which would not constitute a data-protection carte-blanche for researchers, but instead would prescribe high demands for the ethics and the practice of the sciences.

References

Bizer, J. 1992: Forschungsfreiheit und Informationelle Selbstbestimmung, Baden-Baden.

Deutsche Gesellschaft fuer Soziologie und Berufsverband Deutscher Soziologen 1992: Ethik-Kodex (reprinted in: *Soziologie (DGS Nachrichten)*, Heft 4, 1998, 79-85).

Dillman, Don A. 1996a. "Why Innovation is Difficult in Government Surveys," *Journal of Official Statistics*, 12 (2), 113-124.

Finetti, M./Himmelrath, A. 1999, *Der Suendenfall – Betrug und Faelschung in der deutschen Wissenschaft*, Stuttgart: dtv.

Fitzenberger, B./Speckesser, St. 2000: Zur wissenschaftlichen Evaluation der Aktiven Arbeitsmarktpolitik in Deutschland – Ein Ueberblick, *ZEW Discussion Paper Nr. 00-06*, Mannheim: mimeo.

Hamm, R./Moeller, K.-P. (eds.) 1999: *Datenschutz und Forschung*, Baden-Baden: Nomos.

Kaase, M./Krupp, H.-J./Pflanz, M./Scheuch, E. K./Spiros, S. 1980: *Datenzugang und Datenschutz – Konsequenzen fuer die Forschung*, Koenigstein/Taunus: Athenaeum.

Seufert, W./Wagner, G. G. 1998: Problems of Dissemination and Examples of "Public Use Micro-Data" in Germany, in: *American Statistical Association 1997 Proceedings for the Government and Social Statistics Sections*, 58-64.

Simitis, S. 1999: Diskussionsbeitrag, in: R. Hamm/Moeller, K.-P. (eds.) 1999: *Datenschutz und Forschung*, Baden-Baden: Nomos, 29-31.

Wagner, G. G. 1999a: Ziele und Unabhaengigkeit der Wissenschaft sind Instrumente eines effektiven Datenschutzes, in: R. Hamm/Moeller, K.-P. (eds.) 1999: *Datenschutz und Forschung*, Baden-Baden: Nomos, 14-20.

Wagner, G. G. 1999b: Self-regulation Within the Research Community as a Means for Shaping the Statistical Infrastructure, in: Statistisches Bundesamt (ed.), *Kooperation zwischen Wissenschaft und Amtlicher Statistik*, Wiesbaden: Metzler Poeschel, 59-61.

Westergard-Nielsen, N. 1999: Linking employer-employee data – the Danish experience, , in: Statistisches Bundesamt (ed.), *Kooperation zwischen Wissenschaft und Amtlicher Statistik*, Wiesbaden: Metzler Poeschel, 124-126.

Williamson, S./Jones, W. 1983: Computing the Impact of Social Security Using the Life Cycle Consumption Function, in: American Economic Review, 73, 1036-1052.